Using whole genome sequencing data to identify strain-types, transmission enhancers and novel drug resistance mutations of *Mycobacterium tuberculosis*

Gary Napier

Thesis submitted in accordance with the requirements for the degree of

Doctor of Philosophy

University of London

September 2022

Department of Infection Biology

Faculty of Infectious Tropical Disease

LONDON SCHOOL OF HYGIENE & TROPICAL MEDICINE, UNIVERSITY OF LONDON

Funded by BBSRC

Research group affiliation(s): Taane G Clark
Martin L Hibberd
Jody E Phelan

I, Gary Napier, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in this thesis.

Signed [redacted] Date 18/01/2023

Abstract

Tuberculosis disease (TB), caused by bacteria in the *Mycobacterium tuberculosis* complex (MTBC) including *M. tuberculosis* (*Mtb*), is a leading cause of global morbidity and mortality. Drug resistance, especially to first-line rifampicin (RIF) and isoniazid (INH) drugs, is making the control of the disease difficult. To understand aspects of the genomic epidemiology of TB, with insights for disease control, this thesis analyses whole genome sequences from a large global dataset (n=~32k). Understanding genetic variation in the MTBC genome informs strain-typing, phylogenetic clustering and transmission patterns, and predicts genotypic drug resistance. In turn, these analyses can assist diagnostic design and provide epidemiological insights, as well as improve clinical and surveillance decision making, leading to improvements in TB control.

To enhance strain-typing, the 32k dataset was used to infer synonymous single nucleotide polymorphisms (SNPs) which uniquely identify 90 MTBC clades (lineages and sub-lineages). By finding those SNPs with perfect sub-population differentiation (fixation index F_{st} values = 1), a new barcode was inferred, providing greater resolution of the MTBC phylogenetic tree than previous work, including by identifying 30 new sub-lineages with associated barcoding SNPs. Spoligotyping is a method of strain-typing, where a 42-place binary barcode can be generated from the presence or absence of so-called 'spacers' in repeat regions of the MTBC genome. Spoligotypes can be inferred from whole genome sequencing data, and software was developed (*Spolpred2*) to rapidly do this. Spoligotyping has lower resolution and precision in its ability to discern a sample's place on the MTBC phylogenetic tree compared to the SNP-based barcoding of lineages outlined above, but is nevertheless widely used. Therefore, correlations between various levels of phylogenetic lineage and spoligotypes were investigated, and revealed high concordance between the two systems at the highest lineage levels.

Pakistan is a high burden nation for TB, and the profiling of *Mtb* drug resistance and transmission was conducted on 535 samples across that country. High relatedness of samples based on genome-wide SNP differences was used to infer transmission clusters, which provided a proxy phenotype for increased transmissibility. Using these transmitted and other potentially non-transmitted samples, a genome-wide association study (GWAS) was conducted to find associations between SNPs and increased transmission, revealing the *nusG* gene to be the most significant ($P=5.8x10^{-10}$), after adjustment for population structure. In terms of drug resistance, there were mismatches between the phenotypic drug susceptibility tests (DST) in the data and genotypic predicted drug resistance, revealing putative SNPs conferring drug resistance in Pakistan.

Mutations in MTBC bacteria that cause drug resistance often come with a fitness cost. To compensate for this cost, the bacteria can develop changes in genes which have similar roles to that of the (pro-)drug targets. To improve genotypic predictions for drug resistance to RIF and INH, samples with compensatory mutations, but no known drug resistance mutations were found in the 32k dataset, thereby leading to the identification of novel putative drug resistance mutations in the relevant genes (*rpoB* for RIF and *katG* for INH). Unsurprisingly, there were no new *rpoB* mutations found, but 31 novel *katG* putative resistance mutations were identified. Additional analyses, including *in silico* modeling of the *katG* gene, were undertaken to provide evidence that the putative INH resistance mutations may be causally relevant.

Overall, this thesis has reinforced the benefits of using whole genome sequencing data to provide insights into TB control. Such insights are needed to meet international targets for disease eradication.

Acknowledgements

Many thanks to my supervisors Taane Clark, Martin Hibberd and Jody Phelan.

Thanks also to the BBSRC and all members of the LIDo co-ordination team, especially Nadine Mogford, Sam Alsford and Theresa Ward.

Thanks to past and present members of the Clark-Campino lab for help and support over the course of this thesis, including Amy Ibrahim, Anna Turkiewicz, Emilia Manko, Ernest Diez Benavente, Julian Libiseller-Egger, Matthew Higgins, Pepi, Yaa Oppong, and all the other members.

Thanks also to other members of LSHTM, including Stephane Hue and Dave MacLeod.

Thanks for support to my family, Mum, Dad, Lauren, Mick, Rachel, Tommy, Ada, Bertie, and Reggie (who just made it in time).

List of publications

In the thesis, as Chapters:

[1] **Napier G**, Campino S, Merid Y, Abebe M, Woldeamanuel Y, Aseffa A, Hibberd ML, Phelan JE, Clark TG. Robust barcoding and identification of *Mycobacterium tuberculosis* lineages for epidemiological and clinical studies. Genome Med. 2020 Dec 14;12(1):114. doi: 10.1186/s13073-020-00817-3. PMID: 33317631

[2] **Napier G, et al.** Comparison of *in silico* predicted *Mycobacterium tuberculosis* spoligotypes and lineages from whole genome sequencing data. Submitted, Sci. Rep..

[3] **Napier G**, Khan AS, Jabbar A, Khan MT, Ali S, Qasim M, Mohammad N, Hasan R, Hasan Z, Campino S, Ahmad S, Khattak B, Waddell SJ, Khan TA, Phelan JE, Clark TG. Characterisation of drug-resistant *Mycobacterium tuberculosis* mutations and transmission in Pakistan. Sci Rep. 2022 May 11;12(1):7703. doi: 10.1038/s41598-022-11795-4. PMID: 35545649

[4] **Napier G,** Campino S, Phelan JE, Clark TG. Large-scale genomic analysis of *Mycobacterium tuberculosis* reveals extent of target and compensatory mutations linked to multi-drug resistant tuberculosis. Sci Rep. 2023 Jan 12;13(1):623. doi: 10.1038/s41598-023-27516-4. PMID: 36635309

Additional published papers, contributed to during the PhD.

[5] Deelder W, **Napier G**, Campino S, Palla L, Phelan J, Clark TG. A modified decision tree approach to improve the prediction and mutation discovery for drug resistance in *Mycobacterium tuberculosis*. BMC Genomics. 2022 Jan 11;23(1):46. doi: 10.1186/s12864-022-08291-4. PMID: 35016609.

[6] Gómez-González PJ, Perdigao J, Gomes P, Puyen ZM, Santos-Lazaro D, **Napier G**, Hibberd ML, Viveiros M, Portugal I, Campino S, Phelan JE, Clark TG. Genetic diversity of candidate loci linked to *Mycobacterium tuberculosis* resistance to bedaquiline, delamanid and pretomanid. Sci Rep. 2021 Sep 30;11(1):19431. doi: 10.1038/s41598-021-98862-4. PMID: 34593898

[7] Khan AS, Phelan JE, Khan MT, Ali S, Qasim M, **Napier G**, Campino S, Ahmad S, Cabral-Marques O, Zhang S, Rahman H, Wei DQ, Clark TG, Khan TA. Characterization of rifampicin-resistant *Mycobacterium tuberculosis* in Khyber Pakhtunkhwa, Pakistan. Sci Rep. 2021 Jul 9;11(1):14194. doi: 10.1038/s41598-021-93501-4. PMID: 34244539.

[8] Khan AS, Phelan JE, Khan MT, Ali S, Qasim M, Mohammad N, **Napier G**, Ahmad S, Alam J, Khattak B, Campino S, Clark TG, Khan TA. Genetic mutations underlying isoniazid-resistant *Mycobacterium tuberculosis* in Khyber Pakhtunkhwa, Pakistan. Tuberculosis (Edinb). 2022 Nov 28;138:102286. doi: 10.1016/j.tube.2022.102286. Online ahead of print.

Table of Contents

List of abbreviations

AMK	Amikacin
BCG	Bacilli Calmette-Guerin vaccine
CAP	Capreomycin
DR-TB	Drug-resistant TB
DST	Drug susceptibility testing
EAI	East-African-Indian
ENA	European nucleotide archive
ETH	Ethionamide
FLQ	Fluoroquinolone
GWAS	Genome-wide association study
HIV	Human immunodeficiency virus
IGRA	Interferon-gamma release assay
IMR	Intermediary metabolism and respiration
INH	Isoniazid
indel	Insertion and deletion
KAN	Kanamycin
LAM	Latin American Mediterranean
LPA	Line probe assay
LSPs	Long Sequence Polymorphisms
MDR-TB	Multi-drug resistant TB
Mtb	*Mycobacterium tuberculosis*
MTBC	*Mycobacterium tuberculosis* complex
MIRU-VNTR	Variable-number tandem repeats of mycobacterial interspersed repetitive units
NGS	Next Generation Sequencing
NS	Non-synonymous
PAS	Para-aminosalicylic acid
PCA	Principal component analysis
PCR	Polymerase chain reaction
PE	Proline-glutamate
PPE	Proline-proline-glutamate
PZA	Pyrazinamide
RDs	Regions of difference
RIF	Rifampicin
RRDR	Rifampicin resistance-determining region
RR-TB	Rifampicin-resistant TB
S	Synonymous
SNP	Single nucleotide polymorphism
STM	Streptomycin
TB	Tuberculosis disease
TST	Tuberculin skin test

VCF	Variant call file
VDA	Virulence, detoxification, adaptation
WGS	Whole genome sequencing
WHO	World Health Organization
XDR-TB	Extensively drug-resistant TB

Chapter 1
Introduction

Global burden of tuberculosis disease

The *Mycobacterium tuberculosis* complex (MTBC) is a collection of closely-related mycobacteria strains that include *Mycobacterium tuberculosis sensu stricto*, *Mycobacterium africanum* (together *Mtb*) and animal strains such as *Mycobacterium bovis* (in cattle), which all cause tuberculosis disease (TB) in humans. Every year, 10 million people fall ill with TB, and 1.5 million people die from TB each year, including ~214k deaths in HIV positive people – making it the world's top infectious killer after COVID-19.

By some estimates, *Mtb* is responsible for the most pathogen-related deaths throughout history [1], and is the thirteenth leading cause of death worldwide [2]. Such a high burden persists despite twentieth-century innovations in antibiotics as well as the Bacilli Calmette-Guerin (BCG) vaccine being administered since 1921. Significantly, the number of deaths has risen for the first time in over ten years, and recently because of compromised access to TB diagnosis and treatment during the COVID-19 pandemic [2]. It is estimated that nearly half of the ten million cases did not gain access to care in 2020 and were not even reported. Treatment of drug-resistant TB as well as preventative treatment also 'dropped significantly'. The pandemic has 'reversed years of global progress' in reducing deaths; not since 2005 has there been a year-on-year increase (5.6%), with total deaths at the level of 2017 [2].

TB is curable with drugs, but drug resistance poses threats to disease control. There were ~132k cases of resistance to first-line isoniazid and rifampicin drugs (together called multidrug resistant; MDR-TB) or rifampicin alone (RR-TB). In addition, ~25k detected cases of resistant to rifampicin and a fluoroquinolone (pre-XDR-TB), or further one of bedaquiline and linezolid (XDR-TB). The highest proportions of MDR-TB are in former Soviet Union countries. Globally the burden of MDR-TB is described as 'stable' and current treatment success rates for MDR-TB/RR-

TB is 59%. Nevertheless, these types of TB involve taking multiple drug regimens for nearly two years. The drugs are expensive and toxic, decreasing the likelihood of compliance, and in turn worsening the problem of drug resistance.

Though TB occurs globally, the burden is not evenly spread: 30 high-TB-burden countries accounted for 86% of new cases, with India, China, Indonesia, the Philippines, Pakistan, Nigeria, Bangladesh and South Africa the top eight countries for absolute numbers of cases and accounting for two-thirds of all cases. The World Health Organization (WHO) region of South-East Asia (43%) accounts for most cases, followed by Africa (25%) and the Western Pacific (18%). Overwhelmingly, >95% of cases and deaths are from developing countries (**Figures 1 and 2**).

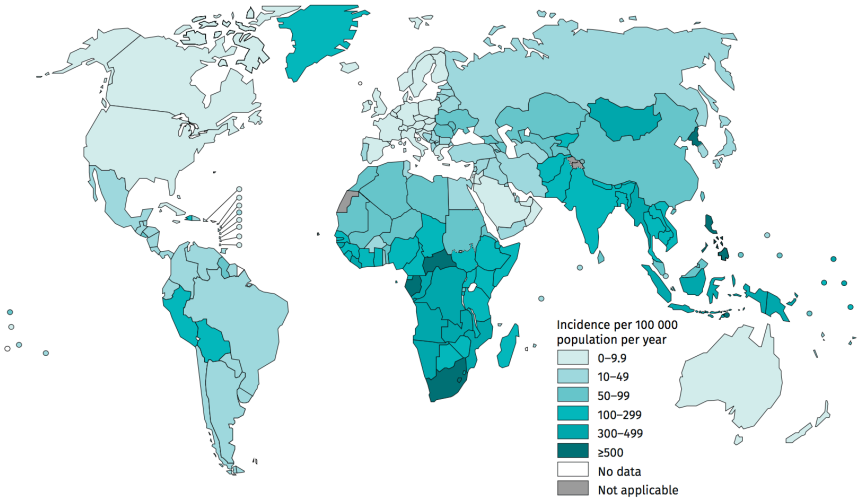


Figure 1: Estimated TB incidence rates per 100,000 population per year (from the WHO Global Tuberculosis Report 2021) [2]

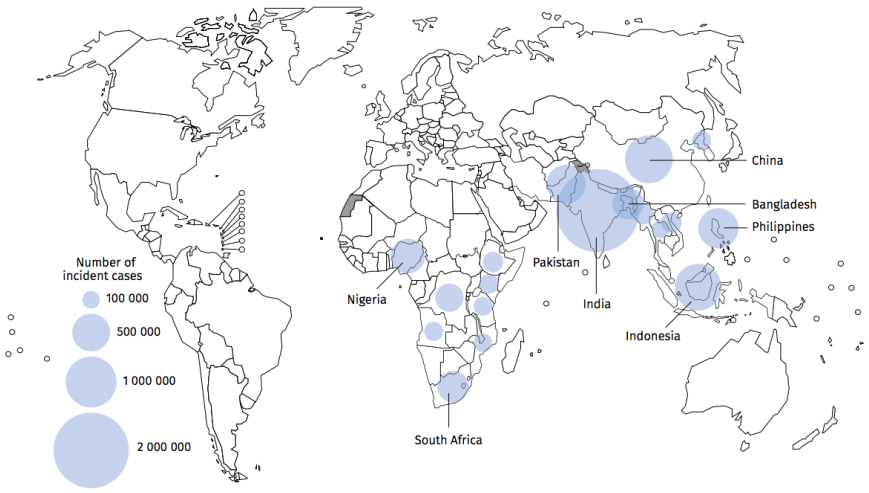


Figure 2: Estimated TB incidence in 2020, for countries with at least 100,000 incident cases. Labelled countries are those in the top eight for number of cases and account for two-thirds of all cases (from WHO Global Tuberculosis Report 2021).

TB is an endemic global burden, to the extent that it is thought that one quarter of the world's population is infected with the *Mtb* bacteria; although infections will remain in a state of low-infectivity (known as 'latent') in all but 5–15% of cases over a lifetime [3]. The global TB incidence is falling at 2% per year, with a cumulative reduction of 11% between 2015 and 2020. However, this was only just over halfway to the 20% End TB Strategy milestone. So, while the overall picture is that incidence rates are falling, the gains are not sufficient to meet the WHO's goal of ending TB by 2035. Therefore, TB disease is likely to remain among the world's biggest health threats for many years in the future.

Disease aetiology, risk factors and host susceptibility

When an infection is active in the lungs, *Mtb* bacteria are spread via aerosol droplets from, for example, a patient's cough. The innate immune response among the pneumocytes is the first obstacle encountered by the bacilli, whereupon pattern recognition receptors on alveolar macrophages recognise pathogen-associated molecular patterns, such as lipomannan, found in the pathogen's cell wall. Neutrophils, a type of granulocyte, are also involved, but play a lesser

role. Macrophages will then engulf the *Mtb* into phagosomes whereupon they attempt to kill the pathogen by way of phago-lysosomal fusion (forming a phagolysosome), inducing exposure to cytotoxic compounds such as reactive oxygen and nitrogen species, in turn leading to a cascade of other compounds hostile to *Mtb* [4].

The *Mtb* pathogen has evolved virulence factors to evade innate immune response destruction, including mechanisms to prevent lysosomal fusion and the hinderance of the formation or binding of enzymes which induce cytotoxic compounds into the phagosome [5]. Hence, *Mtb* become resident inside immune cells (macrophages and neutrophils) whereupon they multiply and cause primary or latent infection. However, not all exposure to *Mtb* leads to infection, with a great majority of people (90%) eliminating the pathogen without priming antigen-specific T-cells, and a significant number of people (10-20%) are even able to clear the pathogen despite sustained exposure. Of those that do become infected, it is estimated that up to 95% develop latent TB, and the remaining 5% developing 'primary' TB, that is, active TB within two years of infection, either with or without short latency [3].

Latency involves the adaptive immune response: macrophages will recruit CD4 (T-helper, T_H) and CD8 (T-cytotoxic, T_C) cells to the site of the infection (via lymph nodes). Here, a Ghon's complex, or granuloma, is formed **(Figure 3)**, with the bacteria inside infected macrophages being at the centre of multiple surrounding immune cells, including epithelioid and multinucleated giant cells, with T_H1 continuously providing support in the periphery. Granulomas can bring the infection to an equilibrium whereby the rate of *Mtb* proliferation equals the rate of killing. At this stage, the bacteria are classically thought to be ‘metabolically quiescent’, having undergone changes that render them phenotypically different from an active

infection. However, the picture of TB infection being either latent or active is perhaps outdated, and a continuum of disease status is now favoured [6] [7] [8].

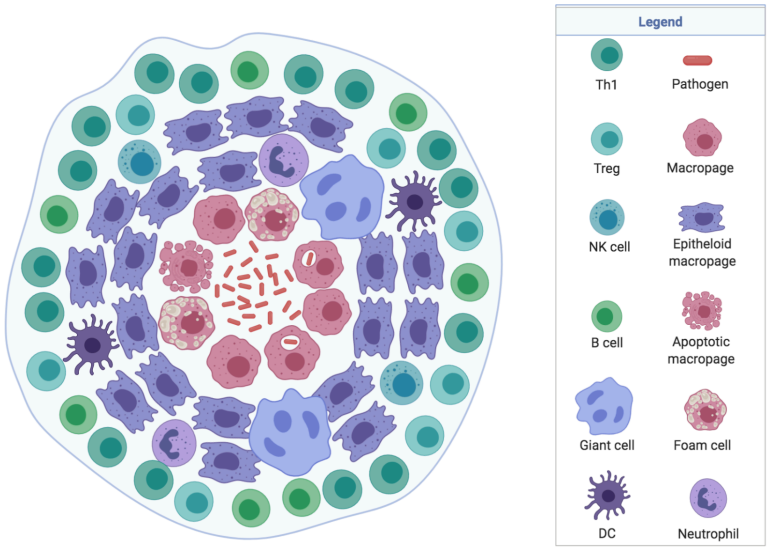


Figure 3: Granuloma. *Mtb* cells are under control in the centre and are surrounded by multiple types of innate and adaptive immune cells. Created with *BioRender.com*

This equilibrium can become fragile when the host is immunocompromised by, say, age, infection, immunosuppressive drugs (as in cancer treatment) or HIV infection. Although, multiple other comorbidities many be associated with the likelihood of active infection, and indeed may be reciprocal [9], [10], [11]. The classical picture of disease progression from latent to active sees the T_H1 cells weaken in immunocompromised patients, and they can no longer maintain the granulomas and the bacteria are free to assume their active state, causing pulmonary or extra-pulmonary TB (including dangerous infections of the meninges or blood, though infections can occur practically anywhere [12]).

Though symptoms of active TB are very noticeable (e.g., fever, night sweats, weight loss, persistent cough), primary infection is difficult to spot in that symptoms are rare or non-specific, which is problematic for TB treatment, given the importance of early diagnosis. Such clinical

aspects are often inadequate in areas of high poverty, thus increasing risks of active infections, in turn worsening the spread in perhaps already endemic areas [13].

Despite environmental factors, host susceptibility is important, though the precise role of genetic factors in innate and adaptive immune response, and the genetic susceptibility to progression from latent to active disease is far from clear [3]. Moreover, there is evidence that strains of MTBC have co-evolved with human populations such that interactions between host and pathogen influence disease outcome and virulence [14], [15].

Diagnosis, treatments, drug resistance

Of critical importance to TB control is swift, accurate diagnosis, and subsequent access to an effective drug regimen. Several methods exist for diagnosis of latent or active TB, though essential to all of them is the detection of the presence of the *Mtb* bacteria. Active and latent diagnoses require different methods, all having certain drawbacks.

For latent TB, the recommended methods are the *tuberculin skin test* (TST) or the *interferon-gamma release assay* (IGRA). While the TST is inexpensive, and therefore widely used in developing countries, the IGRA sees fewer false positives after BCG vaccination [16]. As mentioned above, the binary classification of TB disease into active and latent is possibly outmoded (patients are 'pragmatically classified' [8]), and therefore the decision of when and how to intervene is controversial [6]. It may be more appropriate to treat patients with active but asymptomatic/subclinical TB for example.

For active TB, there are three main methods: smear microscopy, mycobacterial culture, and molecular diagnostic tests. The most widely used diagnostic is smear microscopy. Although cheap and simple, its implementation requires specialists, and its sensitivity/specificity are low [10]. Also widely used is solid or liquid mycobacterial culture. This method is highly accurate

since it directly grows the bacteria if they are present. The main drawbacks are that the process is very slow (around a month for solid medium, or up to 21 days in liquid) due to the slow replication rate of *Mtb*, and trained personnel are required to work in elevated safety conditions. The XpertMTB/RIF assay is a molecular diagnostic test that detects *Mtb* DNA as well as for RR-TB, and was endorsed by the WHO in 2010 [17]. This method can be performed on sputum samples, is very quick (<2 hours), has high sensitivity and specificity, and requires little specialist equipment, however cost is potentially an issue for many countries.

Table 1: Summary of methods for detecting presence of *Mtb* in latent and active states

Test for	Test name	Pros	Cons
Latent TB	TST	Inexpensive	More false positives after BCG
Latent TB	IGRA	More expensive	Fewer false positives after BCG
Active TB	Smear microscopy	Cheap and simple	Requires specialists; low sensitivity/specificity
Active TB	Mycobacterial culture	Accurate	Requires specialists/safety conditions; slow
Active TB	Molecular diagnostic tests (XpertMTB/RIF assay)	Quick; detects RR-TB; high sensitivity/specificity; little specialist equipment required	Expensive

IGRA = Interferon-gamma release assay; RR-TB = Rifampicin resistant TB; TST = Tuberculin skin test

Treating even uncomplicated, drug-susceptible TB requires combination therapy over an extended time. The WHO recommends two months of rifampicin, isoniazid, pyrazinamide and ethambutol followed by four months of rifampicin and isoniazid [2]. However, MDR-TB and XDR-TB, requires a much longer treatment of at least 18 months, at increased cost and toxicity, thereby compromising compliance and outcomes [18], [19].

Determining drug resistant strains via drug susceptibility testing (DST) is therefore of high importance for TB management. Molecular methods such as the above mentioned

XpertMTB/RIF assay as well as line probe assays (LPA), alongside more traditional phenotypic tests can determine drug resistance, however increasingly whole genome sequencing (WGS) offers promising results.

Bioinformatics tools can rapidly determine strain-type and form of drug resistance directly from the *Mtb* genome, having been sequenced from PCR [20], [21], [22], and can even be conducted direct from sputum [23]. This approach contrasts with molecular assays in that the latter only tests for a limited number of loci. Again however, currently costs are often prohibitive for the highest-burden countries but amplicon assays that can sequence many samples and target genes have been developed to run on WGS platforms [24].

Prevention of disease

Administered to 100 million children per year, the BCG vaccine is currently the single prophylactic against TB [10]. Reflecting the global diversity, it is currently recommended to be given to all children where TB is common and only to high-risk children in low-burden countries.

Rates of protection vary widely however, and in children the BCG prevents ~20% of infections; once infected, the vaccine protects about half from developing disease [25]. As of 2018, fourteen candidate vaccines are in various phases of clinical trials [26].

Aside from the BCG vaccine, TB control is largely a question of clinical and environmental intervention. Indeed, the WHO characterised TB as a disease that is "intimately linked to poverty, and control of TB is ultimately a question of justice and human rights" [25]. Therefore, also of importance are epidemiological studies tracking the spread of TB, which includes the tracing of transmission clusters, including through identifying *Mtb* with almost identical genomes using whole genome sequencing (WGS) platforms.

MTBC strain diversity

The MTBC consists of *Mycobacterium tuberculosis sensu stricto* (*Mtb*) (lineages 1, 2, 3, 4 and 7) and *M. africanum* (lineages 5 and 6), which cause human disease, but others including *M. bovis* affect predominantly animals [27]. Recently, new *Mtb* lineages (8 and 9) have been proposed [28] [29]. The MTBC lineages vary in their geographic distribution and spread, being endemic in different locations around the globe, leading to the hypothesis that the strain-types are specifically adapted to different human populations [14]. Lineage 2 is particularly mobile with evidence of recent spread from Asia to Europe and Africa. Lineage 4 is common in Europe and southern Africa, with regions of high TB incidence and high levels of HIV co-infection, while lineages 5/6 and 7 appear isolated within West Africa and Ethiopia, respectively [27] (see **Figure 4**).

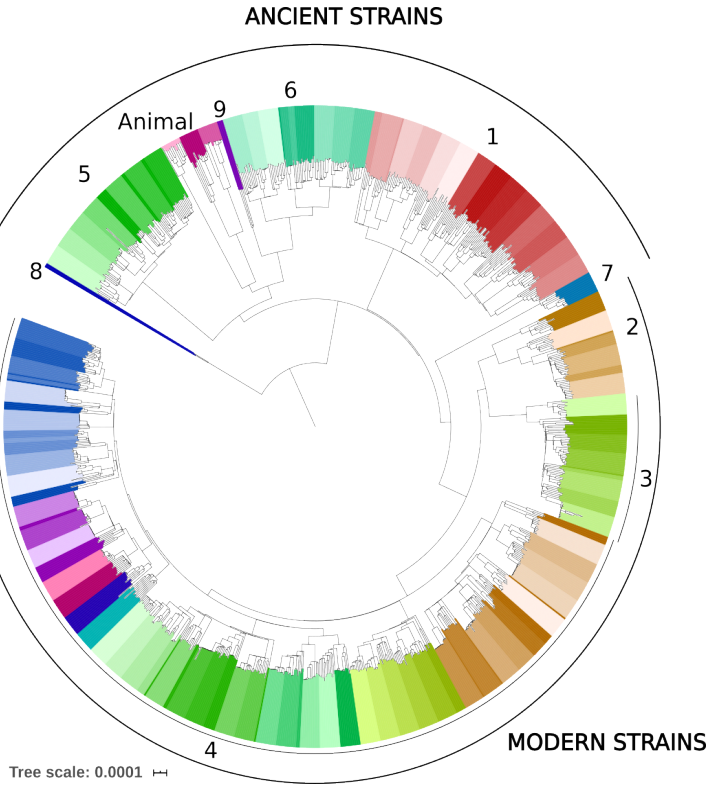


Figure 4: Phylogeny of all main MTBC lineages and sub-lineages (including newly proposed L8 and L9)

MTBC lineages can determine the transmission, control, and clinical outcome of pulmonary and extra-pulmonary TB. Variational phenotypes include differences in the emergence of drug resistance, transmissibility, virulence, host response, disease site and severity [30], [31]. Such phenotypes confer advantages for those MTBC lineages, leading to an increased likelihood of disease spread and poorer prognosis for patients. Of particular concern are the emergence of drug-resistant, MDR-TB, and XDR-TB strains, where Beijing strains show strong lineage-resistance associations [32]. However, there is considerable inter-strain variation within lineages. For example, when comparing two different Beijing sub-lineages, the "ancient" (atypical) and "modern" (typical) strains show differences in geographical distribution, drug resistance, and virulence patterns [5], [33]. In particular, the "modern" sub-lineage is distributed worldwide and has been largely associated with MDR-TB and XDR-TB and hypervirulence [5].

MTBC characteristics, genomics, sequencing and variation

MTBC bacteria do not fit into the classical Gram-positive and Gram-negative categories since their unusual cell wall is impervious to Gram staining [34]. Instead, acid-fast stains must be used. The cell wall consists of an innermost peptidoglycan layer, followed by an arabinogalactan layer, then a layer of lipid mycolic acid (unique to the *Mycobacterium* genus), and finally a lipomannan outer layer. This unusual structure contributes to properties such as resistance to desiccation and virulence [5] (see **Figure 5**).

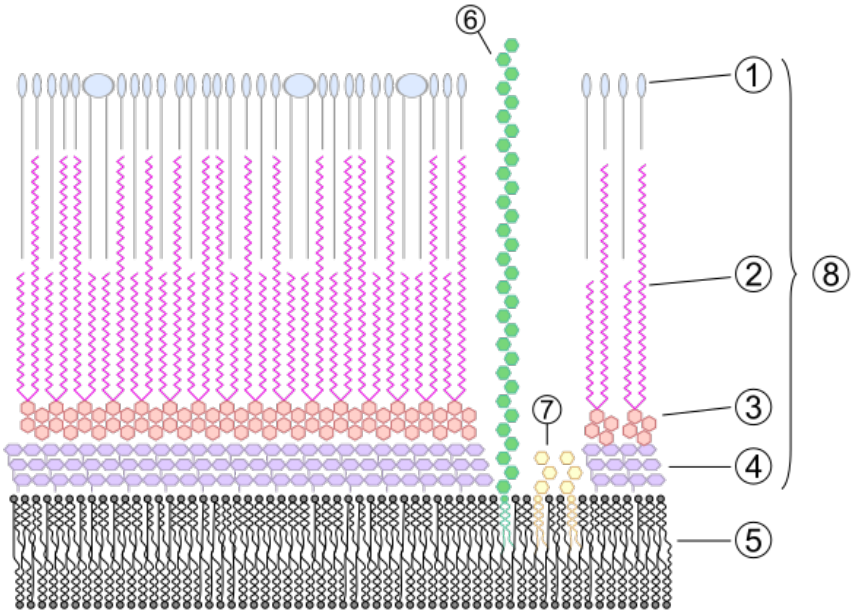


Figure 5: Typical mycobacterial cell wall and membrane: 1) outer lipids, 2) mycolic acid, 3) arabinogalactan, 4) peptidoglycan, 5) plasma membrane, 6) lipoarabinomannan (LAM), 7) phosphatidylinositol mannoside, 8) cell wall skeleton. From https://en.wikipedia.org/wiki/Mycobacterium

First published in 1998, the *Mtb* genome consists of ~4.4 million base pairs (4.4Mb), with 4,111 genes (and six pseudogenes) [35]. Despite intensive genomics research, less than half (40%) of the gene functions are known, though 44% of genes have hypothesised functions. Since there is no recombination and no horizontal gene transfer, and SNPs occur rarely, MTBC has been described as "genetically monomorphic" [36].

Nevertheless, the lack of recombination and low frequency of SNPs can be leveraged since they result in little convergence; many SNPs are unique to clades and can therefore be used as markers for phylogenies and for strain classification, as well as inferring divergence time, if mutation rate is known [36]. Indeed, strain-typing using SNPs is now a ubiquitous part of MTBC WGS pipelines [37]. For example, a minimum 62-SNP barcode has been identified that perfectly discriminates MTBC sub-lineages using a simple fixation-index score [38].

WGS approaches involve the complete DNA sequencing of an organism. These approaches are now quick, reliable, and affordable, and are increasingly used in multiple contexts to solve problems in MTBC control. The beginning of the WGS pipeline involves culturing sputum samples and extracting DNA, followed by library preparation and 'next-generation' short-read (e.g., *Illumina* platform [39]) or long read sequencing (e.g., Oxford Nanopore Technology MinION [40]). Once the raw data are collected, sequences are assessed for their quality, then typically mapped to a reference genome (often strain H37Rv) followed by detection of variants such as SNPs and insertions/deletions (indels) using, for example, *BCFtools* [41] or *GATK* [42]. Pipelines will usually exclude ~10% of the genome due to errors in mapping of certain regions such as *PE/PPE* genes and other repetitive genes which result in false variant calls. Refining the search for variants, criteria are applied statistically to the numerous variant properties, such as read depth, quality ('Phred') score, and numerous others which consider the context of the variant [37], [36]. For example, the GATK's *Variant Quality Score Recalibration (VQSR)* algorithm (https://gatk.broadinstitute.org/hc/en-us/articles/360035531612-Variant-Quality-Score-Recalibration-VQSR) can be applied for SNP-filtering (see **Figure 7**).

Once the raw data have been processed, questions specific to MTBC can be addressed. Despite MTBC being characterised as "genetically monomorphic", there is typically enough variation in the genome to discern genotype-phenotype relations and other biologically informative differences. Areas include diagnosis, genotypical DST, treatment, surveillance, identification of transmission clusters, and strain-typing [43], [44], [45] (see **Figure 9**).

Strain-typing

Using a simple 'fixation index' (F_{ST}) approach to compare allele frequencies between sub-populations, it was possible to identify SNPs uniquely associated to 62 MTBC lineages and sub-

lineages, providing a barcode for rapid strain-typing in the WGS context [27]. Such strain-typing moves on from analysis of partial genome sequencing such as spacer oligonucleotide typing (spoligotyping) and large deletions (Regions of Difference (RDs)) [46] which have the disadvantages of being susceptible to convergent evolution and provide less resolution. For example, spoligotyping looks at the presence or absence of a set of 43 tandem repeats across a small collection of regions. This means a barcode is limited to the binary status of just these regions, and while the number of possible combinations is huge, it is possible to find the same patterns of repeats in distant lineages. In contrast, the WGS approach provides greater resolution by considering the whole genome or genome-wide variants for strain-typing. A genome-wide SNP-based barcode has been implemented to cope with 'low-concentrated, low-quality DNA', potentially suitable for low-resource settings [47].

Transmission

Transmission is typically inferred if a genotypical assay reveals that samples are identical across the characterised markers. Two samples cannot be part of the same transmission event if they are phylogenetically distinct. Analogous to the situation with strain-typing, older methods such as spoligotyping may reliably inform on transmission events, however WGS-based genotyping offers greater temporal resolution. A comparison of clustering methods based on spoligotyping, 24-loci-MIRU-VNTR and WGS-SNP distances found that transmission events can be appropriately determined on time scales of 200 years, three decades and ten years after sampling, respectively [48]. This greater resolution time of WGS can be used effectively by epidemiologists wishing to track the phylodynamics of recent outbreaks.

Although transmission and clustering are usually applied in this epidemiological sense, an analysis of data from Pakistan seeks to infer transmission as a phenotype (or proxy phenotype

for virulence and transmissibility), and associate it with SNPs (**Chapter 4**). Nevertheless, the high temporal resolution of the WGS SNP-based approach will be relevant to establishing this phenotype.

Drug susceptibility testing (DST)

As with strain-typing and transmission, WGS is providing powerful ways in which to determine resistance efficiently and accurately to specific drugs. Given a list of drug resistance mutations which have been experimentally and statistically validated, new samples can be sequenced and analysed for the presence of variants known to be associated with a given drug. Thus, WGS can quickly provide predictions as to which drugs a sample may be resistant, informing clinicians as to most appropriate courses of treatment.

One interesting use of such resistance profiling is in detection of compensatory mechanisms. Resistance mutations often incur a fitness cost to the bacteria, so while a (pro-)drug can no longer bind to a target, the function of that target is compromised. Mutations can occur in other genes which code for similar proteins, compensating for such losses of function. For example, the pro-drug isoniazid binds to the KatG protein (coded by the *katG* gene). When this protein mutates, isoniazid fails to or partially binds and the bacteria is resistant to the pro-drug. However, KatG helps to protect the bacteria from attacks by the immune system, and so is left vulnerable after such mutations. To overcome this, mutations in the promoter region of the *ahpC* gene increase the production of the AhpC protein, which has a similar function to KatG. Not all mutations compromise fitness, and those that do not, such as Ser315Thr in *katG*, are often common since they leave *Mtb* drug resistant yet able to function almost normally. **Chapter 5** takes advantage of the presence of compensatory mutations to detect previously unknown resistance mutations in *katG* (isoniazid) and rpoB (rifampicin). A number of

compensatory and resistance mechanisms in important anti-TB drugs have been found (**Table 2**), including isoniazid and rifampicin (**Figure 6**).

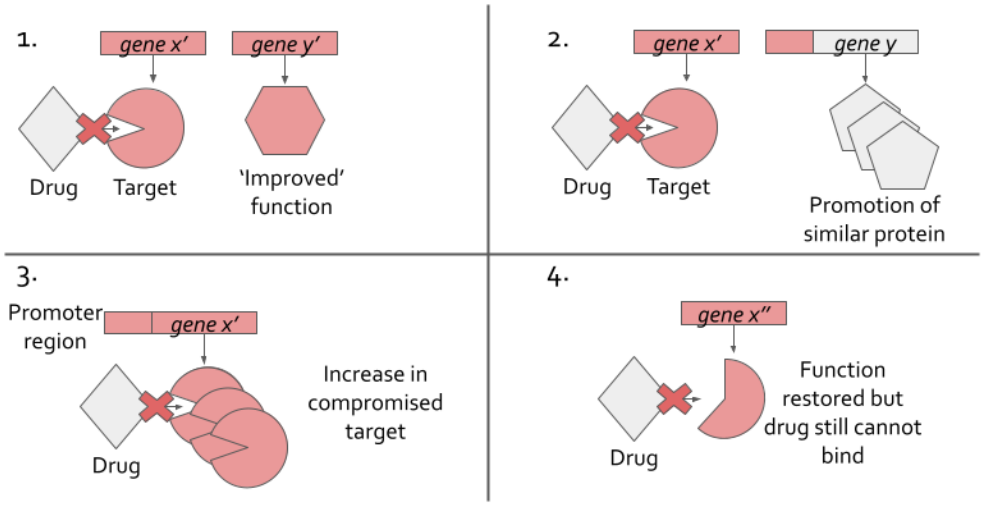
The advances of WGS are allowing for deeper research into MTBC, but are also creating some issues around the standardization of an MTBC WGS pipeline due to numerous methods which may have conflicting results [37]. A review of five automated tools intended to standardise the analysis and interpretation of WGS data (*CASTB*, *KvarQ*, *Mykrobe Predictor TB*, *PhyResSE,* and *TBProfiler*) concluded DST performance was 'highly variable' [49].

Table 2: Key resistance and compensatory mechanisms in anti-TB drugs. Adapted from Emane (2021) [50]

(Pro-)drug	Gene	R/C	Significant mutations	Effect on fitness	In highly transmitted strains?
Isoniazid	*katG*	R	Ser315Thr	minor	Very common
			Numerous	large	
	inhA promoter	R	Promoters: −8, −15, −17, or −47	unclear	Yes
	inhA structure	R	AA 94 or 194	unclear	Yes
	ahpC promoter	C	Promoter region	compensates for KatG protein	
Rifampicin	*rpoB*	R	Ser450Leu	minor	Very common
			Asp435Gly	severe	
			AA 445 & others	moderate	
	rpoC	C	Phe452Leu, Val483Gly & other	restores	Yes
	rpoA	C	Several sites	restores	Yes
	rpoB	C	Ile1106Thr & other	restores	Yes
Streptomycin	*rpsL*	R	Lys43Arg	minimal	Yes
			Lys43Thr/Asp	moderate	No
			Lys88Arg	minimal	Yes
			Lys88Glu	moderate	No
	16S rRNA (rrs)	R	512, 513, or 516	minimal	Sometimes
			514	severe	No
	gidB	R		unknown	
	rpsD; rpsE	C		partial	No
Injectables	16S rRNA	R	A1401G	minor	Yes
Kanamycin, Amikacin		R	C1402A	moderate	Maybe
Viomycin, Capreomycin		C	G1484U	compensatory	
	TlyA	C	methylates rrs 1402	compensatory	
	TlyA	R	loss of TlyA function		
Flouroquinolones	*gyrA*	R	AAs 89, 90, 91, & 94	minor	Yes
			Gly88Asp	minor	No
			Gly88Cys	severe	No
		R	*glgC* mutations	partial	in *M. smegmatis*

R = Resistance; C = Compensatory

A



B

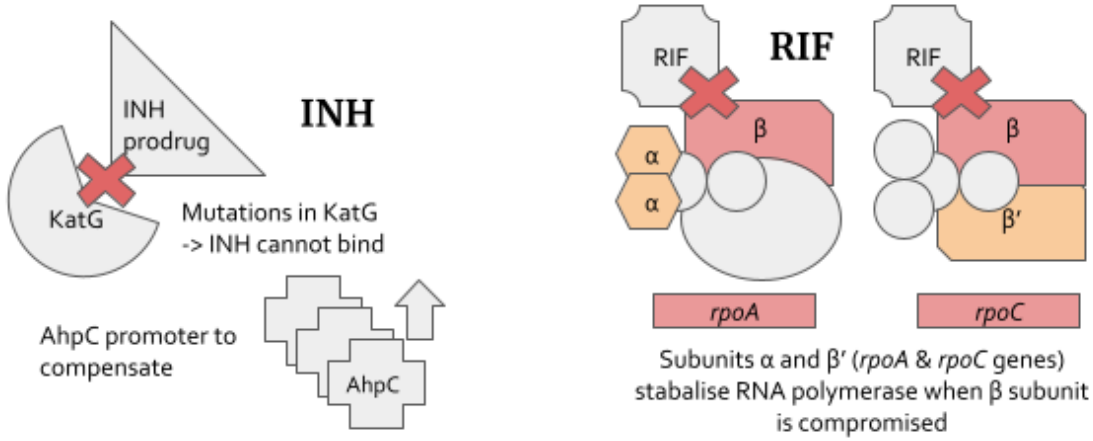


Figure 6: A: Mechanisms of compensatory mutations. 1. Mutation(s) in gene x (x') means the drug can no longer bind to the target. However the function of the target is compromised. Gene y' compensates by simultaneously improving another similarly functioning protein. 2. Here, rather than gene y itself mutating, the target is compensated by mutations in the promotor of gene y. 3. Similarly, the promotor region of gene x' can mutate, increasing levels of the compromised target protein. 4. A mutation in target gene x prevents the drug from binding and a second mutation in the same gene restores the function. B: Mechanisms of resistance and compensatory mutations in isoniazid (INH) and rifampicin (RIF).

GWAS and convergent evolution

Genome-wide association studies (GWAS) aim to identify phenotype-associated mutations by performing a statistical test between the phenotype and genotype for every variant site identified across a dataset. In performing bacterial GWAS, population structure must be accounted for since unintentional over-sampling of lineages (i.e., samples with very similar genotypes) can confound genotype-phenotype relationships [51].

One common method for dealing with this is to include and therefore adjust for the main principal components in regression models involving the phenotype outcome and the SNPs or genes of interest. The principal components summarise the population sub-structure and the regression-based method can be performed in software such as *Plink* [52]. An alternative approach is to use linear mixed models which include a random effect in the form of a between-sample kinship matrix, based on SNP similarities. Because the kinship matrix can represent fine-scale population structure patterns, the linear mixed models appear to be less prone to false positives [53]. This method has been applied to WGS of *Mtb* to analyse associations between mutations and drug resistance [54].

Another approach to find phenotype-associated mutations is to search for convergent evolution. Selective pressure (e.g., due to drug exposure) can force the same mutation to be acquired in different parts of the phylogenetic tree. This data, together with phenotypic data can be used to search for informative mutations. Software such as *PhyC* [55] and more recently *treeWAS* [56] account for population structure and test for significance of phylogenetic-wide mutations, inferring convergence and therefore positive selective pressure of mutations.

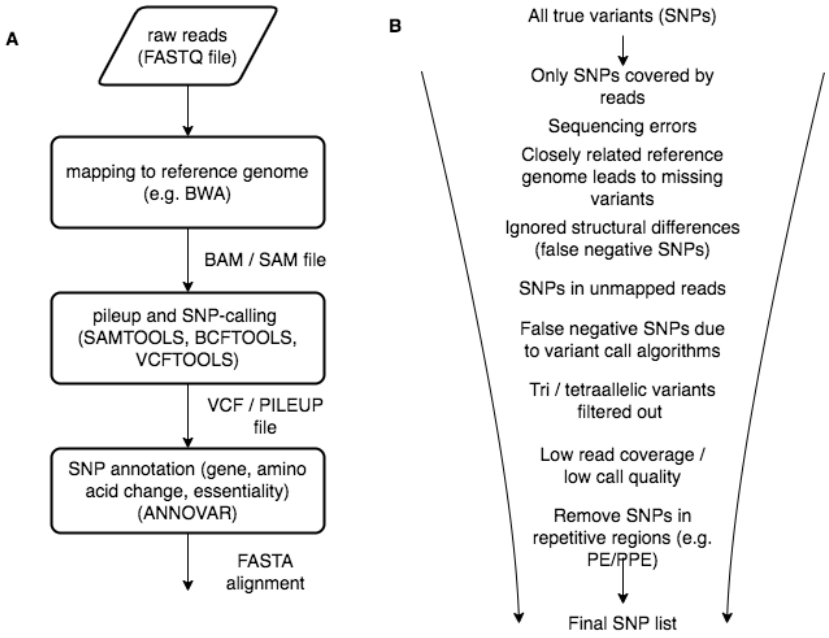


Figure 7: A: Example of sequencing analysis pipeline up to alignment stage. B: Criteria applied for SNP-filtering. Adapted from [36].

Dataset

The genomic data for 32,735 isolates were collected from publicly-available raw sequence (FAST-Q format) datasets. Metadata were collected simultaneously from the same source as the genomic data. **Table 3** and **Figure 8** show sample breakdown by lineage and WHO control region. Lineage 4 comprises the largest proportion of the dataset (51%), followed by lineage 2 (25.3%), lineage 3 (11.5%) and lineage 1 (9.64%), with the other lineages making up 2.53%. Most isolates are from the Northern Europe WHO region (**Table 3**, **Figure 8**). There are data on phenotypic resistance of a total of 23 drugs, these being comprised of 5 first line drugs, 11 second line drugs, 6 third line drugs and 1 drug not categorised in any of the first three (clofazimine) (**Table 4)**. These phenotypes are central to the drug resistance-genotype association analysis (**Chapters 3 and 4**).

Table 3: All samples (N=32,735) and their lineages (L)

Region	No. countries	L1	L2	L3	L4	L5	L6	L7	L9	Animal
Central Africa	6	4	3	7	475	23	1	0	0	4
East Africa	12	309	111	261	1,503	0	0	47	2	6
North Africa	5	0	0	0	57	0	0	0	0	0
Southern Africa	3	71	1,329	131	2,959	0	0	0	0	1
West Africa	11	2	3	0	111	184	95	0	0	0
Caribbean	2	0	0	0	5	0	0	0	0	0
Central America	4	2	4	2	91	0	0	0	0	157
North America	3	288	381	135	1,269	1	0	0	0	0
South America	6	7	90	1	1,254	0	0	0	0	5
Central Asia	3	0	251	3	52	0	0	0	0	0
East Asia	2	8	1,468	28	520	0	0	0	0	0
South Asia	7	252	157	562	163	0	0	0	0	0
Southeast Asia	7	1,170	2,003	11	480	0	0	0	0	0
Western Asia	5	5	7	10	26	0	0	0	0	2
Eastern Europe	9	1	829	3	606	0	0	0	0	2
Northern Europe	7	464	527	1,404	2,503	11	7	1	1	101
Southern Europe	8	5	122	13	658	1	3	0	0	0
Western Europe	6	252	408	530	2,013	19	13	0	0	76
Melanesia	1	0	63	0	6	0	0	0	0	0
Micronesia	1	0	4	0	0	0	0	0	0	0
Oceania	1	5	95	5	49	0	0	1	0	0
Unknown	1	310	413	668	1,909	14	29	3	0	19
Total	110	3,155	8,268	3,774	16,709	253	148	52	3	373

Table 4: Drug resistance (DR) phenotypes in all samples (N=32,735) and their lineages (L).

DR type	L1	L2	L3	L4	L5	L6	L7	L9	Animal	Total
Sensitive	2,337	3,638	2,614	11,548	170	129	49	2	45	20,532
Pre-MDR	390	928	385	1,281	21	9	0	0	0	3,014
MDR	232	1,417	351	2,127	41	5	0	1	0	4,174
Pre-XDR	34	1,164	168	826	3	0	0	0	1	2,196
XDR	6	589	62	367	0	0	0	0	2	1,026
Other	156	532	194	560	18	5	3	0	325	1,793
Total	3,155	8,268	3,774	16,709	253	148	52	3	373	32,735

MDR = Multi drug resistant; XDR = Extensively drug resistant

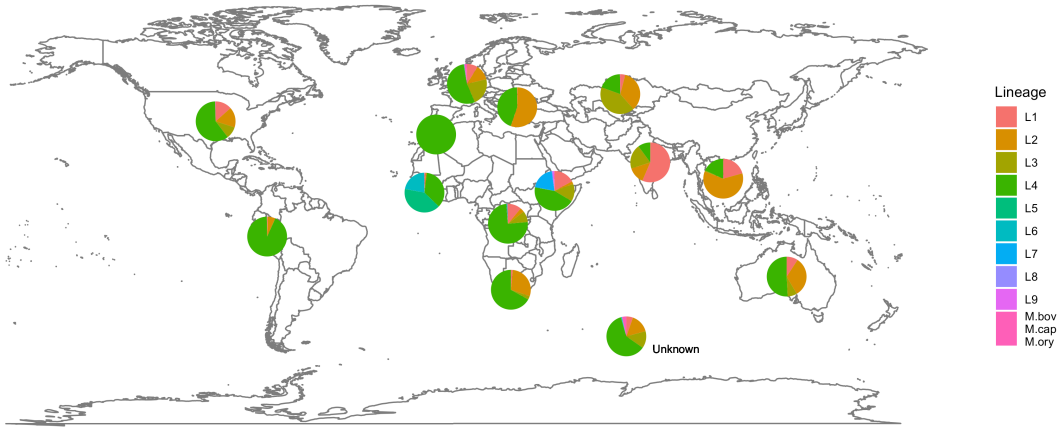


Figure 8: The global distribution of the 32,735 MTBC study isolates

The project structure

The theme of this thesis is to leverage WGS data of *Mtb* to make inferences about its phylogenetic structure and patterns of variation in the resistome. Ultimately, these inferences are valuable to the control of the disease in terms of rapid determination of lineage and drug resistance markers. Knowing information about an *Mtb* isolate's lineage helps determine phenotypic aspects, providing inferences about virulence, pathogenicity, transmissibility and disease outcome. For example L2 and L4 are thought to be more pathogenic and transmissible [57]. Knowing a sample's drug resistance genotype informs as to which drugs will be efficacious, thereby helping treatment outcomes. Knowing genomic resistance locations also benefits drug development since these can be mapped to the final protein structure, informing new drug design based on enriched knowledge of the targets.

A workflow of modern MTBC WGS analyses, from sample collection to profiling of drug resistance, clustering and strain-typing, leads to a coherent report about the specific *Mtb* isolate, as well as the data as a whole (**Figure 9**). Having built new pipelines appropriate to each problem, this thesis seeks to contribute to these main types of MTBC analyses, either through adding new empirical evidence of, say, new drug resistance mutations, or by refining the

framework by which future analyses can be conducted, for example by identifying new lineages to which reports can adhere. The thesis is divided into four central chapters addressing the following:

1. A 'molecular barcode' identifying SNPs specific to existing MTBC clades, as well as extensions to the lineage system.
2. New software updating *in silico* prediction of spoligotypes ('SpolPred2') and association of lineages in (1) to spoligotypes.
3. Characterisation of *Mtb* drug resistance and transmission in Pakistan.
4. Detection of new resistance mutations from the presence of compensatory mutations.

The corresponding research papers in this thesis include:

Chapter	**Title**	**Publication**
2	Robust barcoding and identification of *Mycobacterium tuberculosis* lineages for epidemiological and clinical studies	Napier et al, *Genome Medicine* 12: 114 (2020)
3	Comparison of *in silico* prediction of *Mycobacterium tuberculosis* spoligotypes and lineages from whole genome sequences	Napier et al, submitted, *Sci Rep*
4	Characterisation of drug-resistant *Mycobacterium tuberculosis* mutations and transmission in Pakistan	Napier et al, *Sci Rep* 12: 7703 (2021)
5	Large-scale genomic analysis of *Mycobacterium tuberculosis* reveals extent of target and compensatory mutations linked to isoniazid, rifampicin and multi-drug resistance	Napier et al, *Sci Rep* 13: 623 (2023)

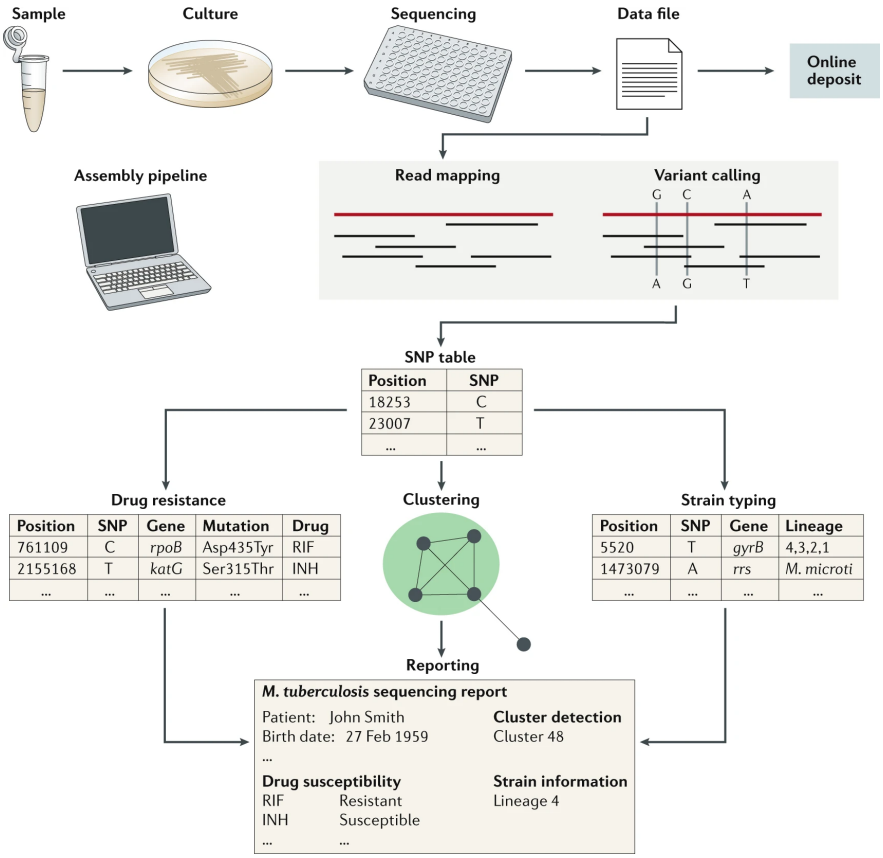


Figure 9: Standard pipeline for MTBC WGS bioinformatic analyses (adapted from [37])

Previous studies (e.g. [27]), have managed to produce comprehensive molecular barcodes for identifying *Mtb* lineages with single SNPs. Predicting a sample's position on the *Mtb* phylogenetic tree in such a manner requires a large dataset to accurately determine those SNPs as uniquely belonging to a clade, especially if the aim is to produce a fine-grained characterisation of clades. The technique used to arrive at such a barcode is a simple one, namely, to find SNPs that perfectly differentiate a (sub-)lineage (fixation index F_{ST} = 1), but it requires a sufficiently large dataset to eliminate spurious SNPs. Adding more data could reveal those F_{ST} = 1 SNPs to be present in other clades and therefore not unique. **Chapter 2** improves on existing barcodes, by analysing a significantly larger globals dataset (n=32k). By updating an already existing list of barcoding SNPs and extending the MTBC phylogenetic landscape with

many more samples and thereby many more SNPs, **Chapter 2** presents a barcode that is a much more accurate and fine-grained determination of MTBC phylogeny. The barcode can be used to position rapidly and cheaply with detection of single SNPs at specific locations, thus contributing to future profiling of strain-types without reconstructing a large tree from scratch.

Having obtained a comprehensive barcode for a hierarchical lineage system, one question that arose was how these lineages and levels relate to other strain-typing systems, especially spoligotypes. Spoligotypes can be readily predicted from WGS data *in silico*, and in the development of software ('SpolPred2'), the same isolate data used for **Chapter 2** could also be genotyped. At each level of the MTBC lineages (lineage 1, lineage 1.1, lineage 1.1.1 for example; from **Chapter 2**), and for each lineage at each level, all known spoligotypes were scored in much the same manner as the F_{ST} method, leading to scores ranging from 1 (exclusively found in a given lineage, at a given level), to anything less than 1 meaning the spoligotype was to be found in at least two lineages. The results reveal that although many individual spoligotypes were indeed predictive of lineage even at the lowest level (e.g., lineage 1.1.1.1), there was also much noise in the system, with low proportions of many lineage's samples having exclusive spoligotypes (score 1), and often a high proportion of samples with high-scoring but impure spoligotypes (score <1). The proportion of spoligotypes scoring 1 decreased markedly from the first lineage level to the lowest level, strongly suggesting that spoligotyping is in general too noisy to be usefully predictive of anything below the first level (lineages 1-7). In turn, these results highlight the strength of the lineage system proposed in **Chapter 2** as a finer resolution, with less noise.

From **Chapter 2**, with the construction of multiple trees, it became apparent that there were potential transmission clades (very closely related samples) in the same country. The potential

transmission clades in Pakistan, a high-burden TB country, had sufficient samples and with a rich diversity of drug-resistant samples. In **Chapter 4**, clustering, in the form of transmission analysis, revealed several closely related clades. To investigate any genetic influence on transmission, samples were divided into those 'transmitted' and 'non-transmitted', and a GWAS performed to find variants most closely associated with the two phenotypes. Also of interest was the drug resistance profile of MDR and (pre-)XDR samples. Multiple discrepancies were found between the phenotypic DST tests and the predicted genotypic resistance status, leading us to believe there were unknown resistance mutations, with further evidence suggesting their causal role beyond just association. These analyses contribute to informing infection control and clinical decision making in Pakistan, and potential other high-burden TB countries, with insights and implications for the mechanisms of drug resistance.

In finding candidate resistance markers in **Chapter 4**, particularly those in isoniazid and rifampicin, there emerged a pattern of association between compensatory mutations and known resistance mutations, as would be expected in any resistance profiling of these drugs. However, it also seemed that several samples had compensatory mutations but no known resistance mutations, yet had mutations in the drug target genes that could potentially explain resistance. It was decided that this pattern could be expanded in **Chapter 5** to the wider dataset to comprehensively find novel resistance mutations in isoniazid and rifampicin target genes, the two drugs with the best-established compensatory-resistance dynamic, and two of the most important first-line drugs. The large, global dataset (n=32k) was analysed for samples with this pattern of compensatory mutations and no known resistance mutation, but with mutations in target genes. The dataset was then re-scanned for samples with these candidate potential resistance mutations to find those that happened to have a potential resistance mutation but not having a compensatory mutation. **Chapter 5** presents a simple pipeline that can detect

potential new resistance mutations that could be extended to other drugs. Tracking where and how *Mtb* is mutating at the drug target sites contributes to rapid and more accurate determination of drug resistance in individual samples, in turn helping clinical decisions about treatment, as well as informing future drug design.

References

1. Daniel TM. The history of tuberculosis. Respiratory Medicine. 2006;100:1862–70.

2. World Health Organization. Global Tuberculosis Report 2021. 2021.

3. Abel L, El-Baghdadi J, Bousfiha AA, Casanova JL, Schurr E. Human genetics of tuberculosis: A long and winding road. 2014;369.

4. Hossain MM, Norazmi MN. Pattern recognition receptors and cytokines in Mycobacterium tuberculosis infection - The double-edged sword? 2013;2013.

5. Forrellad MA, Klepp LI, Gioffré A, García JS, Morbidoni HR, de la Paz Santangelo M, et al. Virulence factors of the mycobacterium tuberculosis complex. Virulence. 2013;4:3–66.

6. Barry CE, Boshoff HI, Dartois V, Dick T, Ehrt S, Flynn JA, et al. The spectrum of latent tuberculosis: Rethinking the biology and intervention strategies. 2009;7:845–55.

7. Dutta NK, Karakousis PC. Latent Tuberculosis Infection: Myths, Models, and Molecular Mechanisms. Microbiology and Molecular Biology Reviews. 2014;78:343–71.

8. Pai M, Behr MA, Dowdy D, Dheda K, Divangahi M, Boehme CC, et al. Tuberculosis. 2016;2:1–23.

9. Chai Q, Zhang Y, Liu CH. Mycobacterium tuberculosis: An adaptable pathogen associated with multiple human diseases. 2018;8:158.

10. Heemskerk D, Caws M, Marais B, Farrar J. Tuberculosis in Adults and Children. Springer; 2015.

11. Delogu G, Sali M, Fadda G. The biology of mycobacterium tuberculosis infection. 2013;5:2013070.

12. Shah M, Chida N. Extrapulmonary tuberculosis. In: Handbook of tuberculosis. Springer International Publishing; 2017. pp. 91–118.

13. Sulis G, Roggi A, Matteelli A, Raviglione MC. Tuberculosis: Epidemiology and control. 2014;6:2014070.

14. Brites D, Gagneux S. Co-evolution of Mycobacterium tuberculosis and Homo sapiens. Immunological Reviews. 2015;264:6–24.

15. Barnes PF, Cave MD. Molecular Epidemiology of Tuberculosis. New England Journal of Medicine. 2003;349:1149–56.

16. Zumla A, Raviglione M, Hafner R, Von Reyn CF. Tuberculosis. 2013;368:745–55.

17. Boehme CC, Nabeta P, Hillemann D, Nicol MP, Shenai S, Krapp F, et al. Rapid Molecular Detection of Tuberculosis and Rifampin Resistance. New England Journal of Medicine. 2010;363:1005–15.

18. Dheda K, Gumbo T, Gandhi NR, Murray M, Theron G, Udwadia Z, et al. Global control of tuberculosis: From extensively drug-resistant to untreatable tuberculosis. 2014;2:321–38.

19. Bastos ML, Hussain H, Weyer K, Garcia-Garcia L, Leimane V, Leung CC, et al. Treatment outcomes of patients with multidrug-resistant and extensively drug-resistant tuberculosis according to drug susceptibility testing to first- and second-line drugs: An individual patient data meta-analysis. Clinical Infectious Diseases. 2014;59:1364–74.

20. Coll F, McNerney R, Preston MD, Guerra-Assunção JA, Warry A, Hill-Cawthorne G, et al. Rapid determination of anti-tuberculosis drug resistance from whole-genome sequences. Genome Medicine. 2015;7:51.

21. Witney AA, Gould KA, Arnold A, Coleman D, Delgado R, Dhillon J, et al. Clinical application of whole-genome sequencing to inform treatment for multidrug-resistant tuberculosis cases. Journal of Clinical Microbiology. 2015;53:1473–83.

22. Phelan JE, Coll F, Bergval I, Anthony RM, Warren R, Sampson SL, et al. Recombination in pe/ppe genes contributes to genetic variation in Mycobacterium tuberculosis lineages. BMC Genomics. 2016;17:151.

23. Doyle RM, Burgess C, Williams R, Gorton R, Booth H, Brown J, et al. Direct Whole-Genome Sequencing of Sputum Accurately Identifies Drug-Resistant Mycobacterium tuberculosis Faster than MGIT Culture Sequencing. Journal of Clinical Microbiology. 2018;56:666–84.

24. Gómez-González PJ, Campino S, Phelan JE, Clark TG. Portable sequencing of Mycobacterium tuberculosis for clinical and epidemiological applications. Briefings in Bioinformatics. 2022. https://doi.org/10.1093/BIB/BBAC256.

25. World Health Organization. BCG vaccine: WHO position paper, February 2018 – Recommendations. 2018;36:3408–10.

26. Méndez-Samperio P. Development of tuberculosis vaccines in clinical trials: Current status. 2018;88.

27. 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. Nature Communications. 2014;5:4812.

28. Ngabonziza JCS, Loiseau C, Marceau M, Jouet A, Menardo F, Tzfadia O, et al. A sister lineage of the Mycobacterium tuberculosis complex discovered in the African Great Lakes region. Nature Communications. 2020;11:1–11.

29. Coscolla M, Brites D, Menardo F, Loiseau C, Darko Otchere I, Asante-Poku A, et al. Phylogenomics of Mycobacterium africanum reveals a new lineage and a complex evolutionary history. bioRxiv. 2020;17:19.

30. Ford CB, Shah RR, Maeda MK, Gagneux S, Murray MB, Cohen T, et al. Mycobacterium tuberculosis mutation rate estimates from different lineages predict substantial differences in the emergence of drug-resistant tuberculosis. Nature Genetics. 2013;45:784–90.

31. Reiling N, Homolka S, Walter K, Brandenburg J, Niwinski L, Ernst M, et al. Clade-Specific Virulence Patterns of Mycobacterium tuberculosis Complex Strains in Human Primary Macrophages and Aerogenically Infected Mice. mBio. 2013;4.

32. Oppong Y, Phelan J, Perdigão J, MacHado D, Miranda A, Portugal I, et al. Genome-wide analysis of Mycobacterium tuberculosis polymorphisms reveals lineage-specific associations with drug resistance. BMC Genomics. 2019;20.

33. Ribeiro SC, Gomes LL, Amaral EP, Andrade MR, Almeida FM, Rezende AL, et al. Mycobacterium tuberculosis strains of the modern sublineage of the beijing family are more likely to display increased virulence than strains of the ancient sublineage. Journal of Clinical Microbiology. 2014;52:2615–24.

34. Fu LM, Fu-Liu CS. Is Mycobacterium tuberculosis a closer relative to Gram-positive or Gram-negative bacterial pathogens? Tuberculosis. 2002;82:85–90.

35. Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, et al. Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. Nature. 1998;393:537–44.

36. Stucki D, Gagneux S. Single nucleotide polymorphisms in Mycobacterium tuberculosis and the need for a curated database. 2013;93:30–9.

37. Meehan CJ, Goig GA, Kohl TA, Verboven L, Dippenaar A, Ezewudo M, et al. Whole genome sequencing of Mycobacterium tuberculosis: current standards and open issues. 2019;17:533–45.

38. Coll F, Preston M, Guerra-Assunção JA, Hill-Cawthorn G, Harris D, Perdigão J, et al. PolyTB: a genomic variation map for Mycobacterium tuberculosis. Tuberculosis (Edinburgh, Scotland). 2014;94:346–54.

39. WHO. The use of next-generation sequencing technologies for the detection of mutations associated with drug resistance in Mycobacterium tuberculosis complex: technical guide. 2018;112 p.

40. Jain M, Olsen HE, Paten B, Akeson M. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. Genome Biology 2016 17:1. 2016;17:1–11.

41. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078–9.

42. Depristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature Genetics. 2011;43:491–501.

43. Allix-Béguec C, Arandjelovic I, Bi L, Beckert P, Bonnet M, Bradley P, et al. Prediction of Susceptibility to First-Line Tuberculosis Drugs by DNA Sequencing. New England Journal of Medicine. 2018;379:1403–15.

44. Satta G, Lipman M, Smith GP, Arnold C, Kon OM, McHugh TD. Mycobacterium tuberculosis and whole-genome sequencing: how close are we to unleashing its full potential? 2018;24:604–9.

45. Gardy JL, Johnston JC, Sui SJH, Cook VJ, Shah L, Brodkin E, et al. Whole-Genome Sequencing and Social-Network Analysis of a Tuberculosis Outbreak. New England Journal of Medicine. 2011;364:730–9.

46. Jagielski T, Ingen J van, Rastogi N, Dziadek J, Mazur PK, Bielecki J. Current Methods in the Molecular Typing of Mycobacterium tuberculosis and Other Mycobacteria. BioMed Research International. 2014;2014.

47. Cancino-Muñoz I, Gil-Brusola A, Torres-Puente M, Mariner-Llicer C, Dogba J, Akinseye V, et al. Development and application of affordable SNP typing approaches to genotype Mycobacterium tuberculosis complex strains in low and high burden countries. Scientific Reports. 2019;9:1–12.

48. Meehan CJ, Moris P, Kohl TA, Pečerska J, Akter S, Merker M, et al. The relationship between transmission time and clustering methods in Mycobacterium tuberculosis epidemiology. EBioMedicine. 2018;37:410–6.

49. Schleusener V, Köser CU, Beckert P, Niemann S, Feuerriegel S. Mycobacterium tuberculosis resistance prediction and lineage classification from genome sequencing: comparison of automated analysis tools. Scientific Reports. 2017;7:46327.

50. Alame Emane AK, Guo X, Takiff HE, Liu S. Drug resistance, fitness and compensatory mutations in Mycobacterium tuberculosis. Tuberculosis. 2021;129:102091.

51. Hoffman GE. Correcting for Population Structure and Kinship Using the Linear Mixed Model: Theory and Extensions. PLoS ONE. 2013;8.

52. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: A tool set for whole-genome association and population-based linkage analyses. American Journal of Human Genetics. 2007;81:559–75.

53. Hyun MK, Zaitlen NA, Wade CM, Kirby A, Heckerman D, Daly MJ, et al. Efficient control of population structure in model organism association mapping. Genetics. 2008;178:1709–23.

54. Phelan J, Coll F, McNerney R, Ascher DB, Pires DE, Furnham N, et al. Mycobacterium tuberculosis whole genome sequencing and protein structure modelling provides insights into anti-tuberculosis drug resistance. BMC Medicine. 2016;14:31.

55. Farhat MR, Shapiro BJ, Kieser KJ, Sultana R, Jacobson KR, Victor TC, et al. Genomic analysis identifies targets of convergent positive selection in drug-resistant Mycobacterium tuberculosis. Nature Genetics. 2013;45:1183–9.

56. Collins C, Didelot X. A phylogenetic method to perform genome-wide association studies in microbes that accounts for population structure and recombination. PLoS Computational Biology. 2018;14:e1005958.

57. Coscolla M, Gagneux S. Consequences of genomic diversity in mycobacterium tuberculosis. 2014;26:431–44.

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

London School of Hygiene & Tropical Medicine
Keppel Street, London WC1E 7HT

T: +44 (0)20 7299 4646
F: +44 (0)20 7299 4656
www.lshtm.ac.uk

RESEARCH PAPER COVER SHEET

Please note that a cover sheet must be completed for each research paper included within a thesis.

SECTION A – Student Details

Student ID Number	1807750	**Title**	Mr
First Name(s)	Gary		
Surname/Family Name	Napier		
Thesis Title	Using whole genome sequencing data to identify strain-types, transmission enhancers and novel drug resistance mutations of *Mycobacterium tuberculosis*		
Primary Supervisor	Prof. Taane G. Clark		

If the Research Paper has previously been published please complete Section B, if not please move to Section C.

SECTION B – Paper already published

Where was the work published?	Genome Medicine		
When was the work published?	14/12/2020		
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion			
Have you retained the copyright for the work?*	Yes	Was the work subject to academic peer review?	Yes

*If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	
Please list the paper's authors in the intended authorship order:	
Stage of publication	Choose an item.

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	I performed the bioinformatic and statistical analysis, and wrote the first draft of the manuscript. I worked with co-authors on subsequent drafts and finalisation of the paper.

SECTION E

Student Signature	
Date	30/09/2022

Supervisor Signature	
Date	22/09/2022

Chapter 2

Robust barcoding and identification of *Mycobacterium tuberculosis* lineages for epidemiological and clinical studies

Napier et al. Genome Medicine (2020) 12:114
https://doi.org/10.1186/s13073-020-00817-3

Genome Medicine

RESEARCH Open Access

Robust barcoding and identification of *Mycobacterium tuberculosis* lineages for epidemiological and clinical studies

Gary Napier[1], Susana Campino[1], Yared Merid[2,3,4], Markos Abebe[2], Yimtubezinash Woldeamanuel[3], Abraham Aseffa[2], Martin L. Hibberd[1], Jody Phelan[1] and Taane G. Clark[1,5*]

Abstract

Background: Tuberculosis, caused by bacteria in the *Mycobacterium tuberculosis* complex (MTBC), is a major global public health burden. Strain-specific genomic diversity in the known lineages of MTBC is an important factor in pathogenesis that may affect virulence, transmissibility, host response and emergence of drug resistance. Fast and accurate tracking of MTBC strains is therefore crucial for infection control, and our previous work developed a 62-single nucleotide polymorphism (SNP) barcode to inform on the phylogenetic identity of 7 human lineages and 64 sub-lineages.

Methods: To update this barcode, we analysed whole genome sequencing data from 35,298 MTBC isolates (~ 1 million SNPs) covering 9 main lineages and 3 similar animal-related species (*M. tuberculosis* var. *bovis*, *M. tuberculosis* var. *caprae* and *M. tuberculosis* var. *orygis*). The data was partitioned into training ($N = 17{,}903$, 50.7%) and test ($N = 17{,}395$, 49.3%) sets and were analysed using an integrated phylogenetic tree and population differentiation (F_{ST}) statistical approach.

Results: By constructing a phylogenetic tree on the training MTBC isolates, we characterised 90 lineages or sub-lineages or species, of which 30 are new, and identified 421 robust barcoding mutations, of which a minimal set of 90 was selected that included 20 markers from the 62-SNP barcode. The barcoding SNPs (90 and 421) discriminated perfectly the 86 MTBC isolate (sub-)lineages in the test set and could accurately reconstruct the clades across the combined 35k samples.

Conclusions: The validated 90 SNPs can be used for the rapid diagnosis and tracking of MTBC strains to assist public health surveillance and control. To facilitate this, the SNP markers have now been incorporated into the *TB-Profiler* informatics platform (https://github.com/jodyphelan/TBProfiler).

Keywords: Tuberculosis, Diagnostics, Profiling, SNPs, Barcoding, Mycobacteria tuberculosis complex

* Correspondence: taane.clark@lshtm.ac.uk
[1]Faculty of Infectious and Tropical Diseases, London School of Hygiene & Tropical Medicine, Keppel Street, London WC1E 7HT, UK
[5]Faculty of Epidemiology and Population Health, London School of Hygiene & Tropical Medicine, London WC1E 7HT, UK
Full list of author information is available at the end of the article

Background

Tuberculosis, caused by bacteria in the *Mycobacterium tuberculosis* complex (MTBC), is a major global burden causing approximately ten million active cases and killing 1.5 million people in 2018 (www.who.int/tb). The MTBC consists of *Mycobacterium tuberculosis* sensu stricto (*Mtb*) (lineages 1, 2, 3, 4 and 7) and *M. tuberculosis* var. *africanum* (lineages 5 and 6; *M. africanum*), which cause human disease, but others including *M. tuberculosis* var. *bovis* affect predominantly animals [1]. Recently, new *Mtb* lineages (8, 9) have been proposed [2, 3]. The MTBC lineages vary in their geographic distribution and spread, being endemic in different locations around the globe, leading to the hypothesis that the strain types are specifically adapted to different human populations [4]. Lineage 2 is particularly mobile with evidence of recent spread from Asia to Europe and Africa. Lineage 4 is common in Europe and southern Africa, with regions of high TB incidence and high levels of HIV co-infection, whilst lineages 5, 6 and 7 appear isolated within West Africa and Ethiopia, respectively [1].

There is some evidence to suggest that MTBC lineages can determine the transmission, control, and clinical outcome of pulmonary and extra-pulmonary tuberculosis. In particular, variational phenotypes include differences in the emergence of drug resistance, transmissibility, virulence, host response, disease site and severity [5, 6]. Such phenotypes confer advantages for those MTBC lineages and may lead to an increased likelihood of disease spread and poorer prognosis for patients. Whether increased virulence is associated with poorer prognosis is unclear, with some studies reporting increased mortality risk with strains thought to be less virulent [7]. Of particular concern are the emergence of drug-resistant, multidrug-resistant (MDR-TB) and extensively drug-resistant (XDR-TB) strains, where Beijing strains show strong linear-resistance associations [8]. However, there is considerable inter-strain variation within lineages. For example, when comparing two different Beijing sub-lineages, the "ancient" (atypical) and "modern" (typical) strains show differences in geographical distribution, drug resistance and virulence patterns [9]. In particular, the "modern" sub-lineage is distributed worldwide and has been largely associated with MDR-TB and XDR-TB and hypervirulence [9].

Tracking the spread of lineages is of great importance in tuberculosis research and control. Rapid lineage identification enables the analysis of phenotypic associations, informs on likely provenance and can assist in the prediction of potential future outbreaks. The molecular barcoding of lineages and sub-lineages can be used to classify clinical isolates to aid in the evaluation of tools to control the disease, including therapeutics and vaccines, whose effectiveness may vary by strain type [1, 5]. Historically, strain identification has involved the genotyping of tandem repeats (e.g. spoligotypes) and large deletions (regions of difference (RDs)) [10], but these approaches are being replaced by methods analysing data from whole genome sequencing (WGS) technologies. These approaches include in silico spoligotyping and RD detection, the characterisation of lineage-associated single nucleotide polymorphisms (SNPs) and higher resolution methods such as core genome MLST [11]. SNP-based approaches can be applied in silico or implemented within a laboratory typing assay [12, 13]. Although the SNP-defined lineages do not offer the same resolution as using the whole genome, they provide a valuable insight into the epidemiology of circulating strains. A 62-SNP barcode was developed using WGS data for 1601 MTBC isolates and was the first to position samples within clades of a global phylogeny of 7 human lineages and 64 sub-lineages, covering all common strain types [1].

Here, we update the 62-SNP barcode using WGS for 35,298 MTBC isolates. In particular, we use WGS data for 17,903 (50.7%) isolates to reconstruct a global phylogeny, resulting in 30 new (sub-)lineages. This analysis led to the 62-SNP barcode being modified and extended to ninety robust SNPs to cover 90 MTBC (sub-)lineages or species, including animal-related *M. tuberculosis* var. *bovis* (*M. bovis*), *M. tuberculosis* var. *caprae* (*M. caprae*) and *M. tuberculosis* var. *orygis* (*M. orygis*), which are similar and sometimes misclassified. The new barcode was validated on the 17,395 (49.3%) remaining MTBC isolates. The ninety SNP markers have been incorporated into the *TB-Profiler* software (https://github.com/jodyphelan/TBProfiler) [14], which has been used to profile more than fifty thousand MTBC for strain types and drug resistance, and will thereby assist with barcode implementation for research and infection control activities.

Methods

Sample, raw data and sequence analysis

Illumina whole genome sequencing data was publicly available across 35,298 MTBC isolates, which encompassed *Mtb* lineages (1, 2, 3, 4 and 7), *M. africanum* (lineages 5 and 6), *M. bovis*, *M. caprae* and *M. orygis* [14], and the recently proposed lineages 8 [2] and 9 [3] (Additional file 1: Table S1). The data were convenience sampled with the first processed set (n = 17,903; 50.7%) serving as a training dataset, and the second set collated subsequently (n = 17,395; 49.3%) serving as a testing dataset (Additional file 1: Table S1). The test set covers all the sub-lineages in the training set with at least 10 isolates (range 10–917), except (sub-)lineages 3.1.2.2, 4.6.2.1, 8 and 9, but for these the number of training samples is relatively small.

All raw sequences were trimmed using *trimmomatic* software [15] (v0.36, parameters: PE -phred33 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:20 MINLEN:

36). Trimmed reads were then aligned with *BWA-MEM* software [16] (v0.7.17-r1188, default parameters) using the H37Rv reference sequence (Genbank accession number: NC_000962.3). Alignments from *BWA-MEM* were converted to "bam" format and sorted using *samtools* software [17] (v1.9, default parameters). SNPs were identified by applying *BCFtools* [17] (v1.9, mpileup parameters: default, call parameters: -mv) and *GATK* software [18] (version: 4.1.3.0) using the HaplotypeCaller function (parameters: -ERC GVCF). Individual sample "vcf" files were merged using *GATK* GenomicsDBImport (default parameters) and *GATK* CombineGVCFs (default parameters) to perform joint calling using all samples. The resulting multi-sample vcf file was filtered to remove indels and heterozygous calls and monomorphic SNPs. A multi-FASTA file containing all isolates was generated from the filtered SNP file (N = 1,014,762 SNPs; training 620,652 SNPs; test 533,152 SNPs) and H37Rv reference genome using *bedtools* (v2.28.0) [19] and in-house python scripts. The regions of difference (RDs) were detected using *delly* software [20] and confirmed using de novo assembly by applying *Spades* software [21]. Spoligotypes were called using *spolpred* software [22].

Principal component analysis and phylogenetic tree

Distance matrices and the principal components of the multi-FASTA files were computed with *Plink* software (v1.90b4; https://www.cog-genomics.org/plink2) [23]. The distance matrices were used for the new cluster identification. Maximum likelihood phylogenetic trees were constructed from the multi-FASTA file using *IQ-TREE* (v1.6.12) (http://www.iqtree.org/) [24]. A general time reversible model with rate heterogeneity set to a discrete Gamma model and an ascertainment bias correction were used (parameters *-m GTR+G+ASC*), with 1000 bootstrap samples used to measure branch quality and robustness. Phylogenetic trees were generated for all MTBC isolates, as well as for each main lineage separately. The resulting Newick-formatted tree files were visualised and annotated with metadata in *iTOL* (v5.2; https://itol.embl.de/) [25]. These metadata included the 62-SNP barcode sub-lineage predictions [1], allowing for the rapid identification of outliers. By annotating the branches with ancestral mutations, it was possible to inform on SNP markers for barcoding.

Lineage revision and new sub-lineage identification

The visual inspection of the phylogenetic trees (and principal component analysis plots) revealed that some pre-existing (sub-)lineages (as defined using the 62-SNP barcode) could be merged or split, as well as new ones created. The original 62-SNP barcode was constructed to reflect the original strain-type families used by researchers based on spoligotypes and RDs. We sought to analyse the phylogenetic tree to further divide these clades where obvious splits in the phylogeny existed. To aid in old lineage revision and new lineage identification, phylogenetic trees relating to lineages 1 to 9 and animal strains were analysed using a semi-automated procedure. Each tree was traversed (and each clade inspected) from root to tip using the *ETE3 Toolkit* (v3.1.1) package in Python3 (http://etetoolkit.org/) [26]. We identified metrics and parameters such as branch bootstrap support values and intra/inter-cluster SNP distances to determine splits in the tree, which led to clusters that are separated by long branch lengths from other isolates. Whilst traversing, the following criteria had to be met to establish clades leading to new or revised sub-lineages: (1) a minimum clade size of 20, with a branch supported by a bootstrap value of > 95; (2) differences in the distributions of SNP distances where comparing the isolates within and outside the clade, using a Welch t test assuming unequal variances [27] ($P < 0.05$) and a Cohen's d effect size [28] ($d > 0.5$); (3) the ratio of the branch length of the clade compared to the mean branch length of its descendants (ratio > 1); (4) estimation of the number of clade-informative SNPs, requiring at least 10 SNPs with a fixation index (F_{ST}) [29] value of 1; (5) confirmation of the clade through visual inspection of the tree. Each of the parameter thresholds was based on established cut-offs or determined using standard point of inflection methods [1]. The population differentiation F_{ST} statistic assigns a strength of association between each SNP and (sub-)lineage, with a score of 1 indicating that the SNP allele is fixed in the sub-lineage of interest and not present outside that group. Using the five criteria led to the addition of 87 (27 new) sub-lineages or lineages (including 8 and 9), or changing the branch position of established others (e.g. 1.2 and 1.1.1.1) (see Additional file 1: Fig. S1). The *SNP-IT* tool for identifying species in MTBC [30] was applied to the *M. bovis*, *M. orygis* and *M. caprae* isolates (N = 110; test set), and three barcoding SNPs were required for these mycobacteria. The overall number of (sub-)lineages or species covered was 90.

Barcoding SNPs

To ensure that the required 90 clade-specific mutations ("potential barcoding SNPs", all with F_{ST} = 1) were robust, where possible, we retained synonymous SNPs in essential genes [31], and excluded those in drug resistance loci (from *TB-Profiler* [14]) and non-essential PE/PPE gene families [32]. From those retained "robust" SNPs (n = 421), a minimal set of one per lineage included preferentially those already present in the 62-SNP barcode [1] and, if not possible, (arbitrarily) the lowest position was chosen. The gene functional categories were extracted from *Tuberculist* (tuberculist.epfl.ch), and the frequency of

ontologies across all potential barcoding, robust and minimal SNPs, was assessed for differences across lineage using the chi-squared tests.

Validation of lineage barcode

To validate the final set of robust 421 clade-defining SNPs (Additional file 1: Table S2), the 17,395 samples in the testing set (with 572,021 SNPs) were used. The (sub-)lineage of these samples was predicted with *TB-Profiler* [14]. At the same time, a phylogenetic tree was reconstructed of the training and test samples together using *FastTree2* software [33]. To assess the sensitivity and specificity of the predictions, this tree was traversed in the *ETE3 Toolkit*, and test samples were examined for their presence in the clades defined by the training dataset.

Results

MTBC isolates, SNPs and phylogeny

Across a total of 35,298 MTBC isolates with sequencing data, we identified 1,014,762 high-quality SNPs. The isolates represented all MTBC lineages (1–9), *M. bovis*, *M. orygis* and *M. caprae*, but the majority were from lineages 4 (51.6%), 2 (25.2%), 3 (11.1%) and 1 (9.5%), with the frequency of others being at most 1% (Additional file 1: Table S1). Whilst it is a convenience set of sampled isolates, the geographical distribution of the lineages was as expected, with lineage 2 dominating in Southeast Asia, lineages 1 and 3 predominant in South Asia, lineage 4 abundant in Europe, Americas and Africa and lineages 5 and 6 present in West Africa (Fig. 1). The East Asian lineage 2 had the highest frequency of MDR-TB isolates (36.2%), driven by a higher prevalence in the Beijing sub-lineage (lineage 2.2; 36.5%) compared to the Manu ancestor or proto-Beijing strain type (lineage 2.1, 19.8%) (Table 1).

The 35k isolates were split into training (N = 17,903, 50.7%; all MTBC; 620,652 SNPs) and test (N = 17,395, 49.3%, all MTBC except lineages 8 and 9; 572,021 SNPs) datasets (Table 1; Additional file 1: Table S1). A phylogenetic tree was constructed on the training isolates and confirmed the clustering by lineage and sub-lineages (Fig. 2). Similarly, a principal component analysis of the 35k isolates using the ~ 1 million SNPs revealed the expected clustering by lineage or species (Additional file 1: Fig. S1(a)). Phylogenetic trees were constructed for each lineage separately and confirmed the sub-lineage and strain-type clustering (Additional file 1: Fig. S1(b)-(f)). However, by assessing the fine-scale clustering of sub-lineages predicted by the 62-SNP barcode, outlying samples were revealed and suggested a need for the repositioning of mutations underlying the clades or, alternatively, the creation of new sub-lineages that were on long branches (Additional file 1: Fig. S12(b, c)). In some cases, new sub-lineages reflected existing RD- or spoligotype-based strain classifications which were imperfectly or not captured using the 62-SNP barcode (see Additional file 1: Fig. S2 (d,e)).

Barcoding SNPs

By traversing the whole MTBC and lineage-based phylogenetic trees using a semi-automated algorithm, it was possible to modify sub-lineages within the flexible nomenclature structure of the previous barcode [1], as well as define clade-informative SNPs. The phylogenetic analyses characterised 27 additional (sub-)lineages covering lineages 1 (8), 3 (2), 4 (15), 8 (1) and 9 (1). The final number of (sub-)lineages in *Mtb* was 85 (L (ineage)1 16, L2 7, L3 7, L4 52, L7 1, L8 1, L9 1) and *M. africanum* was 2 (L5 1, L6 1) (Table 1; Fig. 2), requiring 87 SNP markers. A further three SNP markers were required to

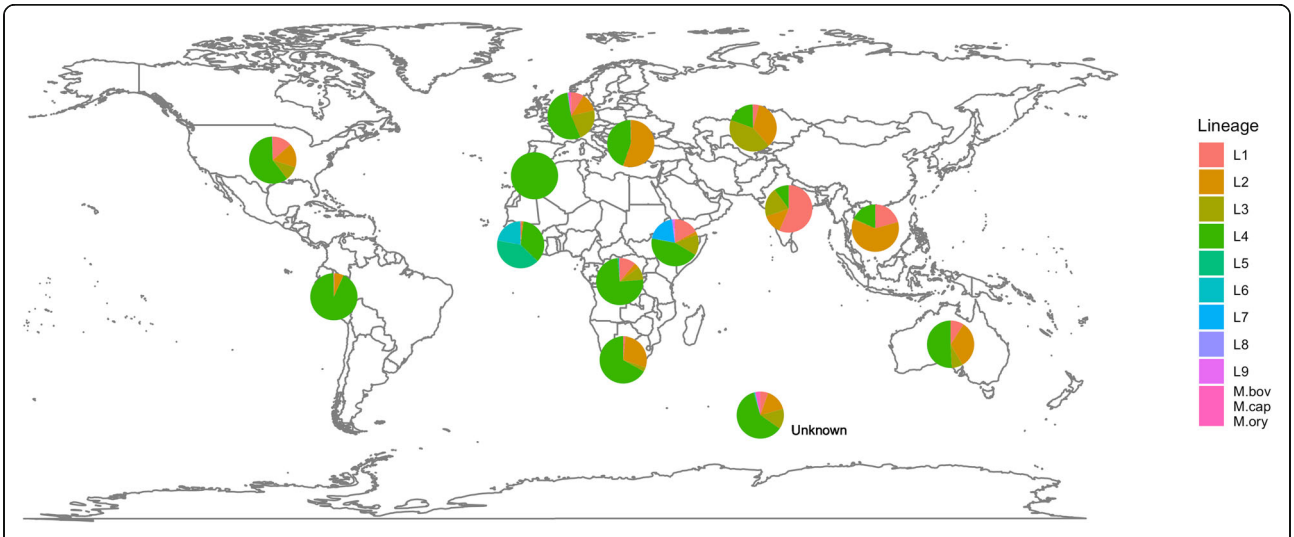


Fig. 1 The global distribution of the 35,298 *Mycobacterium tuberculosis* complex study isolates

Table 1 *Mycobacterium tuberculosis* complex lineages and sub-lineages across the 35,298 isolates

Lineage	No. training (test)	No. countries train (test)	% MDR-TB	No. transmission [clusters]	Potential barcoding SNPs*	Robust SNPs**
1	2162 (1203)	25 (42)	7.8	354 [130]	344	17
1.1	1487 (530)	19 (36)	5.5	218 [82]	23	2
1.1.1	706 (170)	8 (16)	3.8	60 [25]	41	5
1.1.1.1	358 (120)	5 (9)	2.1	28 [11]	52	3
1.1.2	459 (278)	15 (25)	9.0	83 [31]	109	3
1.1.3	299 (80)	11 (16)	3.2	73 [25]	42	2
1.1.3.1	84 (31)	7 (13)	3.5	10 [4]	68	2
1.1.3.2	155 (33)	7 (7)	1.1	57 [18]	113	6
1.1.3.3	32 (7)	5 (4)	10.3	4 [2]	36	2
1.2	309 (550)	13 (21)	7.5	40 [16]	60	2
1.2.1	28 (44)	3 (7)	6.9	6 [2]	78	5
1.2.2	277 (505)	13 (18)	7.5	34 [14]	159	8
1.2.2.1	244 (453)	12 (18)	6.9	34 [14]	34	1
1.3	366 (122)	16 (19)	18.0	96 [32]	71	2
1.3.1	88 (25)	7 (11)	10.6	20 [7]	50	4
1.3.2	278 (97)	16 (17)	20.3	76 [25]	83	4
2	4556 (4322)	45 (56)	36.2	1778 [413]	72	4
2.1	95 (41)	6 (9)	19.8	27 [10]	172	4
2.2	4461 (4281)	45 (56)	36.5	1751 [403]	79	17
2.2.1	4239 (4007)	45 (56)	35.1	1632 [389]	17	2
2.2.1.1	338 (443)	19 (18)	28.0	98 [40]	6	2
2.2.1.2	29 (21)	6 (9)	36.0	10 [3]	5	1
2.2.2	222 (273)	16 (15)	59.0	119 [14]	54	4
3	2654 (1271)	24 (31)	13.4	847 [242]	166	8
3.1	715 (362)	15 (22)	9.5	372 [80]	1	1
3.1.1	387 (280)	11 (16)	6.2	243 [43]	17	2
3.1.2	295 (69)	13 (8)	14.3	124 [35]	8	2
3.1.2.1	98 (25)	8 (7)	19.5	25 [12]	15	7
3.1.2.2	48 (0)	3 (0)	0	36 [2]	85	6
3.2	89 (31)	6 (9)	10.0	31 [7]	85	2
4	8320 (9883)	44 (99)	18.5	3109 [731]	94	3
4.1	2594 (2325)	35 (64)	18.5	1043 [191]	58	3
4.1.1	889 (482)	20 (27)	18.1	403 [72]	30	13
4.1.1.1	210 (158)	14 (16)	9.5	92 [20]	39	2
4.1.1.2	55 (44)	4 (6)	2.0	33 [3]	92	2
4.1.1.3	579 (247)	18 (23)	22.4	266 [44]	58	3
4.1.1.3.1	207 (13)	3 (3)	9.6	158 [5]	46	3
4.1.2	1612 (1743)	32 (61)	17.3	622 [113]	13	1
4.1.2.1	1383 (1087)	32 (60)	22.5	563 [96]	49	3
4.1.2.1.1	231 (18)	1 (1)	97.6	221 [2]	73	3
4.1.3	28 (70)	7 (10)	57.1	4 [2]	124	3
4.1.4	24 (12)	8 (7)	38.9	10 [2]	60	4
4.2	481 (532)	23 (26)	28.0	87 [32]	116	8

Table 1 *Mycobacterium tuberculosis* complex lineages and sub-lineages across the 35,298 isolates *(Continued)*

Lineage	No. training (test)	No. countries train (test)	% MDR-TB	No. transmission [clusters]	Potential barcoding SNPs*	Robust SNPs**
4.2.1	206 (240)	13 (20)	28.3	34 [13]	26	2
4.2.1.1	54 (148)	9 (10)	6.9	2 [1]	36	2
4.2.2	274 (288)	20 (18)	28.1	53 [19]	20	2
4.2.2.1	74 (41)	10 (6)	45.2	22 [7]	26	2
4.2.2.2	120 (139)	11 (14)	27.8	15 [7]	31	10
4.3	2507 (2928)	30 (75)	23.1	993 [244]	38	2
4.3.1	58 (67)	7 (15)	6.4	40 [3]	28	1
4.3.1.1	37 (2)	3 (1)	0.0	36 [1]	52	2
4.3.2	409 (1200)	16 (21)	7.2	75 [32]	75	1
4.3.2.1	291 (917)	6 (7)	3.7	50 [23]	55	4
4.3.3	648 (810)	25 (57)	41.3	210 [66]	33	1
4.3.4	1366 (807)	23 (45)	24.1	664 [142]	8	1
4.3.4.1	194 (170)	14 (30)	28.9	49 [14]	19	4
4.3.4.2	1170 (635)	22 (34)	23.1	614 [128]	26	1
4.3.4.2.1	877 (287)	13 (18)	5.6	457 [103]	11	1
4.4	560 (1059)	24 (29)	15.7	190 [63]	37	2
4.4.1	420 (861)	22 (25)	16.0	149 [48]	38	4
4.4.1.1	379 (755)	21 (24)	17.8	136 [44]	16	1
4.4.1.1.1	75 (206)	5 (4)	19.6	22 [9]	60	3
4.4.1.2	39 (106)	8 (6)	1.4	13 [4]	95	9
4.4.2	112 (181)	7 (9)	14.7	33 [13]	7	2
4.5	293 (357)	17 (17)	15.7	49 [22]	50	1
4.6	340 (442)	21 (25)	22.1	139 [39]	12	1
4.6.1	73 (296)	9 (12)	29.8	24 [8]	53	3
4.6.1.1	29 (126)	6 (7)	1.3	14 [3]	22	1
4.6.1.2	40 (154)	9 (11)	54.6	10 [5]	37	1
4.6.2	164 (89)	16 (17)	15.4	65 [20]	22	1
4.6.2.1	2 (0)	1 (0)	0	2 [1]	45	2
4.6.2.2	150 (89)	14 (17)	15.9	60 [18]	106	6
4.6.3	23 (9)	3 (4)	0	20 [3]	135	3
4.6.4	23 (7)	5 (4)	50.0	10 [2]	49	1
4.6.5	23 (18)	5 (5)	19.5	9 [3]	8	2
4.7	158 (200)	18 (23)	10.3	56 [20]	10	3
4.8	1051 (1807)	29 (55)	7.8	419 [88]	17	1
4.8.1	63 (90)	7 (4)	22.2	21 [5]	46	3
4.8.2	116 (5)	3 (2)	0	113 [1]	42	2
4.8.3	21 (3)	1 (1)	0	19 [1]	34	1
4.9	243 (141)	14 (22)	12.5	114 [24]	37	3
4.9.1	74 (15)	6 (3)	5.6	44 [1]	49	3
5	26 (255)	6 (12)	14.6	2 [1]	460	13
6	32 (135)	6 (13)	3.6	5 [2]	214	10
7	38 (26)	3 (2)	0	3 [1]	837	38
8	2 (0)	1 (0)	0	0 [0]	888	43

Table 1 *Mycobacterium tuberculosis* complex lineages and sub-lineages across the 35,298 isolates *(Continued)*

Lineage	No. training (test)	No. countries train (test)	% MDR-TB	No. transmission [clusters]	Potential barcoding SNPs*	Robust SNPs**
9	3 (0)	1 (0)	0	0 [0]	160	5
M. bovis	81 (281)	9 (12)	0.8	42 [11]	93	3
M. caprae	3 (7)	2 (3)	0	0 [0]	225	5
M. orygis	26 (12)	4 (4)	0	0 [0]	743	28
Totals	17,903 (17,395)	165 (269)	21.0	6140 [1531]	8128	421

Bolded are changes from the barcode in reference [1]—either new sub-lineages or new barcoding SNPs; MDR-TB multidrug-resistant TB, which is resistant to at least rifampicin and isoniazid drugs. *All potential barcoding SNPs ($F_{ST} = 1$). **Final robust SNP set, based on synonymous changes in essential and non-drug resistance genes only (except 12 sub-lineages which had no informative SNPs in essential genes; see Additional file 1: Table S2)

discriminate *M. bovis*, *M. caprae* and *M. orygis*, which have highly similar mycobacterial genomes, and therefore, their accurate typing will greatly assist with the misclassification of *M. bovis* infections.

To find informative SNPs for each of the 90 MTBC clades, we used the population differentiation metric F_{ST} to identify mutations that were only present in the isolates in the selected (sub-)lineage of interest ($F_{ST} = 1$). We identified 8128 potential barcoding SNPs (with $F_{ST} = 1$) across the 90 clades (Table 1). These barcoding SNPs were distributed evenly genome-wide, with no visible clustering of informative mutations for individual lineages (Additional file 1: Fig. S3). Of these SNPs, 7282 (89.6%) were in genic regions, with mutations leading to 4699 non-synonymous (NS) and 2564 synonymous (S) amino acid changes, as well as 20 changes in non-coding genes. By focusing on essential genes, 889 (10.9%) SNPs remained (499 NS, 390 S). Furthermore, variants in

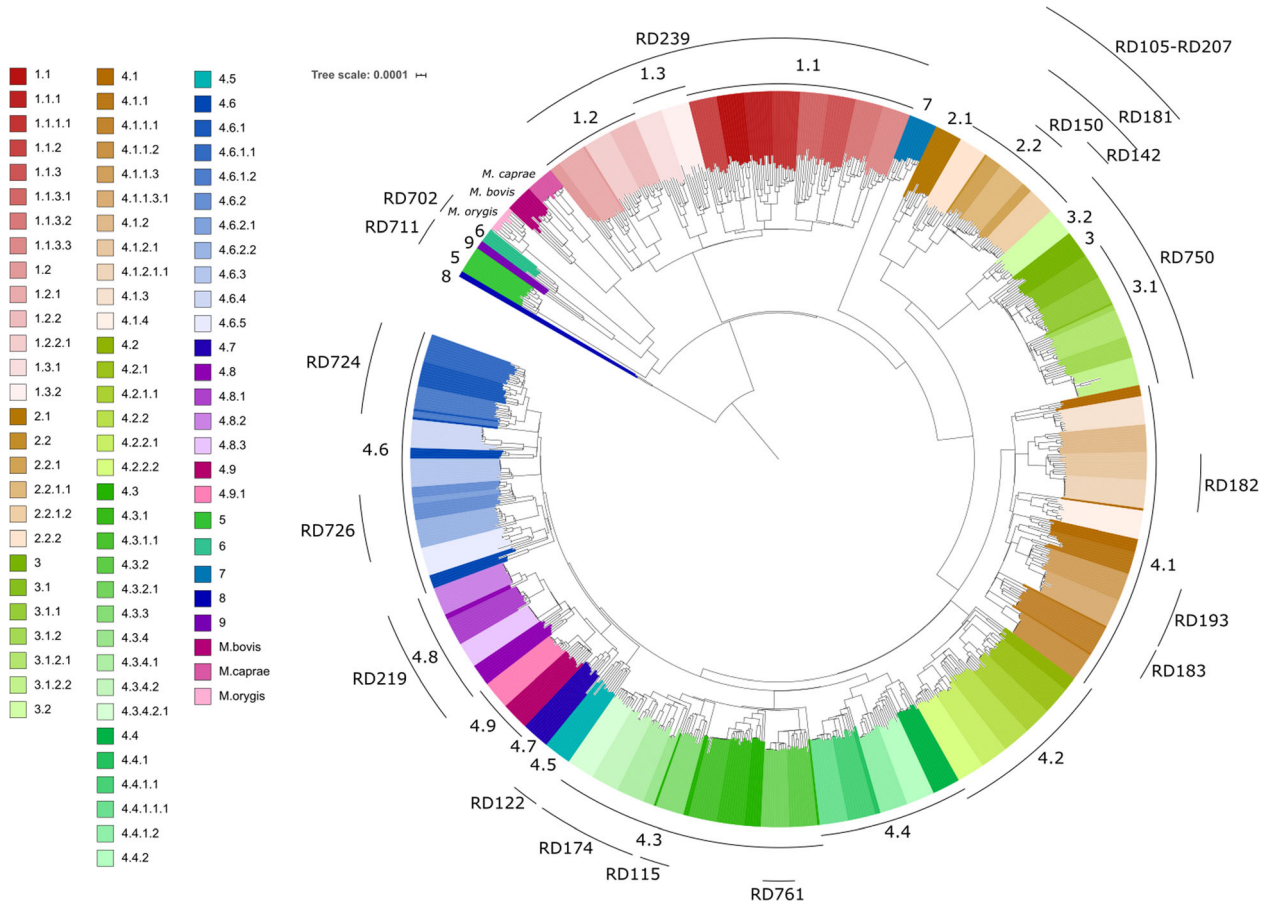


Fig. 2 Phylogenetic tree of *Mycobacterium tuberculosis* complex isolates. A representative tree with a maximum of 10 isolates per sub-lineage (important regions of difference (RDs) are also highlighted)

drug-resistance-associated genes were removed, leaving 824 SNPs (464 NS and 360 S mutations). Across all lineages, except lineages 8 ($N = 2$) and 9 ($N = 3$) which had small sample sizes, we compared the distribution of gene functions for all potential barcoding SNPs in all characterised genes (7060/7282 SNPs) with only those in essential (and non-drug resistance) loci (790/824 SNPs) (Additional file 1: Fig. S4). The distribution of gene function for all potential barcoding SNPs is similar across all lineages. However, after filtering for essential and non-drug-resistant genes, lineage 2 has a relatively high proportion of non-synonymous SNP mutations in cell wall and cell process genes, whilst for lineage 6, *M. bovis*, *M. caprae* and *M. orygis*, there are relatively higher proportions of non-synonymous SNP mutations in intermediary metabolism and pathway genes. For 11 (sub-)lineages, there were no potential barcoding SNPs lying within essential and non-drug resistance genes, so they were identified in non-essential and non-PE/PPE loci (Additional file 1: Table S3) (180 SNPs, 61 synonymous mutations).

By considering only the SNPs with synonymous changes, similar to the selection strategy applied in [1], a total of 421 SNPs were considered suitable for barcoding the 90 (sub-)lineages (Table 1; Additional file 1: Table S2). Of these, 20 SNPs represented (sub-)lineages in the 62-SNP barcode [1] and were therefore retained, leading to 70 new SNPs chosen for final (sub-)lineage classification (Additional file 1: Table S3). Across the 60 (sub-)lineages common to the 62- and 90-SNP barcodes, the 40 new SNPs had higher F_{ST} values than those in the old barcode (Additional file 1: Fig. S5). Using the test set ($N = 17{,}395$) which had representation of 86 of the 90 (sub-)lineages, we found that the minimal set of 90 SNPs had perfect predictive performance for all clades (all sensitivities and specificities of value 1). This analysis excluded four (sub-)lineages (3.1.2.2, 4.6.2.1, 8 and 9), which had no test samples.

Comparisons to other software

The barcode was compared to lineage predictions from SNP-IT [30] software, a 27 strain-type system covering MTBC, including 6 animal lineages that are not present in our large dataset. First, we assessed the assigned major MTBC lineages (1–6) by both barcodes and found complete concordance. Second, we quantified how the increased number of strain types in our barcode ($n = 90$) improved the resolution of sub-lineage assignment over the SNP-IT tool. For 14 of the 21 SNP-IT strain types present in our data, the 90-SNP approach provides higher resolution of clades (range 2 to 15 sub-lineages per SNP-IT clade) (Additional file 1: Fig. S6). Six other strain types have direct mapping between our barcode and SNP-IT, and there is one instance where isolates classified as *M. bovis* with our barcode are further classified into *M. bovis BCG* and *M. bovis bovis* using SNP-IT.

Discussion

MTBC strain types and lineages are distributed phylogeographically and have been associated with differences in the emergence of drug resistance, transmissibility, virulence, host response, vaccine efficacy, disease site and severity [5, 6, 34]. However, further research into lineage, genotype–phenotype associations are required. Such research needs to be underpinned by molecular barcodes of MTBC (sub-)lineages, strain types and species. Here, we updated a 62-SNP barcode that forms a highly resolved phylogenetic identification system that determines 7 lineages, 64 sub-lineages and *M. bovis*, but was constructed using ~ 1600 MTBC isolates with WGS data [1]. Using twenty-fold more MTBC isolates with WGS data, we identified and validated a set of 90 robust SNPs (of 421 alternatives) to cover a global phylogeny of 9 lineages, 87 sub-lineages, *M. bovis*, *M. caprae* and *M. orygis*. These SNPs can be used to construct high-resolution and reproducible phylogenies, which can be incorporated within diagnostic assays and assess genotype–phenotype associations. By extending an established 62-SNP barcode system with a flexible nomenclature [1], it was possible to update and add seamlessly (sub-)lineages and species and in the future include potentially novel strain types should they be reported. Such modifications could involve inclusion of SNPs to barcode other MTBC animal lineages or partitioning of *M. africanum* lineages 5 and 6 into sub-lineages [3]. Further, incorporating drug resistance loci will further enhance the usefulness of the 90-SNP barcode as an important tool for tuberculosis control and elimination activities worldwide. To assist this, the 90-SNP variants have been incorporated into the publicly available *TB-Profiler* informatics tool [14], which predicts resistance to 14 anti-tuberculosis drugs from WGS data.

Our barcode development focused on SNPs, but future work could include other types of strain-specific polymorphisms (e.g. insertions, deletions and large structural variants), which are less common than SNPs, but may have major functional consequences. An analysis of the gene ontologies of the barcoding SNPs revealed some differences across lineages, but there is a need to the characterise functional effects of the lineage-specific SNP variants, as these could provide insights into disease control measures. Overall, we have provided an updated molecular barcode for MTBC strain types, with ninety robust markers that can be detected from applications of WGS or integrated within high-throughput genotyping or sequencing (e.g. amplicon) platforms to inform ongoing TB surveillance and control.

Conclusions

The use of molecular barcoding of MTBC bacteria causing tuberculosis can provide insights into outbreaks and help to reveal strain types that are more virulent and prone to drug resistance. In an analysis of 35,298 isolates from MTBC, we update an established 62-SNP barcode with a minimal set of 90 genetic markers, which now cover *M. tuberculosis* (7 lineages, 85 sub-lineages), *M. africanum* (2 lineages), *M. bovis*, *M. caprae* and *M. orygis* bacteria. The new barcode has been implemented within the publicly available *TB-Profiler* informatics tool, to assist the rapid, simple and reliable phylogenetic identification of individual MTBC isolates, thereby aiding clinical studies in the tracking, maintenance and phenotypic determination of MTBC pathogens.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13073-020-00817-3.

Additional file 1: Table S1. The study samples (*N*=35,298) used and their lineages. **Table S2.** Robust barcoding SNPs (421 SNPs, including the 90 SNPs in **Table S3**). **Table S3.** The ninety minimal barcoding SNPs. **Figure S1.** Population structure of the *Mycobacterium tuberculosis* complex isolates by lineage. **Figure S2.** Examples of discrepancies using the 62-SNP barcode**. Figure S3.** The genome-wide distribution of barcoding SNPs (F_{ST}=1) for each lineage. **Figure S4.** Functional differences between genes containing lineage-barcoding (F_{ST}=1) SNPs. **Figure S5.** Differentiation of sub-lineages when comparing the 62- versus 90-SNP barcodes. **Figure S6.** The increased resolution of our 90-SNP barcode (implemented in *TB-Profiler* software) over the comparable (sub-)lineages of the *SNP-IT* tool.

Abbreviations

MDR-TB: Multidrug-resistant TB; MTBC: *Mycobacterium tuberculosis* complex; RD: Region of difference; SNP: Single nucleotide polymorphism; TB: Tuberculosis; WGS: Whole genome sequencing; XDR-TB: Extensively drug-resistant TB

Acknowledgements

The MRC eMedLab computing resource was used for bioinformatics and statistical analysis.

Authors' contributions

JP and TGC conceived and directed the project. GN and JP performed bioinformatic and statistical analyses under the supervision of SC, MLH and TGC. GN, SC, JP and TGC interpreted results. YM, MA, AA, and YW contributed sequence data. GN, SC, JP and TGC wrote the first draft of the manuscript. All authors commented and edited on various versions of the draft manuscript. GN, JP and TGC compiled the final manuscript. All authors read and approved the final manuscript.

Funding

GN is supported by a BBSRC LiDO PhD studentship. TGC is funded by the Medical Research Council UK (Grant no. MR/M01360X/1, MR/N010469/1, MR/R025576/1, and MR/R020973/1) and BBSRC UK (Grant no. BB/R013063/1). SC is funded by Medical Research Council UK (MR/M01360X/1, MR/R025576/1, and MR/R020973/1) and BBSRC UK (Grant no. BB/R013063/1) grants. The study was funded in part from the core AHRI budget (NORAD and SIDA grants) and the National Institutes of Health (NIH) Fogarty International Center Global Infectious Diseases grant entitled "Ethiopia-Emory TB Research Training Program" (D43TW009127). These funding bodies did not have a role in the design of the study and collection, analysis and interpretation of data and in writing the manuscript.

Availability of data and materials

All raw sequence data is available from the EBI short read archive. A dedicated GitHub repository (https://github.com/GaryNapier/tb-lineages) [35] contains the list of accession numbers and code. The new (sub-)lineages have been implemented within the TB-Profiler tool https://github.com/jody-phelan/TBProfiler [14].

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Faculty of Infectious and Tropical Diseases, London School of Hygiene & Tropical Medicine, Keppel Street, London WC1E 7HT, UK. [2]Armauer Hansen Research Institute, Addis Ababa, Ethiopia. [3]Department of Microbiology, Immunology and Parasitology, College of Health Sciences, Addis Ababa University, Addis Ababa, Ethiopia. [4]Hawassa University College of Medicine and Health Sciences, Hawassa, Ethiopia. [5]Faculty of Epidemiology and Population Health, London School of Hygiene & Tropical Medicine, London WC1E 7HT, UK.

Received: 19 July 2020 Accepted: 3 December 2020
Published online: 14 December 2020

References

1. 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;5:4812 [cited 2017 Jul 17] Available from: http://www.nature.com/articles/ncomms5812.
2. Ngabonziza JCS, Loiseau C, Marceau M, Jouet A, Menardo F, Tzfadia O, et al. A sister lineage of the Mycobacterium tuberculosis complex discovered in the African Great Lakes region. Nat Commun. 2020;11:1–11.
3. Coscolla M, Brites D, Menardo F, Loiseau C, Darko Otchere I, Asante-Poku A, et al. Phylogenomics of Mycobacterium africanum reveals a new lineage and a complex evolutionary history. bioRxiv. 2020;17:19.
4. Brites D, Gagneux S. Co-evolution of *Mycobacterium tuberculosis* and *Homo sapiens*. Immunol Rev 2015;264:6–24. [cited 2018 Sep 3] Available from: http://www.ncbi.nlm.nih.gov/pubmed/25703549.
5. Ford CB, Shah RR, Maeda MK, Gagneux S, Murray MB, Cohen T, et al. *Mycobacterium tuberculosis* mutation rate estimates from different lineages predict substantial differences in the emergence of drug-resistant tuberculosis. Nat Genet ; 2013;45:784–90. [cited 2020 Oct 26] Available from: /pmc/articles/PMC3777616/?report=abstract.
6. Reiling N, Homolka S, Walter K, Brandenburg J, Niwinski L, Ernst M, et al. Clade-specific virulence patterns of *Mycobacterium tuberculosis* complex strains in human primary macrophages and aerogenically infected mice. MBio. American Society for Microbiology; 2013;4. [cited 2020 Oct 26] Available from: /pmc/articles/PMC3735190/?report=abstract.
7. Smittipat N, Miyahara R, Juthayothin T, Billamas P, Dokladda K, Imsanguan W, et al. Indo-Oceanic *Mycobacterium tuberculosis* strains from Thailand associated with higher mortality. Int J Tuberc Lung Dis. 2019;23:972–9 [cited 2020 Oct 26] Available from: https://pubmed.ncbi.nlm.nih.gov/31615603/.
8. Oppong YEA, Phelan J, Perdigão J, MacHado D, Miranda A, Portugal I, et al. Genome-wide analysis of *Mycobacterium tuberculosis* polymorphisms reveals lineage-specific associations with drug resistance. BMC Genomics; 2019;20.
9. Forrellad MA, Klepp LI, Gioffré A, García JS, Morbidoni HR, de la Paz Santangelo M, et al. Virulence factors of the *Mycobacterium tuberculosis* complex. Virulence. Taylor and Francis Inc.; 2013. p. 3–66.
10. Jagielski T, van Ingen J, Rastogi N, Dziadek J, Mazur PK, Bielecki J. Current methods in the molecular typing of Mycobacterium tuberculosis and other mycobacteria. Biomed Res Int. 2014;2014:645802. https://doi.org/10.1155/2014/645802. Epub 2014 Jan 5.
11. Kohl TA, Harmsen D, Rothgänger J, Walker T, Diel R, Niemann S. Harmonized genome wide typing of tubercle bacilli using a web-based gene-by-gene nomenclature system. EBioMedicine; 2018;34:131–8. [cited 2020 Oct 26] Available from: /pmc/articles/PMC6116475/?report=abstract.

12. Conceição EC, Refregier G, Gomes HM, Olessa-Daragon X, Coll F, Ratovonirina NH, et al. *Mycobacterium tuberculosis* lineage 1 genetic diversity in Pará, Brazil, suggests common ancestry with east-African isolates potentially linked to historical slave trade. Infect Genet Evol. 2019;73:337–41 [cited 2019 Jul 29] Available from: https://www.sciencedirect.com/science/article/pii/S1567134819301030?via%3Dihub.
13. Cancino-Muñoz I, Gil-Brusola A, Torres-Puente M, Mariner-Llicer C, Dogba J, Akinseye V, et al. Development and application of affordable SNP typing approaches to genotype *Mycobacterium tuberculosis* complex strains in low and high burden countries. Sci Rep. 2019;9:1–12 [cited 2020 Oct 29] Available from: https://doi.org/10.1038/s41598-019-51326-2.
14. Phelan JE, O'Sullivan DM, Machado D, Ramos J, Oppong YEA, Campino S, et al. Integrating informatics tools and portable sequencing technology for rapid detection of resistance to anti-tuberculous drugs. Genome Med. 2019;11:41. [cited 2019 Jun 28] Available from: http://www.ncbi.nlm.nih.gov/pubmed/31234910.
15. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30:2114–20 [cited 2018 Sep 5] Available from: http://www.ncbi.nlm.nih.gov/pubmed/24695404.
16. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. 2013; [cited 2017 Sep 6] Available from: http://arxiv.org/abs/1303.3997.
17. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics. 2011;27:2987–93. [cited 2017 Sep 6] Available from: http://www.ncbi.nlm.nih.gov/pubmed/21903627.
18. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20:1297–303.
19. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841–2.
20. Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, Korbel JO. DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics. 2012;28:i333–9.
21. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 2012;19:455–77.
22. 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;28:2991–3 [cited 2017 Sep 7] Available from: http://www.ncbi.nlm.nih.gov/pubmed/23014632.
23. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81:559–75 [cited 2017 Sep 6] Available from: http://www.ncbi.nlm.nih.gov/pubmed/17701901.
24. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 2015;32:268–74.
25. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. Nucleic Acids Res. 2019;47:W256–9 [cited 2019 Sep 12] Available from: https://academic.oup.com/nar/article/47/W1/W256/5424068.
26. Huerta-Cepas J, Serra F, Bork P. ETE 3: reconstruction, analysis, and visualization of phylogenomic data. Mol Biol Evol. 2016;33:1635–8 [cited 2020 Mar 16] Available from: http://www.ncbi.nlm.nih.gov/pubmed/26921390.
27. Welch BL. The generalization of 'Student's' problem when several different population variances are involved. Biometrika. 1947;34:28.
28. Cohen J. Statistical power analysis for the behavioral sciences. New York: Routledge Academic; 1988.
29. Weir BS, Cockerham CC. Estimating F-statistics for the analysis of population structure. Evolution. 1984;38:1358.
30. Lipworth S, Jajou R, De Neeling A, Bradley P, Van Der Hoek W, Maphalala G, et al. SNP-IT tool for identifying subspecies and associated lineages of *Mycobacterium tuberculosis* complex. Emerg Infect Dis. 2019;25:482–8.
31. Dejesus MA, Gerrick ER, Xu W, Park SW, Long JE, Boutte CC, et al. Comprehensive essentiality analysis of the Mycobacterium tuberculosis genome via saturating transposon mutagenesis. mBio. 2017;8(1):e02133-16. https://doi.org/10.1128/mBio.02133-16.
32. Phelan JE, Coll F, Bergval I, Anthony RM, Warren R, Sampson SL, et al. Recombination in pe/ppe genes contributes to genetic variation in *Mycobacterium tuberculosis* lineages. BMC Genomics. 2016;17:151 [cited 2017 Jul 17] Available from: http://www.ncbi.nlm.nih.gov/pubmed/26923687.
33. Price MN, Dehal PS, Arkin AP. FastTree 2 – approximately maximum-likelihood trees for large alignments. Poon AFY, editor. PLoS One; 2010;5:e9490.
34. Coscolla M, Gagneux S. Consequences of genomic diversity in *Mycobacterium tuberculosis*. Semin. Immunol; 2014; 431–44. [cited 2020 Oct 26] Available from: https://pubmed.ncbi.nlm.nih.gov/25453224/. Accessed 1 Nov 2020.
35. Napier G. tb-lineages. GitHub. https://github.com/GaryNapier/tb-lineages (2020). Accessed 1 Nov 2020.

Publisher's Note

Table S1: **The study samples (N=35,298) used and their lineages (L)**

Region	No. countries	L1	L2	L3	L4	L5	L6	L7	L8	L9	*M.bovis, caprae, orygis*
North America	10	317	390	228	1414	3	-	-	-	-	8
South America	6	4	57	1	851	-	-	-	-	-	2
Western Europe	14	749	1047	1880	4551	17	27	-	-	-	178
Eastern Europe	15	3	974	5	788	-	-	-	-	-	2
North Africa	5	-	-	-	60	-	-	-	-	-	-
West Africa	12	5	5	-	184	212	114	-	-	-	1
East Africa	9	43	6	36	112	-	1	64	2	3*	1
South Africa	3	80	1451	145	3380	-	-	-	-	-	1
Central Africa	10	291	93	238	1965	25	2	-	-	-	11
Central Asia	15	40	344	418	191	-	-	-	-	-	3
South Asia	2	309	74	109	55	-	-	-	-	-	-
East Asia	15	1202	3513	50	1060	-	-	-	-	-	-
Oceania	1	6	21	5	33	-	-	-	-	-	-
Unknown	-	316	903	810	3559	24	23	-	-	-	203
Total	118	3365	8878	3925	18203	281	167	64	2	3	410
Training N	-	2162	4556	2654	8320	26	32	38	2	3	110
Test N	-	1203	4322	1271	9883	255	135	26	-	-	300

* predicted in reference [3]

Table S2: Robust barcoding SNPs (421 SNPs, including the 90 minimal barcoding SNPs in Table S3)

Lineage	Position	Change	Strand	Amino acid	Gene	Locus	Functional Category
1	272678	C->T	-	54A	*Rv0227c*	Rv0227c	cell wall and cell processes
1	344288	C->G	+	89S	*eccB3*	Rv0283	cell wall and cell processes
1	615938	G->A	+	368E	*hemL*	Rv0524	IMR
1	646531	A->T	+	78T	*menD*	Rv0555	IMR
1	811492	C->G	+	40V	*rplN*	Rv0714	information pathways
1	812502	C->T	+	148V	*rplE*	Rv0716	information pathways
1	865761	C->T	+	392H	*purD*	Rv0772	IMR
1	1560912	G->A	+	156E	*pyrF*	Rv1385	IMR
1	1590555	C->T	+	53T	*ribA2*	Rv1415	IMR
1	2897528	G->A	-	92A	*aspS*	Rv2572c	information pathways
1	3233605	G->A	-	179L	*ftsY*	Rv2921c	cell wall and cell processes
1	3647591	A->G	-	73N	*rmlD*	Rv3266c	IMR
1	3830566	G->A	-	318S	*guaB2*	Rv3411c	IMR
1	4022652	G->A	-	384S	*cysS1*	Rv3580c	information pathways
1	4081987	G->C	-	245A	*Rv3644c*	Rv3644c	information pathways
1	4081996	G->C	-	242P	*Rv3644c*	Rv3644c	information pathways
1	4155266	C->G	+	509G	*leuA*	Rv3710	IMR
1.1	2989683	C->T	+	131A	*aftC*	Rv2673	cell wall and cell processes
1.1	4404247	G->A	+	352L	*Rv3915*	Rv3915	IMR
1.1.1	529363	C->T	+	252V	*groEL2*	Rv0440	VDA
1.1.1	870112	C->T	+	35A	*purB*	Rv0777	IMR
1.1.1	1261056	C->T	-	97G	*metE*	Rv1133c	IMR
1.1.1	1924765	T->C	+	313L	*pyrG*	Rv1699	IMR
1.1.1	2078024	G->A	+	716P	*gcvB*	Rv1832	IMR
1.1.1.1	1750465	T->C	+	924L	*dnaE1*	Rv1547	information pathways
1.1.1.1	2412584	G->A	-	256A	*murG*	Rv2153c	cell wall and cell processes
1.1.1.1	2994964	C->T	-	33G	*hemE*	Rv2678c	IMR
1.1.2	2622402	G->A	-	17A	*dnaG*	Rv2343c	information pathways
1.1.2	3879882	G->A	-	63N	*rpsM*	Rv3460c	information pathways
1.1.2	4157259	G->A	+	93V	*Rv3712*	Rv3712	IMR
1.1.3	15177	C->G	+	88A	*trpG*	Rv0013	IMR
1.1.3	1491275	G->A	-	346H	*glgB*	Rv1326c	IMR
1.1.3.1	403481	C->T	-	787R	*Rv0338c*	Rv0338c	IMR
1.1.3.1	1345104	G->A	-	22L	*dapD*	Rv1201c	IMR
1.1.3.2	285096	G->T	-	586R	*aftD*	Rv0236c	cell wall and cell processes
1.1.3.2	2369187	C->T	-	181L	*prcA*	Rv2109c	IMR
1.1.3.2	2369460	G->A	-	90D	*prcA*	Rv2109c	IMR
1.1.3.2	2418554	G->A	-	151L	*murF*	Rv2157c	cell wall and cell processes
1.1.3.2	4084405	G->A	+	533R	*Rv3645*	Rv3645	cell wall and cell processes

1.1.3.2	4154816	G->A	+	359P	*leuA*	Rv3710	IMR
1.1.3.3	2738352	C->T	-	445A	*obg*	Rv2440c	IMR
1.1.3.3	3337585	G->A	-	111D	*ddlA*	Rv2981c	cell wall and cell processes
1.2	1136017	A->G	-	155G	*prsA*	Rv1017c	IMR
1.2	1553855	C->T	+	208T	*pyrB*	Rv1380	IMR
1.2.1	590595	G->A	+	171Q	*proC*	Rv0500	IMR
1.2.1	2640960	C->T	-	35A	*glyS*	Rv2357c	information pathways
1.2.1	2847191	G->A	-	714D	*fas*	Rv2524c	lipid metabolism
1.2.1	3387252	G->A	-	260A	*fixB*	Rv3028c	IMR
1.2.1	4402048	C->T	+	107T	*trxB2*	Rv3913	IMR
1.2.2	528781	G->A	+	58E	*groEL2*	Rv0440	VDA
1.2.2	1567985	A->G	+	387E	*metK*	Rv1392	IMR
1.2.2	2639868	A->G	-	399D	*glyS*	Rv2357c	information pathways
1.2.2	2840849	G->C	-	2828G	*fas*	Rv2524c	lipid metabolism
1.2.2	3629612	G->A	-	12T	*sahH*	Rv3248c	IMR
1.2.2	3862181	C->T	-	70Q	*rplM*	Rv3443c	information pathways
1.2.2	4024368	C->T	-	224L	*ispD*	Rv3582c	IMR
1.2.2	4237383	C->A	+	73T	*dprE2*	Rv3791	lipid metabolism
1.2.2.1	2737201	A->C	-	349S	*proB*	Rv2439c	IMR
1.3	2763624	G->A	-	51L	*clpP1*	Rv2461c	IMR
1.3	4238120	G->A	+	63Q	*aftA*	Rv3792	cell wall and cell processes
1.3.1	1245275	C->T	+	49A	*gnd2*	Rv1122	IMR
1.3.1	1651063	C->G	+	116R	*Rv1463*	Rv1463	cell wall and cell processes
1.3.1	3323665	C->A	-	13P	*Rv2968c*	Rv2968c	cell wall and cell processes
1.3.1	3787156	G->A	+	281Q	*otsB2*	Rv3372	VDA
1.3.2	61842	T->C	+	483L	*dnaB*	Rv0058	information pathways
1.3.2	1492049	C->T	-	88L	*glgB*	Rv1326c	IMR
1.3.2	3147316	G->A	-	186G	*infB*	Rv2839c	information pathways
1.3.2	4200993	T->G	+	191A	*tyrA*	Rv3754	IMR
2	282892	C->T	-	1320T	*aftD*	Rv0236c	cell wall and cell processes
2	811753	C->T	+	4H	*rplX*	Rv0715	information pathways
2	4254431	G->A	-	506D	*accD4*	Rv3799c	lipid metabolism
2	4308395	G->A	-	174L	*serS*	Rv3834c	information pathways
2.1	648465	A->G	+	169A	*Rv0556*	Rv0556	cell wall and cell processes
2.1	1135798	C->T	-	228L	*prsA*	Rv1017c	IMR
2.1	2737453	A->C	-	265R	*proB*	Rv2439c	IMR
2.1	4165481	G->C	-	417P	*dnaZX*	Rv3721c	information pathways
2.2	195682	C->G	+	230V	*fadD5*	Rv0166	lipid metabolism
2.2	363464	G->A	+	71R	*Rv0298*	Rv0298	conserved hypotheticals
2.2	465300	C->T	+	630F	*Rv0386*	Rv0386	regulatory proteins
2.2	892416	C->T	-	286V	*Rv0799c*	Rv0799c	conserved hypotheticals

2.2	1288698	G->A	+	457G	*narG*	Rv1161	IMR
2.2	1695037	G->A	-	36F	*Rv1504c*	Rv1504c	conserved hypotheticals
2.2	1849051	C->T	-	995P	*lysX*	Rv1640c	information pathways
2.2	2112832	A->C	-	45A	*Rv1865c*	Rv1865c	IMR
2.2	2202500	C->T	+	121H	*higA*	Rv1956	VDA
2.2	2505085	G->A	-	205A	*cobC*	Rv2231c	IMR
2.2	2775361	C->T	+	30R	*Rv2472*	Rv2472	conserved hypotheticals
2.2	2903439	G->A	-	31S	*Rv2578c*	Rv2578c	conserved hypotheticals
2.2	3477942	A->G	+	98T	*moaA1*	Rv3109	IMR
2.2	3587446	G->A	-	32L	*Rv3210c*	Rv3210c	conserved hypotheticals
2.2	4050811	G->A	-	691Y	*ftsH*	Rv3610c	cell wall and cell processes
2.2	4186678	G->A	+	15L	*Rv3736*	Rv3736	regulatory proteins
2.2	4189210	G->T	+	504P	*Rv3737*	Rv3737	cell wall and cell processes
2.2.1	2078246	C->G	+	790G	*gcvB*	Rv1832	IMR
2.2.1	4158493	C->T	+	89I	*cobQ2*	Rv3713	IMR
2.2.1.1	1947282	A->G	-	46R	*vapC12*	Rv1720c	VDA
2.2.1.1	4080525	C->T	-	12R	*fic*	Rv3641c	cell wall and cell processes
2.2.1.2	1692069	A->G	+	60A	*Rv1501*	Rv1501	conserved hypotheticals
2.2.2	346693	G->T	+	353S	*eccC3*	Rv0284	cell wall and cell processes
2.2.2	1565566	C->T	+	42P	*dfp*	Rv1391	IMR
2.2.2	2640807	G->A	-	86V	*glyS*	Rv2357c	information pathways
2.2.2	3147511	T->G	-	121A	*infB*	Rv2839c	information pathways
3	342873	C->T	+	248V	*eccA3*	Rv0282	cell wall and cell processes
3	652950	T->C	+	60R	*grcC1*	Rv0562	IMR
3	1450316	C->T	+	314A	*thrA*	Rv1294	IMR
3	1764225	C->T	+	266A	*ilvA*	Rv1559	IMR
3	1925136	G->A	+	436V	*pyrG*	Rv1699	IMR
3	2738221	G->A	-	9I	*proB*	Rv2439c	IMR
3	2782498	G->A	-	515D	*Rv2477c*	Rv2477c	cell wall and cell processes
3	4396495	C->A	+	768G	*Rv3909*	Rv3909	conserved hypotheticals
3.1	958362	C->G	+	690A	*fadB*	Rv0860	lipid metabolism
3.1.1	1591545	G->T	+	383P	*ribA2*	Rv1415	IMR
3.1.1	3023684	G->A	+	40T	*ideR*	Rv2711	regulatory proteins
3.1.2	1914217	C->A	+	206R	*tyrS*	Rv1689	information pathways
3.1.2	3722702	G->C	-	310L	*trpS*	Rv3336c	information pathways
3.1.2.1	1237818	C->G	-	125L	*Rv1111c*	Rv1111c	conserved hypotheticals
3.1.2.1	2020120	C->T	+	288T	*eccC5*	Rv1783	cell wall and cell processes
3.1.2.1	2185538	G->A	-	217V	*fadE17*	Rv1934c	lipid metabolism
3.1.2.1	2245499	C->T	+	97S	*Rv2000*	Rv2000	conserved hypotheticals
3.1.2.1	2271143	G->A	-	202V	*Rv2025c*	Rv2025c	cell wall and cell processes
3.1.2.1	3557911	C->T	-	145L	*Rv3191c*	Rv3191c	insertion seqs and phages

3.1.2.1	4053713	C->G	-	45R	*Rv3612c*	Rv3612c	conserved hypotheticals
3.1.2.2	343092	C->T	+	321F	*eccA3*	Rv0282	cell wall and cell processes
3.1.2.2	346898	C->T	+	422L	*eccC3*	Rv0284	cell wall and cell processes
3.1.2.2	1862121	C->T	+	788H	*pheT*	Rv1650	information pathways
3.1.2.2	2415042	G->T	-	451P	*murD*	Rv2155c	cell wall and cell processes
3.1.2.2	2874344	G->A	-	714R	*alaS*	Rv2555c	information pathways
3.1.2.2	3629291	A->G	-	119G	*sahH*	Rv3248c	IMR
3.2	17842	G->C	-	307A	*pknA*	Rv0015c	regulatory proteins
3.2	1919627	G->A	+	294V	*ppnK*	Rv1695	IMR
4	2825466	A->G	+	263K	*Rv2509*	Rv2509	IMR
4	2994187	C->T	-	292L	*hemE*	Rv2678c	IMR
4	3830695	G->A	-	275A	*guaB2*	Rv3411c	IMR
4.1	62657	G->A	+	754P	*dnaB*	Rv0058	information pathways
4.1	284623	G->A	-	743T	*aftD*	Rv0236c	cell wall and cell processes
4.1	902413	C->T	+	101V	*purF*	Rv0808	IMR
4.1.1	265968	C->G	+	154R	*echA1*	Rv0222	lipid metabolism
4.1.1	514245	C->T	-	359V	*ctpH*	Rv0425c	cell wall and cell processes
4.1.1	869440	C->T	-	108L	*Rv0776c*	Rv0776c	conserved hypotheticals
4.1.1	1256806	C->T	+	225D	*prpC*	Rv1131	IMR
4.1.1	1952601	C->T	+	250H	*Rv1726*	Rv1726	IMR
4.1.1	2158582	G->A	-	170G	*fadB5*	Rv1912c	lipid metabolism
4.1.1	2603797	G->A	-	142I	*lppP*	Rv2330c	cell wall and cell processes
4.1.1	2752854	G->A	-	47D	*Rv2452c*	Rv2452c	conserved hypotheticals
4.1.1	3129359	C->T	-	805K	*Rv2823c*	Rv2823c	conserved hypotheticals
4.1.1	3231091	C->A	-	472V	*amt*	Rv2920c	cell wall and cell processes
4.1.1	3597737	C->T	-	10V	*TB7.3*	Rv3221c	lipid metabolism
4.1.1	4003130	G->A	+	498V	*fadD3*	Rv3561	lipid metabolism
4.1.1	4306767	G->A	-	15G	*Rv3832c*	Rv3832c	conserved hypotheticals
4.1.1.1	1006080	G->A	-	161V	*prrA*	Rv0903c	regulatory proteins
4.1.1.1	2824839	C->T	+	54A	*Rv2509*	Rv2509	IMR
4.1.1.2	1109535	G->A	+	88K	*galU*	Rv0993	IMR
4.1.1.2	3213615	G->A	-	80Y	*lepB*	Rv2903c	cell wall and cell processes
4.1.1.3	896356	C->T	+	179T	*purL*	Rv0803	IMR
4.1.1.3	4154051	G->A	+	104R	*leuA*	Rv3710	IMR
4.1.1.3	4229087	C->T	+	247N	*glfT1*	Rv3782	cell wall and cell processes
4.1.1.3.1	286300	C->A	-	184A	*aftD*	Rv0236c	cell wall and cell processes
4.1.1.3.1	2739087	C->T	-	200V	*obg*	Rv2440c	IMR
4.1.1.3.1	4269540	G->A	-	98P	*ubiA*	Rv3806c	cell wall and cell processes
4.1.2	3147742	A->G	-	44V	*infB*	Rv2839c	information pathways
4.1.2.1	342340	C->T	+	71L	*eccA3*	Rv0282	cell wall and cell processes
4.1.2.1	4256758	G->C	-	1463P	*pks13*	Rv3800c	lipid metabolism

4.1.2.1	4331585	G->A	-	1499D	*gltB*	Rv3859c	IMR
4.1.2.1.1	2488724	C->G	+	370P	*glnA1*	Rv2220	IMR
4.1.2.1.1	2640369	G->A	-	232Y	*glyS*	Rv2357c	information pathways
4.1.2.1.1	4404313	G->A	+	374L	*Rv3915*	Rv3915	IMR
4.1.3	1564799	C->T	+	133P	*gmk*	Rv1389	IMR
4.1.3	2450245	G->A	-	302A	*pimB*	Rv2188c	lipid metabolism
4.1.3	4228101	C->T	+	191D	*rfbE*	Rv3781	cell wall and cell processes
4.1.4	58786	G->C	+	67V	*ssb*	Rv0054	information pathways
4.1.4	590250	G->T	+	56T	*proC*	Rv0500	IMR
4.1.4	2844014	G->A	-	1773F	*fas*	Rv2524c	lipid metabolism
4.1.4	4391663	G->A	-	471L	*pcnA*	Rv3907c	information pathways
4.2	1466779	C->T	+	313I	*atpD*	Rv1310	IMR
4.2	1568018	C->T	+	398D	*metK*	Rv1392	IMR
4.2	1670814	C->T	+	134G	*Rv1480*	Rv1480	conserved hypotheticals
4.2	1872211	G->A	+	283A	*argG*	Rv1658	IMR
4.2	2748087	G->A	-	713S	*valS*	Rv2448c	information pathways
4.2	3198496	G->A	-	204I	*tsf*	Rv2889c	information pathways
4.2	3469694	G->T	-	30A	*smpB*	Rv3100c	VDA
4.2	3666905	C->T	+	183T	*accA3*	Rv3285	lipid metabolism
4.2.1	783601	A->C	+	373R	*fusA1*	Rv0684	information pathways
4.2.1	3646964	C->G	-	282L	*rmlD*	Rv3266c	IMR
4.2.1.1	870187	C->T	+	60D	*purB*	Rv0777	IMR
4.2.1.1	1652216	A->G	+	233K	*csd*	Rv1464	IMR
4.2.2	353766	T->C	+	228I	*eccD3*	Rv0290	cell wall and cell processes
4.2.2	2420503	A->G	-	36L	*murE*	Rv2158c	cell wall and cell processes
4.2.2.1	1131	C->A	+	377I	*dnaA*	Rv0001	information pathways
4.2.2.1	1455780	T->C	+	96L	*prfA*	Rv1299	information pathways
4.2.2.2	611463	G->A	-	204T	*Rv0519c*	Rv0519c	cell wall and cell processes
4.2.2.2	1233285	G->A	-	224Y	*Rv1106c*	Rv1106c	IMR
4.2.2.2	1880850	G->A	+	1849L	*pks7*	Rv1661	lipid metabolism
4.2.2.2	2153246	T->G	-	213R	*Rv1907c*	Rv1907c	conserved hypotheticals
4.2.2.2	2156847	G->A	-	151T	*Rv1910c*	Rv1910c	cell wall and cell processes
4.2.2.2	2923264	G->A	-	324L	*ruvB*	Rv2592c	information pathways
4.2.2.2	3141827	C->G	-	132L	*ugpA*	Rv2835c	cell wall and cell processes
4.2.2.2	3416734	G->A	+	10L	*dinP*	Rv3056	information pathways
4.2.2.2	3497369	G->A	+	273L	*Rv3131*	Rv3131	conserved hypotheticals
4.2.2.2	4298106	C->A	-	500S	*pks2*	Rv3825c	lipid metabolism
4.3	1452071	C->A	+	25G	*thrB*	Rv1296	IMR
4.3	3191027	G->A	-	199L	*cdsA*	Rv2881c	lipid metabolism
4.3.1	825585	T->C	+	262Y	*secY*	Rv0732	cell wall and cell processes
4.3.1.1	1647807	T->C	+	273I	*Rv1461*	Rv1461	conserved hypotheticals

4.3.1.1	3360032	G->A	-	185T	*ilvC*	Rv3001c	IMR
4.3.2	3414791	G->C	-	56A	*nrdH*	Rv3053c	information pathways
4.3.2.1	784581	G->C	+	699T	*fusA1*	Rv0684	information pathways
4.3.2.1	1451542	C->T	+	282A	*thrC*	Rv1295	IMR
4.3.2.1	1592015	C->T	+	115G	*ribH*	Rv1416	IMR
4.3.2.1	2844689	G->C	-	1548L	*fas*	Rv2524c	lipid metabolism
4.3.3	2077253	G->A	+	459T	*gcvB*	Rv1832	IMR
4.3.4	1297327	G->A	+	392V	*lpqW*	Rv1166	cell wall and cell processes
4.3.4.1	1274335	G->A	+	327L	*mmpL13b*	Rv1146	cell wall and cell processes
4.3.4.1	2199684	G->A	-	117F	*Rv1949c*	Rv1949c	conserved hypotheticals
4.3.4.1	3244674	G->A	+	326R	*fadD26*	Rv2930	lipid metabolism
4.3.4.1	3285945	C->G	+	292A	*mmpL7*	Rv2942	cell wall and cell processes
4.3.4.2	784440	G->T	+	652A	*fusA1*	Rv0684	information pathways
4.3.4.2.1	225495	T->C	-	359V	*Rv0193c*	Rv0193c	conserved hypotheticals
4.4	4238963	C->T	+	344H	*aftA*	Rv3792	cell wall and cell processes
4.4	4307886	G->A	-	343R	*serS*	Rv3834c	information pathways
4.4.1	2905505	G->A	-	196T	*hisS*	Rv2580c	information pathways
4.4.1	3147376	G->A	-	166P	*infB*	Rv2839c	information pathways
4.4.1	3664135	G->A	+	149R	*accE5*	Rv3281	lipid metabolism
4.4.1	3813473	G->A	-	202L	*guaA*	Rv3396c	IMR
4.4.1.1	355181	G->A	+	228K	*mycP3*	Rv0291	IMR
4.4.1.1.1	15036	C->G	+	41A	*trpG*	Rv0013	IMR
4.4.1.1.1	1126895	G->A	-	37L	*metS*	Rv1007c	information pathways
4.4.1.1.1	1221479	C->T	+	302V	*glyA1*	Rv1093	IMR
4.4.1.2	342201	C->G	+	24P	*eccA3*	Rv0282	cell wall and cell processes
4.4.1.2	345697	C->T	+	21T	*eccC3*	Rv0284	cell wall and cell processes
4.4.1.2	1297981	G->C	+	610V	*lpqW*	Rv1166	cell wall and cell processes
4.4.1.2	1494231	G->A	-	65L	*glgE*	Rv1327c	IMR
4.4.1.2	1803959	G->T	+	222G	*hisA*	Rv1603	IMR
4.4.1.2	1808124	A->C	+	74P	*trpE*	Rv1609	IMR
4.4.1.2	2410938	A->T	-	395A	*murC*	Rv2152c	cell wall and cell processes
4.4.1.2	3349093	G->C	-	395P	*gltS*	Rv2992c	information pathways
4.4.1.2	4392120	G->A	-	318H	*pcnA*	Rv3907c	information pathways
4.4.2	985287	G->A	+	495P	*fprB*	Rv0886	IMR
4.4.2	2913091	C->T	-	307V	*secF*	Rv2586c	cell wall and cell processes
4.5	620029	C->T	+	47L	*ccsA*	Rv0529	IMR
4.6	18091	G->A	-	224T	*pknA*	Rv0015c	regulatory proteins
4.6.1	435708	G->A	-	354T	*purA*	Rv0357c	IMR
4.6.1	2440953	G->T	-	256R	*aroG*	Rv2178c	IMR
4.6.1	4260268	G->C	-	293A	*pks13*	Rv3800c	lipid metabolism
4.6.1.1	4406749	G->A	-	261L	*parA*	Rv3918c	cell wall and cell processes

4.6.1.2	1098698	C->G	+	397G	*mprB*	Rv0982	regulatory proteins
4.6.2	4260742	G->A	-	135P	*pks13*	Rv3800c	lipid metabolism
4.6.2.1	896119	C->T	+	100F	*purL*	Rv0803	IMR
4.6.2.1	2897684	A->G	-	40D	*aspS*	Rv2572c	information pathways
4.6.2.2	118469	G->A	+	252A	*Rv0102*	Rv0102	cell wall and cell processes
4.6.2.2	1352350	C->T	+	69V	*gpgS*	Rv1208	IMR
4.6.2.2	2369118	G->A	-	204G	*prcA*	Rv2109c	IMR
4.6.2.2	2875883	C->T	-	201L	*alaS*	Rv2555c	information pathways
4.6.2.2	3354625	G->A	-	149L	*serA1*	Rv2996c	IMR
4.6.2.2	3360152	C->A	-	145P	*ilvC*	Rv3001c	IMR
4.6.3	734562	G->A	+	103K	*nusG*	Rv0639	information pathways
4.6.3	2516158	C->G	+	285T	*Rv2242*	Rv2242	conserved hypotheticals
4.6.3	2516365	T->C	+	354Y	*Rv2242*	Rv2242	conserved hypotheticals
4.6.4	4236903	G->A	+	375A	*dprE1*	Rv3790	lipid metabolism
4.6.5	17665	G->A	-	366N	*pknA*	Rv0015c	regulatory proteins
4.6.5	1553876	C->G	+	215A	*pyrB*	Rv1380	IMR
4.7	716918	G->A	+	85T	*vapC30*	Rv0624	VDA
4.7	3270289	C->A	+	851R	*ppsE*	Rv2935	lipid metabolism
4.7	4112595	G->A	-	307A	*Rv3671c*	Rv3671c	IMR
4.8	1130526	G->A	+	112A	*ispE*	Rv1011	IMR
4.8.1	2914906	G->C	-	277T	*secD*	Rv2587c	cell wall and cell processes
4.8.1	3348870	G->A	-	470L	*gltS*	Rv2992c	information pathways
4.8.1	3389922	T->G	+	274P	*Rv3030*	Rv3030	conserved hypotheticals
4.8.2	2417281	G->A	-	65Y	*murX*	Rv2156c	cell wall and cell processes
4.8.2	3404883	G->A	-	13A	*ctaD*	Rv3043c	IMR
4.8.3	616408	C->G	+	62A	*Rv0525*	Rv0525	conserved hypotheticals
4.9	420008	G->A	+	58A	*dnaK*	Rv0350	VDA
4.9	903913	C->T	+	63G	*purM*	Rv0809	IMR
4.9	3367765	A->G	-	343G	*gatB*	Rv3009c	information pathways
4.9.1	119600	C->G	+	629V	*Rv0102*	Rv0102	cell wall and cell processes
4.9.1	1940611	G->C	+	108A	*engA*	Rv1713	IMR
4.9.1	4165205	C->T	-	509A	*dnaZX*	Rv3721c	information pathways
5	345317	G->C	+	432V	*eccB3*	Rv0283	cell wall and cell processes
5	352646	G->A	+	166P	*espG3*	Rv0289	cell wall and cell processes
5	801959	C->T	+	166R	*rplD*	Rv0702	information pathways
5	1505806	C->T	+	244P	*murI*	Rv1338	cell wall and cell processes
5	1555432	C->T	+	415T	*pyrC*	Rv1381	IMR
5	1578212	A->C	+	200A	*priA*	Rv1402	information pathways
5	1649265	G->A	+	759L	*Rv1461*	Rv1461	conserved hypotheticals
5	1799921	C->A	+	113G	*hisD*	Rv1599	IMR
5	2485956	G->A	+	228E	*lipA*	Rv2218	IMR

5	2859147	C->T	-	48K	*Efp*	Rv2534c	Information pathways
5	3882025	G->A	+	63L	*rmlB*	Rv3464	IMR
5	4086604	G->A	-	218Y	*topA*	Rv3646c	information pathways
5	4387392	G->A	-	168F	*Rv3902c*	Rv3902c	conserved hypotheticals
6	982363	G->T	-	64G	*serC*	Rv0884c	IMR
6	1069146	C->T	+	314G	*purH*	Rv0957	IMR
6	1372002	C->T	-	316L	*mrp*	Rv1229c	IMR
6	1811964	C->A	+	280R	*trpB*	Rv1612	IMR
6	1867707	C->T	+	359P	*argJ*	Rv1653	IMR
6	2847737	G->A	-	532I	*fas*	Rv2524c	lipid metabolism
6	3213255	C->T	-	200K	*lepB*	Rv2903c	cell wall and cell processes
6	3862148	G->A	-	81P	*rplM*	Rv3443c	information pathways
6	4086697	G->T	-	187A	*topA*	Rv3646c	information pathways
6	4236891	C->A	+	371P	*dprE1*	Rv3790	lipid metabolism
7	349081	G->A	+	1149L	*eccC3*	Rv0284	cell wall and cell processes
7	784143	A->C	+	553A	*fusA1*	Rv0684	information pathways
7	811642	C->T	+	90D	*rplN*	Rv0714	information pathways
7	896431	G->A	+	204L	*purL*	Rv0803	IMR
7	1125894	A->C	-	370L	*metS*	Rv1007c	information pathways
7	1137518	G->A	-	181N	*glmU*	Rv1018c	cell wall and cell processes
7	1297084	G->A	+	311L	*lpqW*	Rv1166	cell wall and cell processes
7	1365895	G->A	+	7T	*htrA*	Rv1223	IMR
7	1463776	C->T	+	183V	*atpA*	Rv1308	IMR
7	1561245	C->T	+	267A	*pyrF*	Rv1385	IMR
7	1663221	T->G	-	942S	*acn*	Rv1475c	IMR
7	1799774	C->T	+	64A	*hisD*	Rv1599	IMR
7	1867937	C->T	+	32V	*argB*	Rv1654	IMR
7	2086202	C->G	-	260V	*glcB*	Rv1837c	IMR
7	2380244	G->A	-	139A	*hisG*	Rv2121c	IMR
7	2406193	G->C	-	217L	*Rv2147c*	Rv2147c	conserved hypotheticals
7	2621157	T->C	-	432A	*dnaG*	Rv2343c	information pathways
7	2759363	A->G	-	42C	*clpX*	Rv2457c	IMR
7	2842238	C->A	-	2365A	*fas*	Rv2524c	lipid metabolism
7	2999030	G->A	-	313G	*dxs1*	Rv2682c	IMR
7	3182040	C->T	-	324A	*dxr*	Rv2870c	IMR
7	3225566	C->T	-	240A	*ffh*	Rv2916c	cell wall and cell processes
7	3324231	G->C	-	82S	*Rv2969c*	Rv2969c	cell wall and cell processes
7	3360233	G->A	-	118A	*ilvC*	Rv3001c	IMR
7	3473482	C->T	-	141R	*prfB*	Rv3105c	information pathways
7	3603631	G->T	+	85G	*aroA*	Rv3227	IMR
7	3635935	C->G	-	111R	*manA*	Rv3255c	IMR

7	3666497	C->T	+	47A	*accA3*	Rv3285	lipid metabolism
7	3667883	C->A	+	509V	*accA3*	Rv3285	lipid metabolism
7	3837064	G->A	-	75G	*groES*	Rv3418c	VDA
7	3860696	G->A	-	225D	*mrsA*	Rv3441c	IMR
7	3878040	A->G	-	156G	*rpoA*	Rv3457c	information pathways
7	4022163	C->T	-	77L	*Rv3579c*	Rv3579c	IMR
7	4086748	C->G	-	170L	*topA*	Rv3646c	information pathways
7	4262608	G->A	-	153I	*fadD32*	Rv3801c	lipid metabolism
7	4267649	A->G	-	396G	*aftB*	Rv3805c	cell wall and cell processes
7	4308411	T->G	-	168L	*serS*	Rv3834c	information pathways
7	4331184	G->A	-	108L	*gltD*	Rv3858c	IMR
8	221190	G->T	-	178V	*ilvD*	Rv0189c	IMR
8	270362	C->A	-	401V	*Rv0226c*	Rv0226c	cell wall and cell processes
8	343314	C->T	+	395A	*eccA3*	Rv0282	cell wall and cell processes
8	344414	G->C	+	131S	*eccB3*	Rv0283	cell wall and cell processes
8	347977	A->C	+	781P	*eccC3*	Rv0284	cell wall and cell processes
8	442072	C->A	-	76A	*fba*	Rv0363c	IMR
8	742270	G->T	-	116S	*Rv0647c*	Rv0647c	conserved hypotheticals
8	896152	C->T	+	111V	*purL*	Rv0803	IMR
8	903262	C->A	+	384V	*purF*	Rv0808	IMR
8	1227518	G->A	-	16T	*fum*	Rv1098c	IMR
8	1260255	C->T	-	364A	*metE*	Rv1133c	IMR
8	1446846	C->T	+	156T	*argS*	Rv1292	information pathways
8	1463923	C->T	+	232T	*atpA*	Rv1308	IMR
8	1505182	C->T	+	36V	*murI*	Rv1338	cell wall and cell processes
8	1553399	C->T	+	56T	*pyrB*	Rv1380	IMR
8	1564640	C->A	+	80L	*gmk*	Rv1389	IMR
8	1629090	C->G	-	370P	*tkt*	Rv1449c	IMR
8	1665285	C->A	-	254P	*acn*	Rv1475c	IMR
8	1805152	C->T	+	100V	*hisF*	Rv1605	IMR
8	1858921	C->G	+	63V	*pheS*	Rv1649	information pathways
8	1914246	C->T	+	215T	*tyrS*	Rv1689	information pathways
8	1922231	C->T	+	230V	*Rv1697*	Rv1697	conserved hypotheticals
8	1941256	C->A	+	323V	*engA*	Rv1713	IMR
8	2086736	G->A	-	82R	*glcB*	Rv1837c	IMR
8	2415402	C->T	-	331K	*murD*	Rv2155c	cell wall and cell processes
8	2440802	A->G	-	307L	*aroG*	Rv2178c	IMR
8	2450263	C->T	-	296V	*pimB*	Rv2188c	lipid metabolism
8	2466760	G->T	+	588T	*asnB*	Rv2201	IMR
8	2862497	G->T	-	60A	*aroB*	Rv2538c	IMR
8	2939600	C->A	-	121L	*Rv2611c*	Rv2611c	lipid metabolism

8	2940079	C->A	-	178S	*pgsA1*	Rv2612c	lipid metabolism
8	3151706	C->T	-	415L	*proS*	Rv2845c	information pathways
8	3333720	C->A	-	23V	*thiL*	Rv2977c	IMR
8	3361394	G->A	-	531A	*ilvB1*	Rv3003c	IMR
8	3362522	G->A	-	155I	*ilvB1*	Rv3003c	IMR
8	3408464	G->A	-	305S	*nrdF2*	Rv3048c	information pathways
8	3469397	G->A	-	129G	*smpB*	Rv3100c	VDA
8	3645324	G->A	-	218S	*manB*	Rv3264c	cell wall and cell processes
8	3786517	G->A	+	68S	*otsB2*	Rv3372	VDA
8	3830815	G->A	-	235D	*guaB2*	Rv3411c	IMR
8	4166096	G->A	-	212S	*dnaZX*	Rv3721c	information pathways
8	4227348	C->T	+	120R	*Rv3780*	Rv3780	conserved hypotheticals
8	4406599	G->A	-	311S	*parA*	Rv3918c	cell wall and cell processes
9	2750052	G->A	-	58T	*valS*	Rv2448c	information pathways
9	3094577	G->A	-	108L	*ribF*	Rv2786c	IMR
9	3370805	G->C	-	210S	*gatA*	Rv3011c	information pathways
9	3414553	G->A	-	44Y	*nrdI*	Rv3052c	information pathways
9	3855303	G->A	-	529I	*glmS*	Rv3436c	IMR
M.bovis	62768	A->G	+	791G	*dnaB*	Rv0058	information pathways
M.bovis	3371401	G->A	-	12L	*gatA*	Rv3011c	information pathways
M.bovis	4229470	T->C	+	71Y	*rfbD*	Rv3783	cell wall and cell processes
M.caprae	904090	T->C	+	122G	*purM*	Rv0809	IMR
M.caprae	918685	C->T	-	22L	*desA1*	Rv0824c	lipid metabolism
M.caprae	1069305	G->C	+	367L	*purH*	Rv0957	IMR
M.caprae	2990241	G->T	+	317S	*aftC*	Rv2673	cell wall and cell processes
M.caprae	3094181	G->A	-	240F	*ribF*	Rv2786c	IMR
M.orygis	44812	G->T	+	417P	*leuS*	Rv0041	information pathways
M.orygis	59181	C->T	+	20C	*rpsR1*	Rv0055	information pathways
M.orygis	268953	C->T	+	97V	*Rv0225*	Rv0225	cell wall and cell processes
M.orygis	500710	G->A	-	103P	*thiE*	Rv0414c	IMR
M.orygis	748320	G->A	+	15A	*rplJ*	Rv0651	information pathways
M.orygis	1125468	C->T	-	512P	*metS*	Rv1007c	information pathways
M.orygis	1236745	G->C	+	187S	*lytB2*	Rv1110	cell wall and cell processes
M.orygis	1344807	G->A	-	121P	*dapD*	Rv1201c	IMR
M.orygis	1449038	C->T	+	337G	*lysA*	Rv1293	IMR
M.orygis	1555342	C->T	+	385A	*pyrC*	Rv1381	IMR
M.orygis	1579197	C->T	+	529L	*priA*	Rv1402	information pathways
M.orygis	1652839	G->C	+	24G	*Rv1465*	Rv1465	IMR
M.orygis	1810542	G->A	+	101S	*trpC*	Rv1611	IMR
M.orygis	2069383	T->C	+	102L	*pgsA2*	Rv1822	lipid metabolism
M.orygis	2076692	G->A	+	272R	*gcvB*	Rv1832	IMR

M.orygis	2078567	T->C	+	897Y	*gcvB*	Rv1832	IMR
M.orygis	2475305	G->A	-	222N	*ilvE*	Rv2210c	IMR
M.orygis	3039261	T->C	-	180A	*dapF*	Rv2726c	IMR
M.orygis	3061003	A->G	-	502L	*ftsK*	Rv2748c	cell wall and cell processes
M.orygis	3147196	G->A	-	226D	*infB*	Rv2839c	information pathways
M.orygis	3148454	G->A	-	325G	*nusA*	Rv2841c	information pathways
M.orygis	3241414	G->A	-	182A	*Rv2927c*	Rv2927c	conserved hypotheticals
M.orygis	3337480	A->G	-	146S	*ddlA*	Rv2981c	cell wall and cell processes
M.orygis	3337675	G->A	-	81G	*ddlA*	Rv2981c	cell wall and cell processes
M.orygis	3667352	C->T	+	332D	*accA3*	Rv3285	lipid metabolism
M.orygis	3770449	G->A	-	67R	*folD*	Rv3356c	IMR
M.orygis	4039853	C->T	-	284L	*clpC1*	Rv3596c	IMR
M.orygis	4311950	T->C	-	240E	*pheA*	Rv3838c	IMR

IMR = Intermediary metabolism and respiration; VDA = virulence, detoxification, adaptation

Table S3: The ninety minimal barcoding SNPs

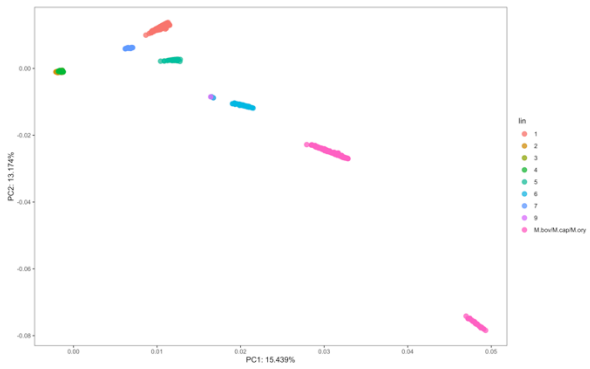
Lineage	Position	Change	Amino Acid	Gene*	Locus	Functional Category
1	615938	G->A	368E	*hemL*	Rv0524	IMR
1.1	4404247	G->A	352L	*Rv3915*	Rv3915	IMR
1.1.1	**529363**	C->T	252V	*groEL2*	Rv0440	VDA
1.1.1.1	**1750465**	T->C	924L	*dnaE1*	Rv1547	information pathways
1.1.2	2622402	G->A	17A	*dnaG*	Rv2343c	information pathways
1.1.3	1491275	G->A	346H	*glgB*	Rv1326c	IMR
1.1.3.1	**403481**	C->T	787R	*Rv0338c*	Rv0338c	IMR
1.1.3.2	**285096**	G->T	586R	*aftD*	Rv0236c	cell wall and cell processes
1.1.3.3	**2738352**	C->T	445A	*obg*	Rv2440c	IMR
1.2	**1136017**	A->G	155G	*prsA*	Rv1017c	IMR
1.2.1	**590595**	G->A	171Q	*proC*	Rv0500	IMR
1.2.2	**528781**	G->A	58E	*groEL2*	Rv0440	VDA
1.2.2.1	**2737201**	A->C	349S	*proB*	Rv2439c	IMR
1.3	**2763624**	G->A	51L	*clpP1*	Rv2461c	IMR
1.3.1	**1245275**	C->T	49A	*gnd2*	Rv1122	IMR
1.3.2	**61842**	T->C	483L	*dnaB*	Rv0058	information pathways
2	**282892**	C->T	1320T	*aftD*	Rv0236c	cell wall and cell processes
2.1	**648465**	A->G	169A	*Rv0556*	Rv0556	cell wall and cell processes
2.2	2505085	G->A	205A	*cobC**	Rv2231c	IMR
2.2.1	**2078246**	C->G	790G	*gcvB*	Rv1832	IMR
2.2.1.1	**1947282**	A->G	46R	*vapC12**	Rv1720c	VDA
2.2.1.2	**1692069**	A->G	60A	*Rv1501**	Rv1501	conserved hypotheticals
2.2.2	346693	G->T	353S	*eccC3*	Rv0284	cell wall and cell processes
3	**342873**	C->T	248V	*eccA3*	Rv0282	cell wall and cell processes
3.1	**958362**	C->G	690A	*fadB**	Rv0860	lipid metabolism
3.1.1	**1591545**	G->T	383P	*ribA2*	Rv1415	IMR
3.1.2	3722702	G->C	310L	*trpS*	Rv3336c	information pathways
3.1.2.1	1237818	C->G	125L	*Rv1111c**	Rv1111c	conserved hypotheticals
3.1.2.2	2874344	G->A	714R	*alaS*	Rv2555c	information pathways
3.2	**17842**	G->C	307A	*pknA*	Rv0015c	regulatory proteins
4	**2825466**	G->A	263K	*Rv2509*	Rv2509	IMR
4.1	62657	G->A	754P	*dnaB*	Rv0058	information pathways
4.1.1	514245	C->T	359V	*ctpH**	Rv0425c	cell wall and cell processes
4.1.1.1	**1006080**	G->A	161V	*prrA*	Rv0903c	regulatory proteins
4.1.1.2	**1109535**	G->A	88K	*galU*	Rv0993	IMR
4.1.1.3	4229087	C->T	247N	*glfT1*	Rv3782	cell wall and cell processes
4.1.1.3.1	**286300**	C->A	184A	*aftD*	Rv0236c	cell wall and cell processes
4.1.2	**3147742**	A->G	44V	*infB*	Rv2839c	information pathways
4.1.2.1	**342340**	C->T	71L	*eccA3*	Rv0282	cell wall and cell processes
4.1.2.1.1	**2488724**	C->G	370P	*glnA1*	Rv2220	IMR
4.1.3	**1564799**	C->T	133P	*gmk*	Rv1389	IMR

4.1.4	**58786**	G->C	67V	*ssb*	Rv0054	information pathways
4.2	**1466779**	C->T	313I	*atpD*	Rv1310	IMR
4.2.1	783601	A->C	373R	*fusA1*	Rv0684	information pathways
4.2.1.1	**870187**	C->T	60D	*purB*	Rv0777	IMR
4.2.2	**353766**	T->C	228I	*eccD3*	Rv0290	cell wall and cell processes
4.2.2.1	1455780	T->C	96L	*prfA*	Rv1299	information pathways
4.2.2.2	**611463**	G->A	204T	*Rv0519c**	Rv0519c	cell wall and cell processes
4.3	**1452071**	C->A	25G	*thrB*	Rv1296	IMR
4.3.1	**825585**	T->C	262Y	*secY*	Rv0732	cell wall and cell processes
4.3.1.1	**1647807**	T->C	273I	*Rv1461*	Rv1461	conserved hypotheticals
4.3.2	**3414791**	G->C	56A	*nrdH*	Rv3053c	information pathways
4.3.2.1	**784581**	G->C	699T	*fusA1*	Rv0684	information pathways
4.3.3	**2077253**	G->A	459T	*gcvB*	Rv1832	IMR
4.3.4	**1297327**	G->A	392V	*lpqW*	Rv1166	cell wall and cell processes
4.3.4.1	**1274335**	G->A	327L	*mmpL13b**	Rv1146	cell wall and cell processes
4.3.4.2	**784440**	G->T	652A	*fusA1*	Rv0684	information pathways
4.3.4.2.1	**225495**	T->C	359V	*Rv0193c**	Rv0193c	conserved hypotheticals
4.4	4307886	G->A	343R	*serS*	Rv3834c	information pathways
4.4.1	**2905505**	G->A	196T	*hisS*	Rv2580c	information pathways
4.4.1.1	355181	G->A	228K	*mycP3*	Rv0291	IMR
4.4.1.1.1	**15036**	C->G	41A	*trpG*	Rv0013	IMR
4.4.1.2	**342201**	C->G	24P	*eccA3*	Rv0282	cell wall and cell processes
4.4.2	**985287**	G->A	495P	*fprB**	Rv0886	IMR
4.5	**620029**	C->T	47L	*ccsA*	Rv0529	IMR
4.6	**18091**	G->A	224T	*pknA*	Rv0015c	regulatory proteins
4.6.1	4260268	G->C	293A	*pks13*	Rv3800c	lipid metabolism
4.6.1.1	**4406749**	G->A	261L	*parA*	Rv3918c	cell wall and cell processes
4.6.1.2	**1098698**	C->G	397G	*mprB*	Rv0982	regulatory proteins
4.6.2	**4260742**	G->A	135P	*pks13*	Rv3800c	lipid metabolism
4.6.2.1	**896119**	C->T	100F	*purL*	Rv0803	IMR
4.6.2.2	2875883	C->T	201L	*alaS*	Rv2555c	information pathways
4.6.3	**734562**	G->A	103K	*nusG*	Rv0639	information pathways
4.6.4	**4236903**	G->A	375A	*dprE1*	Rv3790	lipid metabolism
4.6.5	**17665**	G->A	366N	*pknA*	Rv0015c	regulatory proteins
4.7	**716918**	G->A	85T	*vapC30**	Rv0624	VDA
4.8	**1130526**	G->A	112A	*ispE*	Rv1011	IMR
4.8.1	**2914906**	G->C	277T	*secD*	Rv2587c	cell wall and cell processes
4.8.2	**2417281**	G->A	65Y	*murX*	Rv2156c	cell wall and cell processes
4.8.3	**616408**	C->G	62A	*Rv0525*	Rv0525	conserved hypotheticals
4.9	**420008**	A->G	58A	*dnaK*	Rv0350	VDA
4.9.1	**119600**	C->G	629V	*Rv0102*	Rv0102	cell wall and cell processes
5	1799921	C->A	113G	*hisD*	Rv1599	IMR
6	**982363**	G->T	64G	*serC*	Rv0884c	IMR
7	1137518	G->A	181N	*glmU*	Rv1018c	cell wall and cell processes

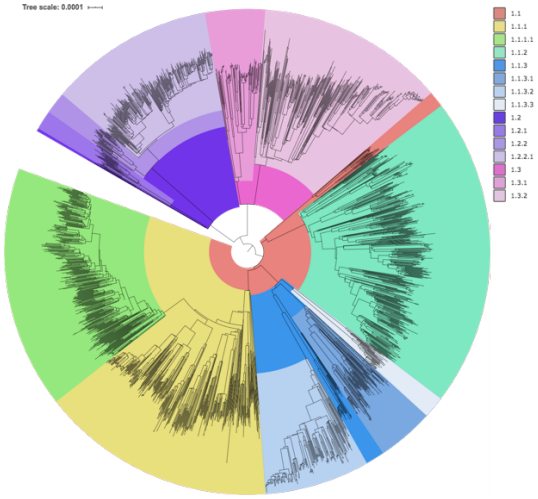
8	**221190**	G->T	178V	*ilvD*	Rv0189c	IMR
9	**2750052**	G->A	58T	*valS*	Rv2448c	information pathways
M.bovis	**62768**	A->G	791G	*dnaB*	Rv0058	information pathways
M.caprae	**904090**	T->C	122G	*purM*	Rv0809	IMR
M.orygis	**44812**	G->T	417P	*leuS*	Rv0041	information pathways

All essential genes (except *) with synonymous changes; Bolded - different from reference [1]; IMR = Intermediary metabolism and respiration; VDA = virulence, detoxification, adaptation

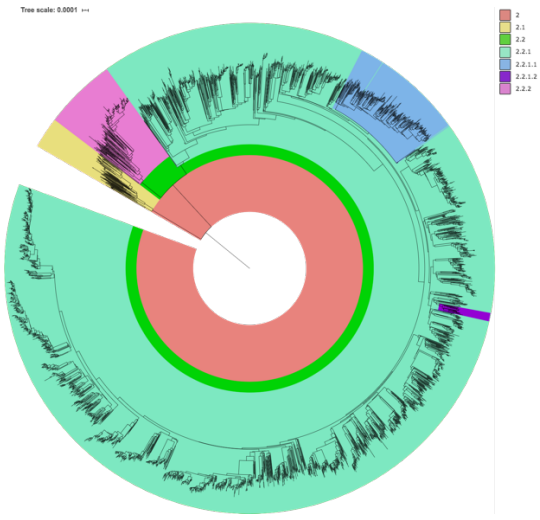
(a) Principal component (PC) analysis plot for lineages 1-9, *M. bovis, M. caprae, M. orygis* (90 sub-lineages; N=35,298)



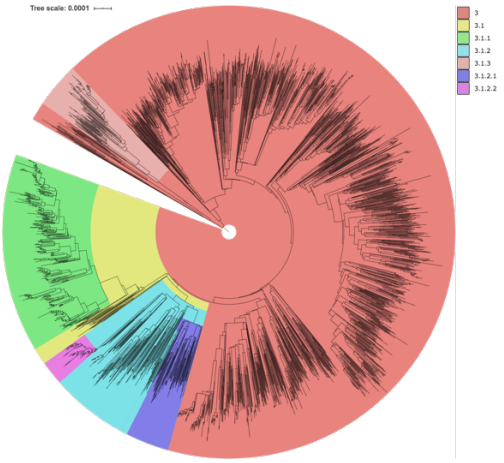
(b) Lineage 1 (16 sub-lineages; N=2,162*)



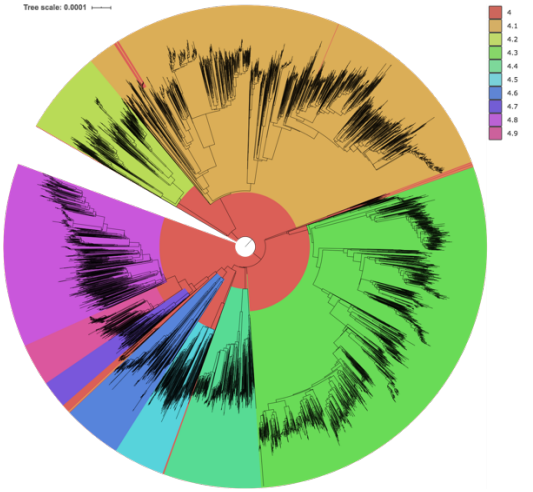
(c) Lineage 2 (7 sub-lineages; N=4,456*)



(d) Lineage 3 (7 sub-lineages; N=2,654*)



(e) Lineage 4 (52 sub-lineages; N=8,320*)



(f) Lineages 5-9, *M.bovis, caprae,* and *orygis**

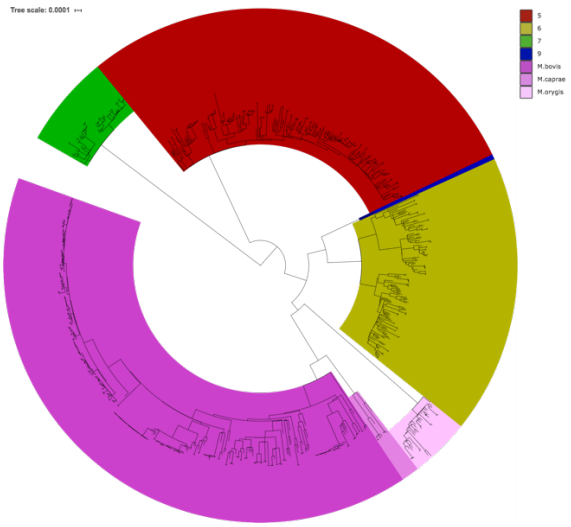
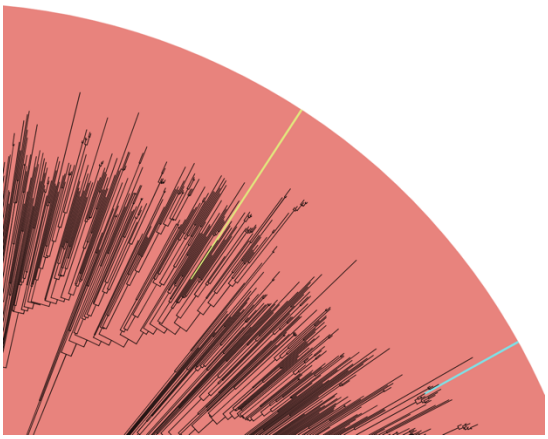
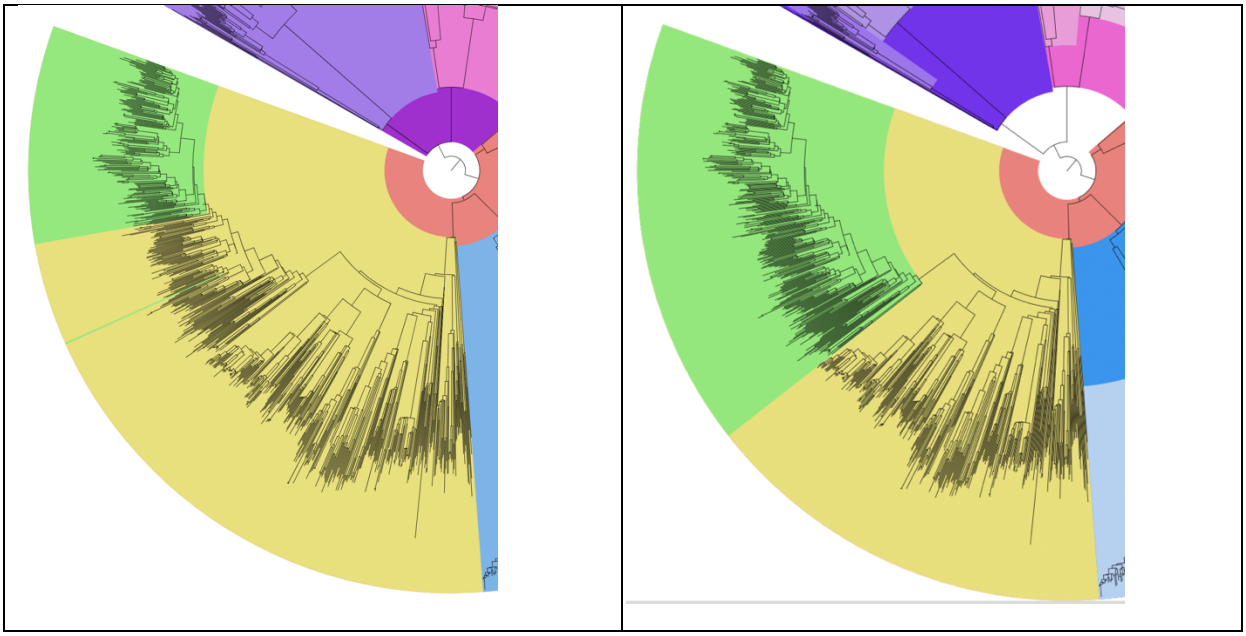


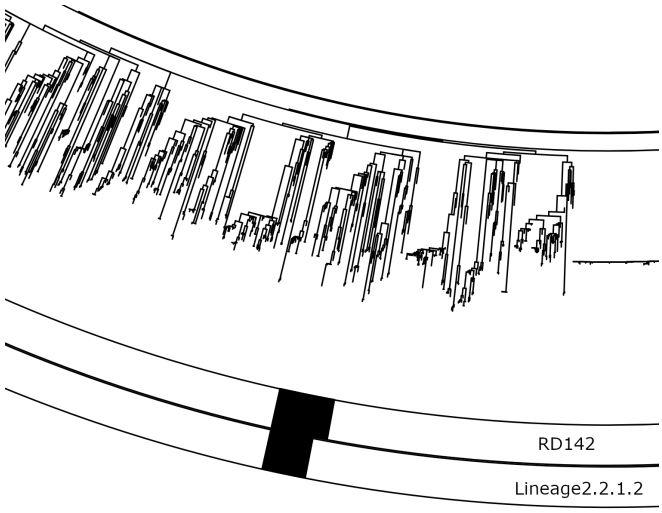
Fig S1: Population structure of the *Mycobacterium tuberculosis* complex isolates by lineage; * training



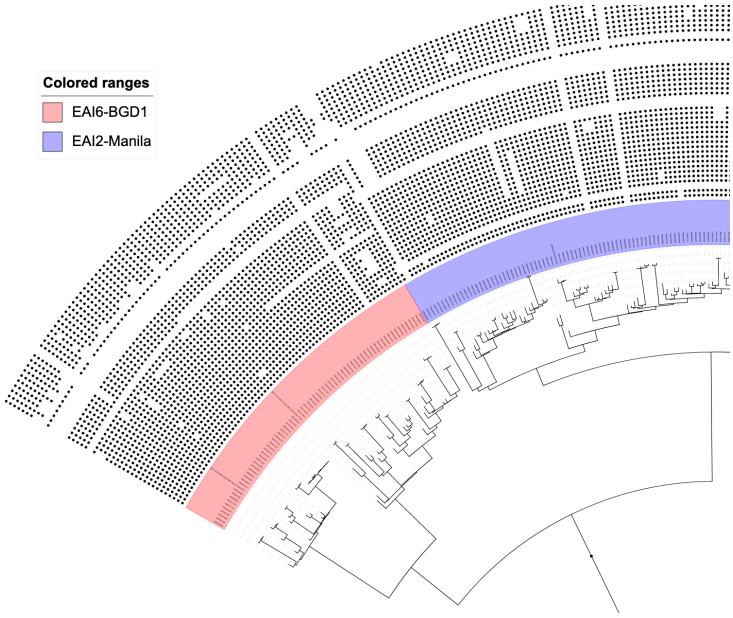
(a) Samples labelled as belonging to 3.1 (yellow) and 3.1.2 (blue) are embedded in the parental clade 3 (red); these samples were therefore re-assigned to the parent clade 3.



(b) Samples originally labelled 1.1.1.1 (green, left) are within a clade deeper in the tree than they ought to be. The long central branch in the middle of the 1.1.1 (yellow) parent clade suggests a more natural position for this lineage, as shown in (right)



(c) Lineage 2.2.1.2 was defined to reflect the strains containing the RD142 deletion. The original barcode SNP did not capture all strains with this RD and needed to be redefined



(d) Lineage 1.2.1 encapsulated a large clade with low imbalance and late diversification timing, which was further partitioned into lineage 1.2.1 and 1.2.2, which now reflect members of the EAI6-BGD1 and EAI2-Manila/EAI2-Nonthaburi families, respectively

Fig S2 (a)-(d): Examples of discrepancies using the 62-SNP barcode

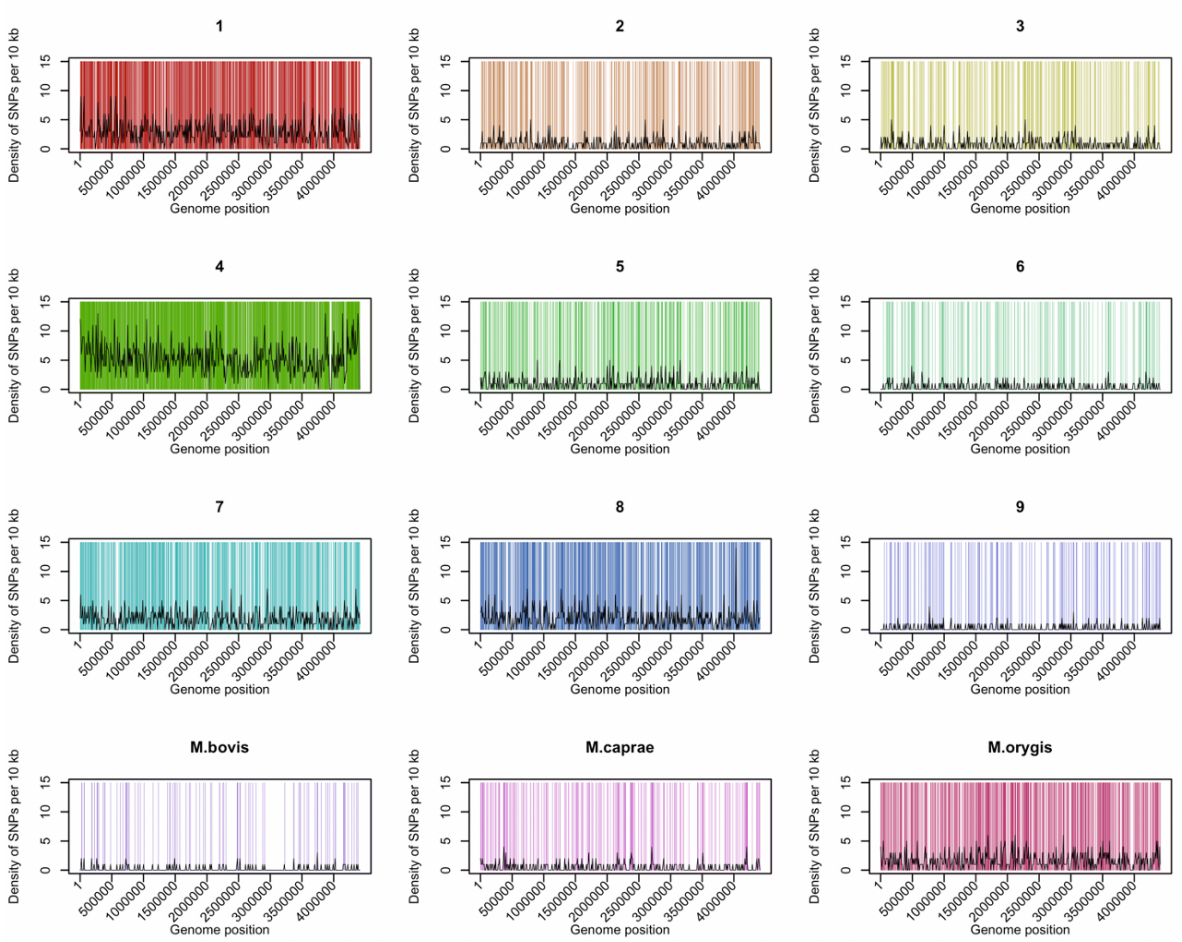


Fig S3: The genome-wide distribution and density (per 10kb) of barcoding SNPs (F_{ST}=1) for each lineage

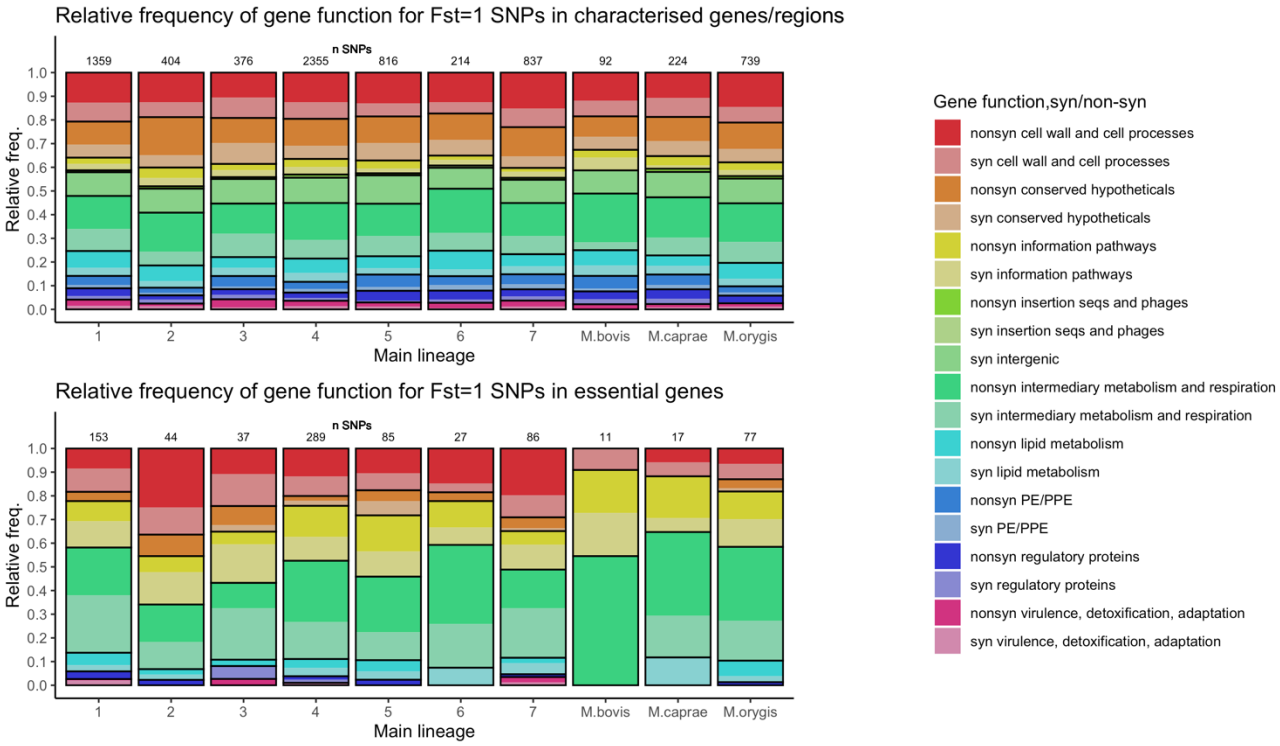


Fig S4: Functional differences between genes containing lineage-barcoding (F_{ST}=1) SNPs*

* Lineages 8 (N=2) and 9 (N=3) have been excluded due to their small sample sizes

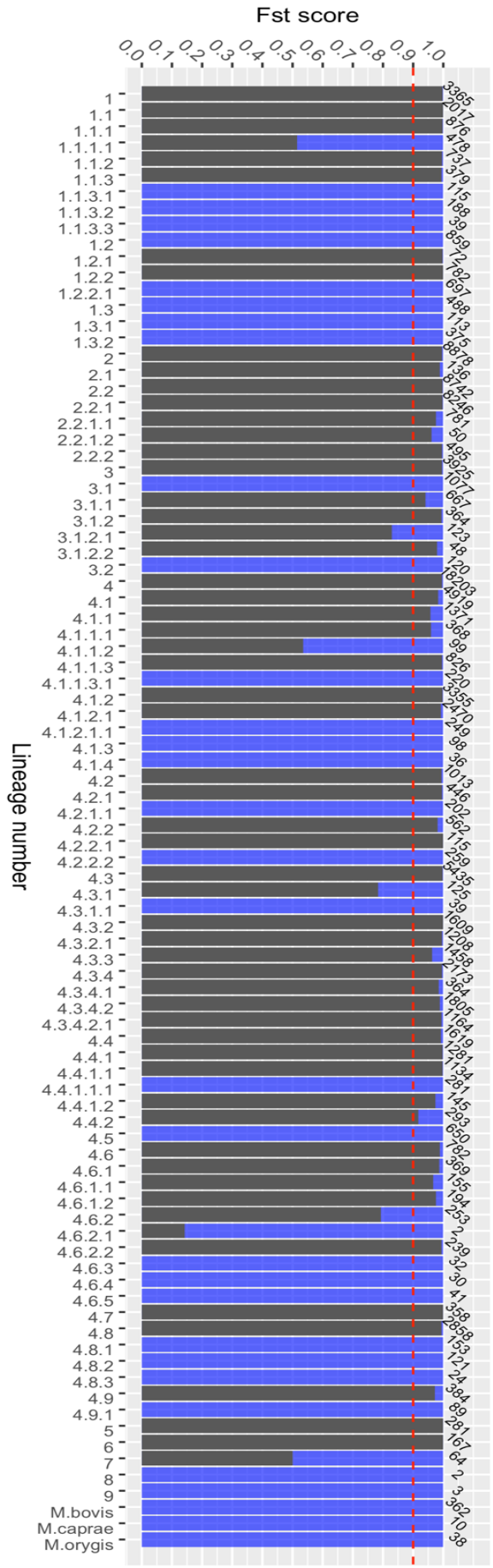


Fig S5: Differentiation of sub-lineages* when comparing the 62- versus 90-SNP barcodes

Blue is the incremental improvement in the new 90-SNP barcode; * based on the F_{ST} measure where values of 1 mean that the barcoding SNPs perfectly differentiate the sub-lineage (versus any other); sub-lineage 4.5 is entirely blue, because we identified a new barcoding SNP which did not lie within a genomic region containing a common deletion

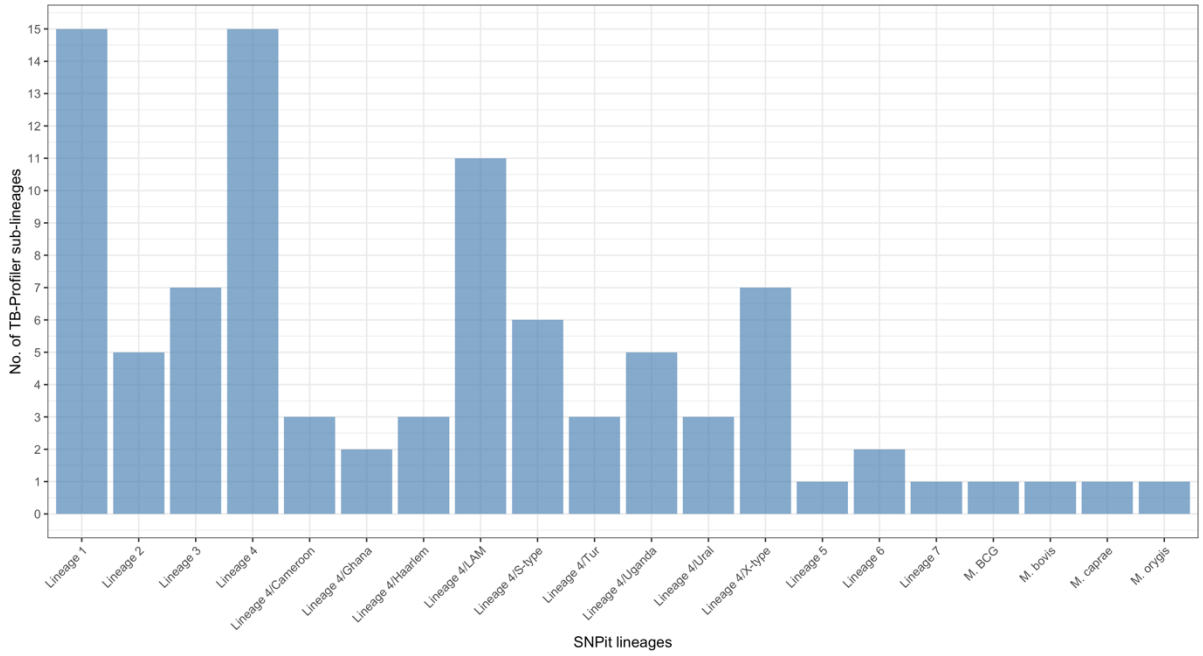


Fig S6: The increased resolution of our 90-SNP barcode (implemented in *TB-Profiler* software) over the comparable (sub-)lineages of the *SNP-IT* tool

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

London School of Hygiene & Tropical Medicine
Keppel Street, London WC1E 7HT

T: +44 (0)20 7299 4646
F: +44 (0)20 7299 4656
www.lshtm.ac.uk

RESEARCH PAPER COVER SHEET

Please note that a cover sheet must be completed for each research paper included within a thesis.

SECTION A – Student Details

Student ID Number	1807750	**Title**	Mr
First Name(s)	Gary		
Surname/Family Name	Napier		
Thesis Title	Using whole genome sequencing data to identify strain-types, transmission enhancers and novel drug resistance mutations of Mycobacterium tuberculosis		
Primary Supervisor	Prof. Taane G. Clark		

If the Research Paper has previously been published please complete Section B, if not please move to Section C.

SECTION B – Paper already published

Where was the work published?			
When was the work published?			
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion			
Have you retained the copyright for the work?*	Choose an item.	Was the work subject to academic peer review?	Choose an item.

*If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	Scientific Reports
Please list the paper's authors in the intended authorship order:	Gary Napier, David Couvin, Guislaine Refrégier, Christophe Guyeux, Conor Meehan, Christophe Sola, Susana Campino, Jody E. Phelan, Taane G. Clark
Stage of publication	Submitted

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	I performed the bioinformatic and statistical analysis, and wrote the first draft of the manuscript. I worked with co-authors on subsequent drafts and finalisation of the paper for submission.

SECTION E

Student Signature	
Date	30/09/2022

Supervisor Signature	
Date	22/09/2022

Chapter 3

Comparison of *in silico* prediction of *Mycobacterium tuberculosis* spoligotypes and lineages from whole genome sequences

Comparison of *in silico* predicted *Mycobacterium tuberculosis* spoligotypes and lineages from whole genome sequencing data

Gary Napier[1] gary.napier@lshtm.ac.uk
David Couvin[2] dcouvin@pasteur-guadeloupe.fr
Guislaine Refrégier[3] guislaine.refregier@universite-paris-saclay.fr
Christophe Guyeux[4] guyeux@gmail.com
Conor Meehan[5] conor.meehan@ntu.ac.uk
Christophe Sola[3] christophe.sola@universite-paris-saclay.fr
Susana Campino[1] susana.campino@lshtm.ac.uk
Jody Phelan[1,*] jody.phelan@lshtm.ac.uk
Taane G. Clark[1,6,*] taane.clark@lshtm.ac.uk

1. Faculty of Infectious and Tropical Diseases, London School of Hygiene & Tropical Medicine, WC1E 7HT London, UK
2. Institut Pasteur de la Guadeloupe, Université Antilles Guyane. Les Abymes, Guadeloupe.
3. CEA, CNRS, Univ. Paris-Sud, Université Paris-Saclay, cedex, 91198, Gif-sur-Yvette, France
4. DISC Computer Science Department, Univ. Bourgogne Franche-Comté (UBFC), 16 Route de Gray, 25000, Besançon, France
5. Nottingham Trent University, NG1 4FQ Nottingham, UK
6. Faculty of Epidemiology and Population Health, London School of Hygiene & Tropical Medicine, WC1E 7HT London, UK

* joint authors

Scientific Reports

Abstract

Mycobacterium tuberculosis complex (MTBC) bacterial strain types underlie tuberculosis disease, and have been associated with drug resistance, transmissibility, virulence, and host-pathogen interactions. Spoligotyping was developed as a molecular genotyping technique used to determine strain types, though recent advances in sequencing technology have led to their characterization using SNP-based sublineage nomenclature. Notwithstanding, spoligotyping remains an important tool and there is a need to characterise the concordance of resulting spoligotypes with sublineages. To achieve this, an *in silico* spoligotype prediction tool ("Spolpred2") was developed and integrated into TB-Profiler software. Lineage and spoligotype predictions were generated for >32k isolates and the overlap between strain types was characterised. Major spoligotype families detected were Beijing (24.7%), T (18.4%), LAM (13.0%), CAS (9.2%), and EAI (7.7%), and these broadly followed known geographic distributions. Most spoligotypes were

perfectly correlated with the main MTBC lineages (L1-L7, plus animal). Conversely, at lower levels of the sublineage system, the relationship breaks down, with only <50% of spoligotypes being perfectly associated with a sublineage at the second or subsequent levels of resolution. Whilst the SNP-based sublineage system may represent a higher resolution system to characterise strain diversity, some spoligotypes (e.g., EAI2-Manila, EAI2-nonthaburi) provide historical and fine-scale insights, and are accessible with the software developed.

Word count: 200/200

Keywords: *Mycobacterium tuberculosis*, spoligotypes, lineages, phylogeny

INTRODUCTION

Tuberculosis is an infectious disease of high global burden caused by members of the *Mycobacterium tuberculosis* complex (MTBC), which includes *M. tuberculosis sensu stricto* (*Mtb*), *M. africanum* and animal strains such as *M. bovis*. Though the MTBC is described as clonal, there is sufficient genetic variation to distinguish strain types within members of the complex. *Mtb* and *M. africanum* are phylogenetically classified in nine main lineages (L1-9), with strain types that are distributed phylo-geographically[1]. Strain identification is crucial to addressing key epidemiological questions, from individual to global scales. Strain typing is informative in the investigation of transmission events and, in the wider context, provides valuable insight into the spread of MTBC variants, indicating potential differences between genotypes and phenotypes. For example, Beijing strains show lineage-specific associations with drug resistance[2], and geographical ubiquity of lineages 2 (Beijing) and 4 (Euro-American-Indian; EAI) can be attributed to virulence and transmissibility[3]. Furthermore, strain typing at higher phylogenetic resolution can reveal within-strain differences, such as between typical and atypical Beijing strains, which vary in geographical distribution, resistance, and virulence[4,5]. Advances in sequencing technologies are leading high-resolution strain typing offered by whole genome sequencing (WGS) data, which improve inference in transmission studies and enable the tracking of between- and within-lineage genotypic-phenotypic differences, as well as assisting with understanding drug resistance mechanisms.

Spoligotyping is a fingerprinting PCR technique[6], which exploits the polymorphism harboured at the direct repeat locus of MTBC. It is based on the PCR amplification of 43 short unique sequences (termed spacers) contained between well-conserved 36-bp direct repeats. Since strains vary in the occurrence of spacers, each sample produces a distinctive spot pattern, which is then translated into a numerical code of 8 digits, leading to >3,800 spoligotypes. The spoligotyping nomenclature[7] reflects the phylogeographical structure of MTBC (e.g. “Beijing”, “Latin-American-Mediterranean”; **Table 1**), and its main families overlap with a SNP-based barcoding system[8], which was recently updated[1]. However, the overwhelming majority of distinct spoligotypes remain uncategorised (“Unknown”) as a result of noise in the spacer patterns (**Table 1**, **Figure S3**). Both spoligotypes and sublineages offer higher resolution than large deletion-based regions of difference (RD). However, the full extent of alignment between spoligotypes and sublineages needs to

be established, potentially leading to improvements in both spoligotyping and sublineage barcoding of strain types using WGS data. Although SNP-based strain typing is more prevalent with the advent of WGS, it is important to maintain concordance with the spoligotyping system for purposes of backwards compatibility and continuity with older studies. Previous work has *in silico* predicted spoligotypes, implemented within the widely applied SpolPred software[9]. With at least 20-fold more *Mtb* WGS data available since its development, we seek to assess the consistency of spoligotypes with the sublineage system[1], and determine their global distribution. This goal is achieved by developing new software to *in silico* genotype isolates, called "Spolpred2", which predicts spoligotypes from raw sequence reads generated by several technological platforms. We incorporate the updated barcodes for spoligotypes, and imbed "Spolpred2" within the TB-Profiler tool[10], widely used to profile MTBC sublineages, strain types, and drug resistance from WGS for clinical and surveillance applications.

RESULTS

Global distribution of spoligotypes families

The dataset consisted of 32,632 *M. tuberculosis* isolates with WGS, drug susceptibility test and geographical source data, with lineages inferred using TB-Profiler software (**Table 1**). The spoligotypes were predicted using new Spolpred2 software, developed as part of this work (see **MATERIALS AND METHODS**) (**Table 1**). Most isolates were from the main lineages (L4 51.1%, L2 25.3%, L3 11.4%, L1 9.7%), and the major spoligotype families identified were Beijing (L2; 24.7%), T (336 spoligotypes; L4; 18.4%), LAM (212 spoligotypes; L4; 13.0%), Central Asian Strain (L3; CAS; 133 spoligotypes; 9.2%), EAI (212 spoligotypes; 7.7%), though many samples had an unknown family (n=4,276, 13.1%). A total of 114 unique (sub-)lineages and 3,817 unique spoligotypes were present. Whilst the isolates represent a convenience sample, they covered all World Health Organization (WHO) Regions, including Europe (36 countries; 33.8%), Africa (29 countries; 23.0%), Western Pacific (8 countries; 13.5%) and the Americas (14 countries; 10.9%).

To improve the stringency of the analysis, all spoligotypes with <5 isolates support were removed, resulting in 27,933 (85.6%) isolates, 105 (92.1%) unique lineages and 452 (11.8%) distinct spoligotypes (**Table 1; Figure 1**). This filtering task reveals the high number of rare spoligotypes (n=3,365; see **S1 Table** for a list),

with representation across most lineages (L4 58.4%; L1 16.7%; L3 16.4%; other 8.6%). After filtering (n=27,933), the most frequent spoligotype families were Beijing (8,023; 28.7%), followed by T (5,589; 20%) and LAM (3,961; 14.2%), consistent with pre-filtering, but the proportion with unknown family decreased (n=1,174; 4.2%) (**Figure 1**, **S2 Table**). The most common WHO geographical regions were Europe (n=9,063; 32.5%), Africa (n=6,795; 24.3%) and Western Pacific (n=4,008; 14.3%) (**Figure 1**, **Figure S2**), also consistent with pre-filtered data. While many isolates occur in their expected geographical regions, such as Beijing strains in Western Pacific and Southeast Asia, there is high variation in the source, reflecting the spread of *Mtb* since spoligotype labels were conceived, and the convenience nature of the sampling, which includes an emphasis on transmission studies or clinically relevant investigations.

Spoligotype families and lineages

There was a strong concordance between spoligotype family and main lineage among the 27,933 samples (**S3 Table**). At the lineage level (L1 - L7), there were 445 (96.5%) spoligotypes appearing exclusively in their respective lineages. For example, the AFRI family only appears in isolates classed as lineages 5 and 6. EAI, CAS, and Ethiopian families are exclusively within lineages 1, 3, and 7 respectively. Similarly, Cameroon, Ethiopian, H, LAM, S, T, Turkey, and URAL spoligotype families appear only in lineage 4, consistent with it being the most genetically diverse lineage (**Figure 2**). There were however a few discrepancies, such as a very small proportion of isolates with a Beijing spoligotype family being classified as lineage 1 (n=1) or 3 (n=21) (22/8023; <0.3%) (**S3 Table**). These discrepancies could not be explained by low coverage in the direct repeat region. Isolates with the Manu spoligotype family straddled lineage 2 (n=43; 40.2%) with the Manu ancestor types, as well as lineage 3 (n=64; 59.2%) corresponding to a Manu3 strain type. Whilst many spoligotypes were found to be exclusive to lineages at each level, in many cases they nevertheless appeared in a relatively small proportion of that lineage's total samples (**S3 Table**). For example, spoligotype EAI2-nonthaburi is only found in lineage 1 but appears in only 17.2% of that lineage's total samples, and is known to be localised to Thailand[11]. EAI2-nonthaburi is similar to the EAI-Manila spoligotype, originally found in the Philippines, and a dominant strain-type in that country[12]. Conversely, as shown above, there are spoligotypes such as Beijing which are highly prevalent in lineage 2 (91.2%) but appear in two other lineages

(**S2 Table**).

Subsequent analysis looked at spoligotypes within secondary, tertiary, and subsequent levels of lineages. At lower levels of sublineages, there were decreasing numbers in perfect concordance with spoligotypes (second level (e.g., L4.2): n = 324 (46.1%); third level (e.g., L4.2.2): n = 310 (39.5%); fourth level (e.g., L4.2.2.1): 289 (33.9%)) (**Figure S1**). The spoligotypes which offered discrimination between lineages at the lowest lineage level and with at least 20 isolates, included EAI2-Manila and EAI2-nonthaburi (L1.2.1.2), Manu ancestor (L2.1), T4-CEU1 (L4.1.2), Turkey (L4.2.2.1), LAM1 and LAM2 (L4.3.4.1), T2-Uganda (L4.6.1.1), and BOV_3 (La1.8.1) (**S4 Table**), which could be used to potentially update the lineage SNP barcode.

DISCUSSION

This work aimed to characterise the global distribution of spoligotypes and correlate this with the lineage system developed previously[1]. To enable this work, a new rapid *in silico* spoligotyping software was developed with speed and flexibility in mind and was integrated into the TB-Profiler analysis platform.

The frequency of spoligotypes and their respective families confirm known common spoligotype families with representation from Beijing, T, LAM, CAS and EAI. The geographic distribution of spoligotype families followed known patterns with T and LAM being most prevalent in Europe, Africa and the Americas, Africanum in West Africa, and Beijing being prevalent across most geographic regions. Interestingly, there were 3,490 spoligotypes that were present in <5 isolates, and may represent either very rare combinations of spacers or isolates with low or unstable coverage around the direct repeat locus. Of the 3,490 rare spoligotypes, 1,140 (32.7%) had been previously seen in the SITVIT2 database, indicating that these are indeed valid spoligotypes, albeit rare. The remaining spoligotypes could represent truly novel spoligotypes or have been generated from samples with spurious coverage. Generally, there was a strong association of spoligotypes to lineage with the majority of spoligotype families associated exclusively to one of the major lineages. The only discrepancies found were twenty-two isolates (twenty-one L3; one L1), assigned as members of a Beijing spoligotype family. These spoligotypes were manually verified, which ruled out poor data quality and confirmed the Beijing spoligotype. As expected, the perfect concordance between spoligotype and lineage diminished as higher resolution sublineages were used for comparison, with only

33.9% of spoligotypes showing perfect concordance at the finest scale of sublineages (4th level). This observation indicates that the spoligotypes are not monophyletic, could have arisen through convergent evolution, or that the sublineage comprises a higher resolution unit than the respective spoligotype(s). Conversely, there were some instances where a sublineage contained multiple major spoligotypes (e.g., EAI2-Manila and EAI2-nonthaburi, both lineage 1.2.1.2), and hence the spoligotype represents the higher resolution unit for the related samples. In these cases, the sublineage system and corresponding SNP-barcode could be further refined to reflect this diversity. These can be explored further through growing WGS datasets, and applications of phylogenetic analysis and *in silico* strain typing using updated TB-Profiler software.

CONCLUSIONS

We have presented a method to predict *in silico* spoligotypes from WGS, called "Spolpred2", which is fast and accurate. This software is freely available as part of the TB-Profiler package. Spoligotypes are useful in tracking the epidemiological spread of MTBC, but do not necessarily agree with the lineage system at lower resolution. We have clarified this relationship, which adds to the power of using a dual approach to strain typing.

MATERIAL AND METHODS

Sequence data and processing

The input dataset consists of 32,632 isolates for which next generation sequences have been deposited on the ENA, and have been previously described elsewhere[10]. All sequence data was aligned to the H37Rv reference genome (NC_000962.3) using BWA mem software (v0.7.17). Variants were called using GATK HaplotypeCaller (v4.1.4.1 -ERC GVCF) and merged using the GATK CombineGVCFs tool. Variants were filtered to remove indels, SNPs in *pe/ppe* genes and those which had >10% missing genotypes across isolates. Filtered variants were transformed to a multi-fasta format file, which was subsequently used as the input to phylogenetic reconstruction by iqtree software (v2.1.2 -m GTR+G+ASC). Lineage assignments were generated using TB-Profiler (v4.3.0). Alignment files in bam format were used for spoligotype generation using the algorithm described below.

Spolpred2 algorithm

The Spolpred2 spoligotype prediction tool is based on k-mer counting. The KMC3 tool[13] is used to count k-mers from either raw fastq, fasta, bam or cram format. A k-mer length equal to the length of the unique spacers (k=25) is chosen. For bam and cram files, alignment against the H37Rv reference genome (AL123456.3)[14] is assumed and only reads falling between positions 3117003 and 3127206 are analysed. A custom Python script then loads the counts and performs a direct look-up of the spacers, accounting for up to two mismatches. The presence or absence of a spacer is determined by comparing the counts against a minimum threshold. The threshold is selected to be 20% of the maximum spacer count. The presence/absence vector represents the binary spoligotype and is converted into an octal form. Finally, the associated family and SIT are reported by performing a look-up in a CSV file, which currently contains data for all samples submitted to SITVIT2[15]. The code was integrated into TB-Profiler (v4.3.0)[10] and can be invoked to perform spoligotyping only, or as part of the standard profiling pipeline, which also reports drug resistance and SNP-based lineage. Using a standard laptop with 8 Gb ram, Spolpred2 can profile from bam and fasta format files with 1000-fold coverage in <10 seconds, whilst perform the same task on raw fastq files with up to 500-fold coverage in <30 seconds.

Association of spoligotypes to lineages

Spoligotypes were inferred using Spolpred2 software across 32,632 MTBC samples with whole genome sequencing data, location, and drug resistance phenotypes. Lineages and sublineages were inferred using the TB-Profiler tool, which implements a published barcode[1]. The number of (sub-)lineages within spoligotype families was estimated. As there were many spoligotypes in low numbers of samples and therefore offering little predictive power, those appearing in <5 isolates were excluded. Since we were interested in the strength of association between spoligotypes and the various levels of the lineage system[1], the lineages were parsed into a hierarchy for each round of analysis. For example, the first level analysed the association between each spoligotype and the main *Mtb* lineages 1-7. Next was the association between each spoligotype and the second level, represented by lineages 1.1, 1.2, 2.1, 2.2, and so on. A concordance correlation coefficient was used to test the statistical strength of association, where a score of 1 is assigned if a spoligotype is unique to a given (sub-)lineage, and anything less than 1 indicates that the spoligotype is

found in at least one other isolate belonging to another (sub-)lineage.

DATA AVAILABILITY

All data used in this work is publicly available. Spolpred2 software is available as part of TB-Profiler (https://github.com/jodyphelan/TBProfiler), but also stand-alone (https://github.com/GaryNapier/spolpred).

ACKNOWLEDGEMENTS

GN is funded by an BBSRC-LIDo PhD studentship. JEP is funded by a Newton Institutional Links Grant (British Council, no. 261868591). TGC is funded by the Medical Research Council UK (Grant no. MR/M01360X/1, MR/N010469/1, MR/R025576/1, and MR/R020973/1). SC is funded by Medical Research Council UK grants (ref. MR/M01360X/1, MR/R025576/1, and MR/R020973/1). The authors declare no conflicts of interest. The funders had no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript should be declared.

AUTHORS CONTRIBUTIONS

JEP and TGC conceived and directed the project. GN performed bioinformatic and statistical analyses under the supervision of SC, JEP and TGC. DC, GR, CG, CM and CS provided resources. GN, CM, SC, JEP and TGC interpreted results. GN wrote the first draft of the manuscript with inputs from JEP and TGC. All authors commented and edited on various versions of the draft manuscript and approved the final version. GN, JEP, and TGC compiled the final manuscript.

ADDITIONAL INFORMATION

Ethics approval and consent

No ethics approvals were required as all data is publicly available.

Consent for publication

All authors have consented to the publication of this manuscript.

Availability of data and materials

Analysis scripts are available at https://github.com/GaryNapier/spolpred

Competing interests

Authors declare no competing interests.

REFERENCES

1. Napier, G. *et al.* Robust barcoding and identification of Mycobacterium tuberculosis lineages for epidemiological and clinical studies. *Genome Med.* **12**, 114 (2020).

2. Oppong, Y. E. A. *et al.* Genome-wide analysis of Mycobacterium tuberculosis polymorphisms reveals lineage-specific associations with drug resistance. *BMC Genomics* **20**, (2019).

3. Coscolla, M. & Gagneux, S. Consequences of genomic diversity in mycobacterium tuberculosis. *Seminars in Immunology* vol. 26 431–444 (2014).

4. Forrellad, M. A. *et al.* Virulence factors of the mycobacterium tuberculosis complex. *Virulence* vol. 4 3–66 (2013).

5. Ribeiro, S. C. M. *et al.* Mycobacterium tuberculosis Strains of the Modern Sublineage of the Beijing Family Are More Likely To Display Increased Virulence than Strains of the Ancient Sublineage. *J. Clin. Microbiol.* **52**, 2615 (2014).

6. Kamerbeek, J. *et al.* Simultaneous detection and strain differentiation of Mycobacterium tuberculosis for diagnosis and epidemiology. *J. Clin. Microbiol.* **35**, 907–914 (1997).

7. Brudey, K. *et al.* Mycobacterium tuberculosis complex genetic diversity: Mining the fourth international spoligotyping database (SpolDB4) for classification, population genetics and epidemiology. *BMC Microbiol.* **6**, 1–17 (2006).

8. Coll, F. *et al.* A robust SNP barcode for typing Mycobacterium tuberculosis complex strains. *Nat. Commun.* **5**, 4812 (2014).

9. Coll, F. *et al.* SpolPred: rapid and accurate prediction of Mycobacterium tuberculosis spoligotypes from short genomic sequences. *Bioinformatics* **28**, 2991–2993 (2012).

10. Phelan, J. E. J. E. J. E. *et al.* Integrating informatics tools and portable sequencing technology for rapid detection of resistance to anti-tuberculous drugs. *Genome Med.* **11**, 41 (2019).

11. Palittapongarnpim, P. *et al.* Evidence for Host-Bacterial Co-evolution via Genome Sequence Analysis of 480 Thai Mycobacterium tuberculosis Lineage 1 Isolates. *Sci. Reports 2018 81* **8**, 1–14 (2018).

12. Phelan, J. E. *et al.* Mycobacterium tuberculosis whole genome sequencing provides insights into the Manila strain and drug-resistance mutations in the Philippines. *Sci. Rep.* **9**, (2019).

13. Kokot, M., Dlugosz, M. & Deorowicz, S. KMC 3: counting and manipulating k-mer statistics. *Bioinformatics* **33**, 2759–2761 (2017).

14. Cole, S. T. *et al.* Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. *Nature* **393**, 537–544 (1998).

15. Couvin, D., David, A., Zozio, T. & Rastogi, N. Macro-geographical specificities of the prevailing tuberculosis epidemic as seen through SITVIT2, an updated version of the Mycobacterium tuberculosis genotyping database. *Infect. Genet. Evol.* **72**, 31–43 (2019).

FIGURE LEGENDS

Figure 1

Spoligotype families and number of samples (a) by lineages; (b) across WHO regions

Figure 2

Spoligotype families and Lineages (n=27,933)

TABLES

Table 1

***Mycobacterium tuberculosis* dataset**

Characteristic		[# members]*	n=32,632	%	n=27,933**	%
Lineage	1	15	3,149	9.7	2,335	8.4
	2	5	8,241	25.3	8,102	29.0
	3	7	3,734	11.4	3,009	10.8
	4	52	16,679	51.1	13,923	49.8
	5	9	253	0.8	169	0.6
	6	10	148	0.5	90	0.3
	7	1	52	0.2	44	0.2
	9	1	3	0.0	-	-
	La	14	373	1.1	261	0.9
Spoligotype	Beijing	32	8,056	24.7	8,023	28.7
	T	336	6,000	18.4	5,589	20.0
	Unknown	2,548	4,276	13.1	1,174	4.2
	LAM	212	4,229	13.0	3,961	14.2
	CAS	133	3,005	9.2	2,842	10.2
	EAI	212	2,509	7.7	2,221	8.0
	X	62	1,147	3.5	1,080	3.9
	H	99	1,032	3.2	892	3.2
	S	36	908	2.8	862	3.1
	Ural	41	491	1.5	443	1.6
	BOV	39	297	0.9	244	0.9
	Other	67	682	2.1	602	2.2
WHO region	Europe	36	11,016	33.8	9,063	32.4
	Africa	29	7,520	23.0	6,795	24.3
	Western Pacific	8	4,412	13.5	4,008	14.3
	Americas	14	3,557	10.9	3,202	11.5
	Unknown	1	3,349	10.3	2,833	10.1
	South-East Asia	7	2,049	6.3	1,726	6.2
	Eastern Mediterranean	11	729	2.2	306	1.1

* sublineage, spoligotype, or number of countries; **excludes isolates with spoligotypes with freq. <5; H = Haarlem; LAM = Latin-American-Mediterranean; EAI = East-African-Indian; CAS = Central Asia

Table 2

Spoligotypes with (sub-)lineages for *Mycobacterium tuberculosis* (n=27,933)

Lineage	Sublineage	No. (%)	No. spoligotypes (%)*	No. families (%)
1	Overall	2,335 (8.4)	72 (15.6)	12 (16.9)
	1.1	1,342 (4.8)	49 (7.0)	9 (5.2)
	1.2	993 (3.6)	33 (4.7)	8 (4.6)
2	Overall	8,102 (29)	23 (5.0)	3 (4.2)
	2.1	96 (0.3)	5 (0.7)	2 (1.2)
	2.2	8,006 (28.7)	18 (2.6)	2 (1.2)
3	Overall	3,009 (10.8)	67 (14.5)	7 (9.9)
	3	2,006 (7.2)	58 (8.3)	5 (2.9)
	3.1	1,003 (3.6)	31 (4.4)	7 (4.0)
4	Overall	13,923 (49.8)	267 (57.9)	38 (53.5)
	4	124 (0.4)	16 (2.3)	4 (2.3)
	4.1	3,865 (13.8)	103 (14.7)	21 (12.1)
	4.2	798 (2.9)	42 (6.0)	13 (7.5)
	4.3	4,263 (15.3)	84 (11.9)	17 (9.8)
	4.4	1,271 (4.6)	40 (5.7)	9 (5.2)
	4.5	426 (1.5)	42 (6.0)	14 (8.1)
	4.6	517 (1.9)	32 (4.6)	9 (5.2)
	4.7	280 (1)	19 (2.7)	9 (5.2)
	4.8	2,083 (7.5)	67 (9.5)	16 (9.2)
	4.9	296 (1.1)	27 (3.8)	11 (6.4)
5	Overall	169 (0.6)	16 (3.5)	3 (4.2)
	5	5 (0)	1 (0.1)	1 (0.6)
	5.1	137 (0.5)	12 (1.7)	3 (1.7)
	5.2	19 (0.1)	2 (0.3)	1 (0.6)
	5.3	8 (0)	1 (0.1)	1 (0.6)
6	Overall	90 (0.3)	4 (0.9)	2 (2.8)
	6	3 (0)	2 (0.3)	1 (0.6)
	6.1	14 (0.1)	2 (0.3)	1 (0.6)
	6.2	25 (0.1)	2 (0.3)	1 (0.6)
	6.3	48 (0.2)	3 (0.4)	2 (1.2)
7	Overall	44 (0.2)	3 (0.7)	2 (2.8)
La	Overall	261 (0.9)	9 (2.0)	4 (5.6)
	La1	244 (0.9)	8 (1.1)	3 (1.7)
	La3	17 (0.1)	1 (0.1)	1 (0.6)

*Number of spoligotypes duplicated on some occasions due to presence in multiple lineages/sublineages.

SUPPLEMENTARY TABLES

S1 Table

Spoligotypes with frequency <5 isolates

S1_Table.csv

S2 Table

Distribution of (sub-)lineages within spoligotype families

S2_Table.csv

S3 Table

Spoligotype families within the main lineages (L) (n=27,933)

Family	L1	L2	L3	L4	L5	L6	L7	La	Total
AFRI	0	0	0	0	**109**	**83**	0	0	192
Beijing	1	**8,001**	21	0	0	0	0	0	8,023
BOV	0	0	0	0	0	0	0	**244**	244
Cameroon	0	0	0	**181**	0	0	0	0	181
CAS	0	0	**2,842**	0	0	0	0	0	2,842
EAI	**2,221**	0	0	0	0	0	0	0	2,221
Ethiopian	0	0	0	0	0	0	**25**	0	25
H	0	0	0	**892**	0	0	0	0	892
LAM	0	0	0	**3,961**	0	0	0	0	3,961
Manu 3	0	0	**64**	0	0	0	0	0	64
Manu ancestor	0	**43**	**0**	0	0	0	0	0	43
S	0	0	0	**862**	0	0	0	0	862
T	0	0	0	**5,589**	0	0	0	0	5,589
Turkey	0	0	0	**97**	0	0	0	0	97
Unknown	**113**	**58**	**82**	**818**	**60**	**7**	**19**	**17**	1,174
Ural	0	0	0	**443**	0	0	0	0	443
X	0	0	0	**1,080**	0	0	0	0	1,080
Total	2,335	8,102	3,009	13,923	169	90	44	261	27,933

Bolded are frequent lineages for each spoligotype family

S4 Table

Spoligotypes discriminating lineages at the lowest level

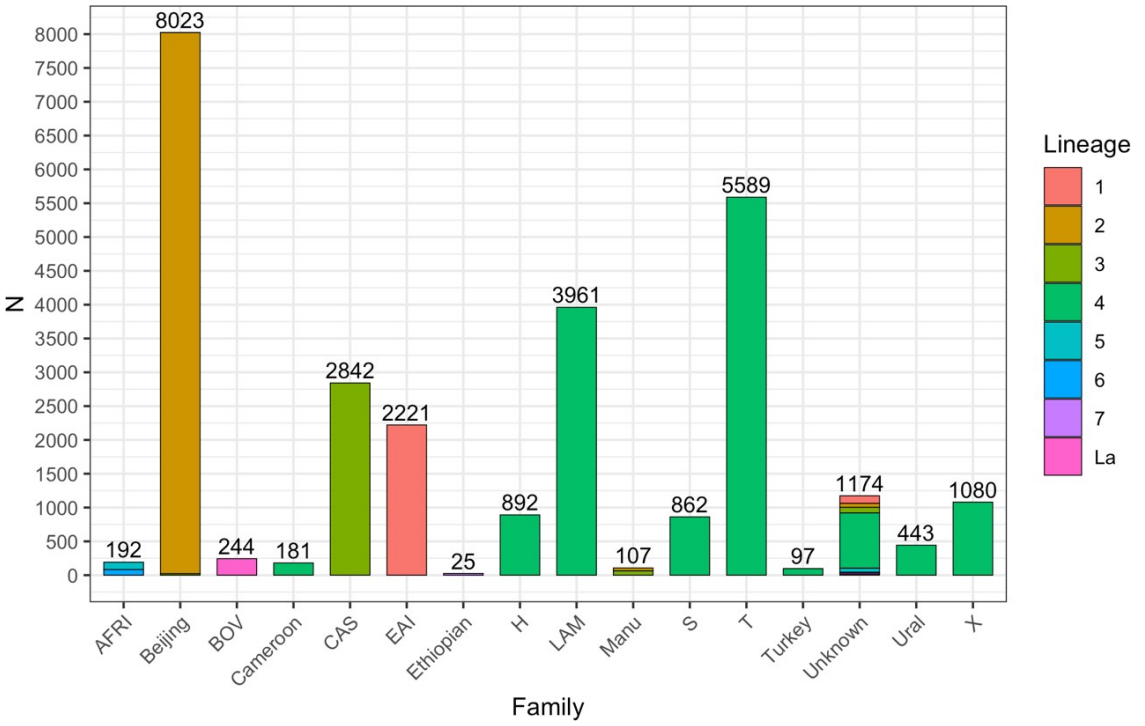
Level	Lineage	Spoligotype	SIT	Family	Proportion in Lineage	No. in lineage	% of lineage
4	1.2.1.2	s110111111111111111100111111100001011111111	19	EAI2-Manila	1.0	402	65.5
4	1.2.1.2	s110111100000000000000000011100001011111111	89	EAI2-nonthaburi	1.0	133	21.7
4	2.1	s111111111111111111111111111111111111111	523	Manu_ancestor	1.0	43	44.8
4	2.1	s111111111111111111111101111111111111111	623	Unknown	1.0	30	31.2
4	4.1.2	s111111111111111111011100111111100001001111	39	T4-CEU1	1.0	249	36.9
4	4.1.2	s111000111111111111011100111111100001001111	1258	T4-CEU1	1.0	135	20.0
4	4.2.2.1	s111111111111111111100000100111110000111111	41	Turkey	1.0	60	57.1
4	4.2.2.1	s111111111111111111100000100111110000111011	1261	Turkey	1.0	37	35.2
4	4.3.4.1	s110111111111011111110000011111110000111111	17	LAM2	1.0	61	21.9
4	4.3.4.1	s110111111111111111110000011111110000111111	20	LAM1	0.9	194	69.5
4	4.6.1.1	s110011111111111100111111111111110000111110	420	T2-Uganda	1.0	36	42.4
4	4.6.1.1	s110011111111111100100111111111110000111110	Unknown	Unknown	1.0	27	31.8
4	La1.8.1	s110100000000001011111111111111111111001100000	479	BOV_3	1.0	90	44.6
4	La1.8.1	s110100000000001011011111111111111111001100000	1158	BOV_3	1.0	49	24.3

SIT spoligotype international type

FIGURES

Figure 1

Spoligotype families and number of samples (n=27,933); by (top) Lineage; (bottom) WHO regions



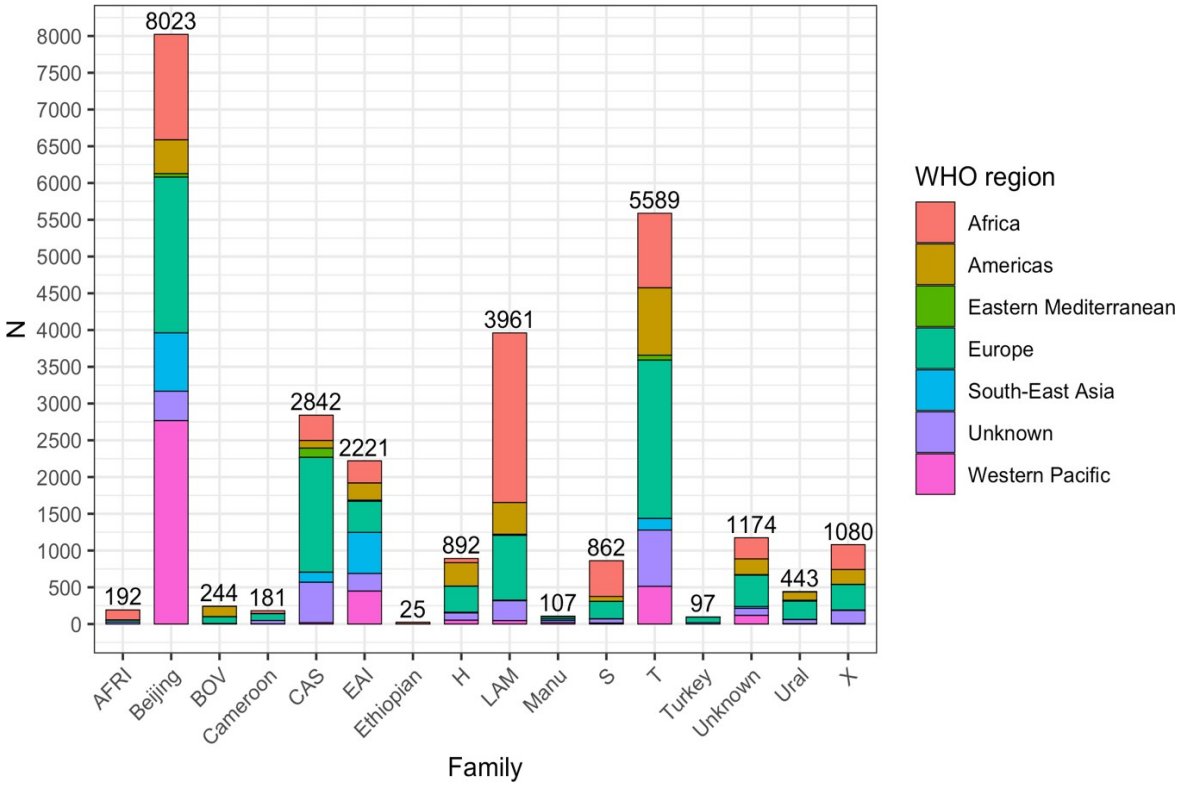
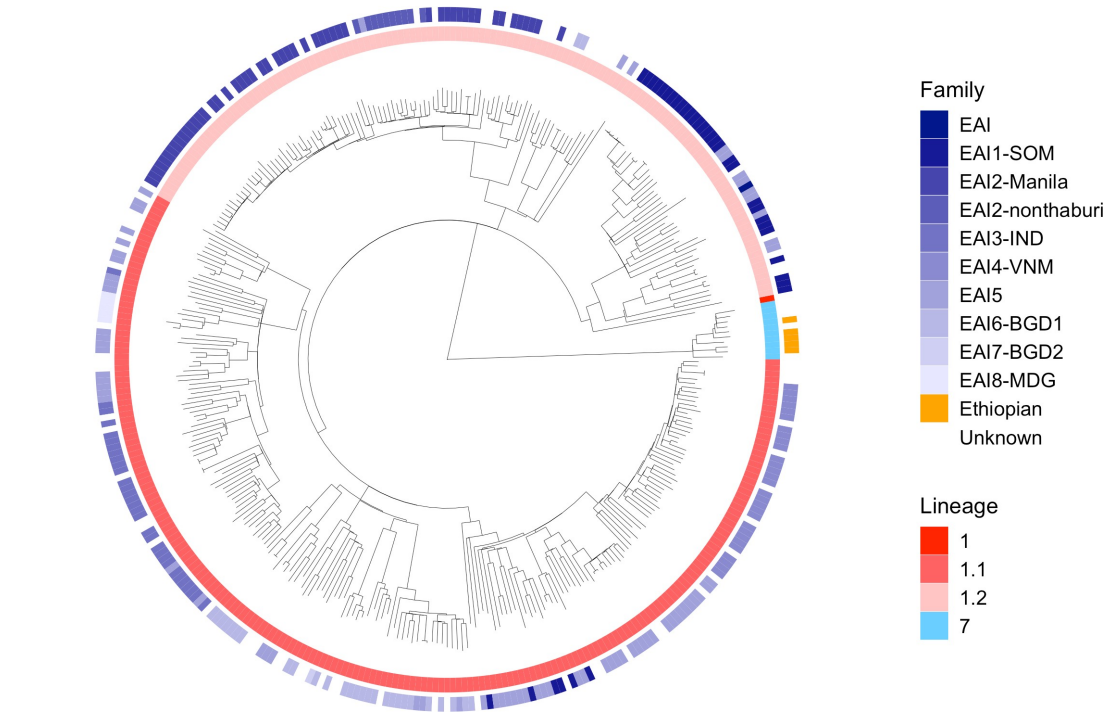


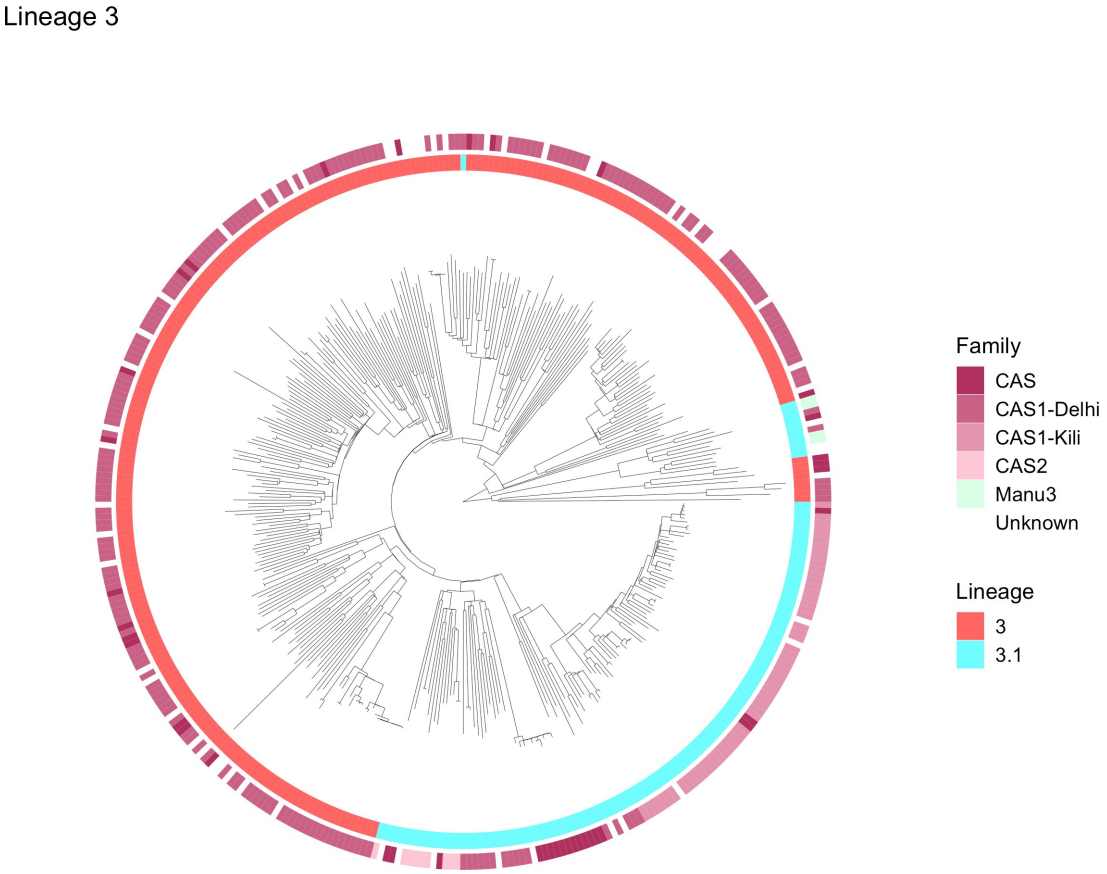
Figure 2

Spoligotype families and Lineages (n=27,933)

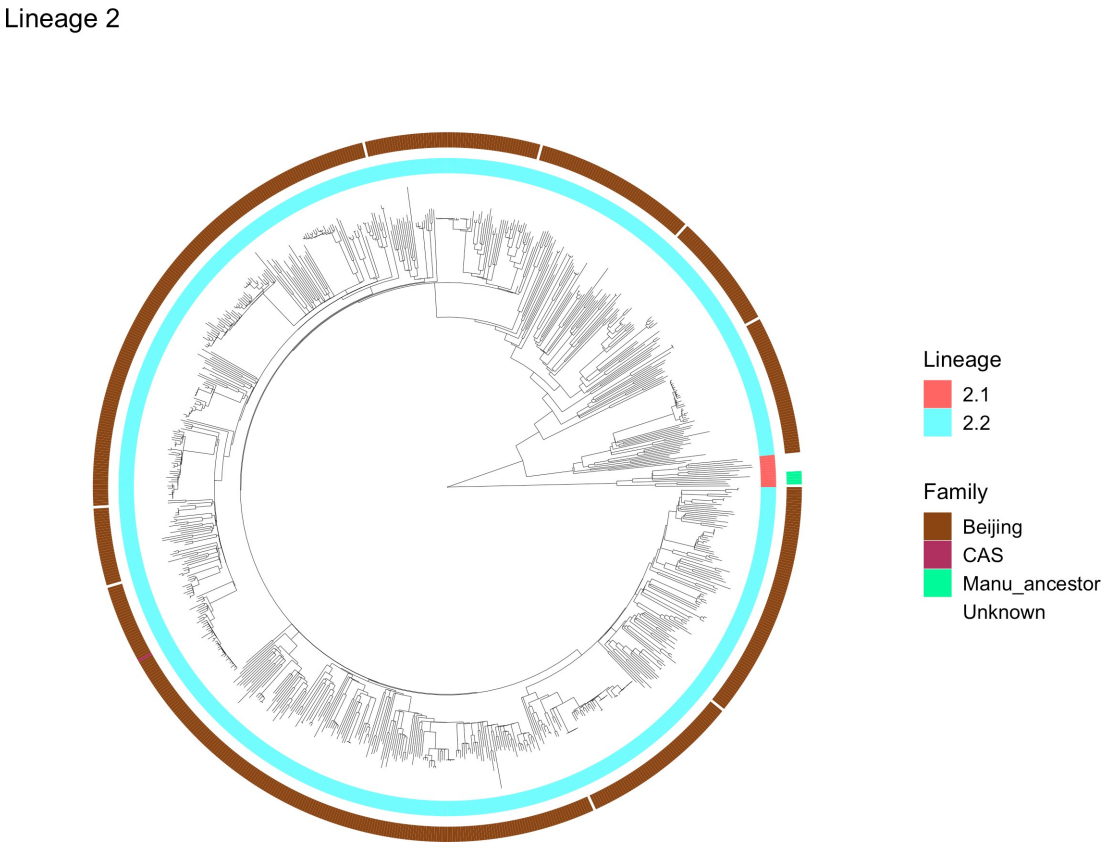
(a) Lineages 1 and 7



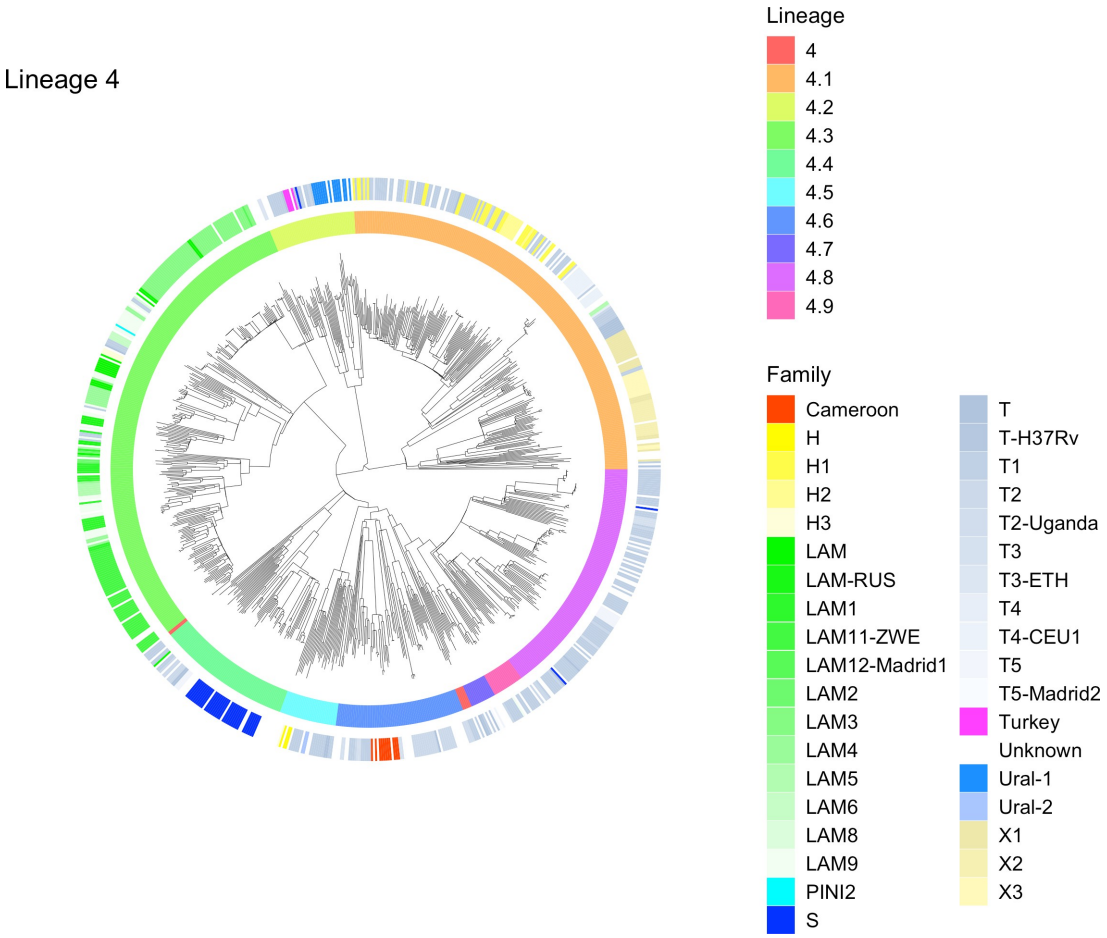
(b) Lineage 3


Lineage 3
Family
CAS
CAS1-Delhi
CAS1-Kili
CAS2
Manu3
Unknown
Lineage
3
3.1

(c) Lineage 2

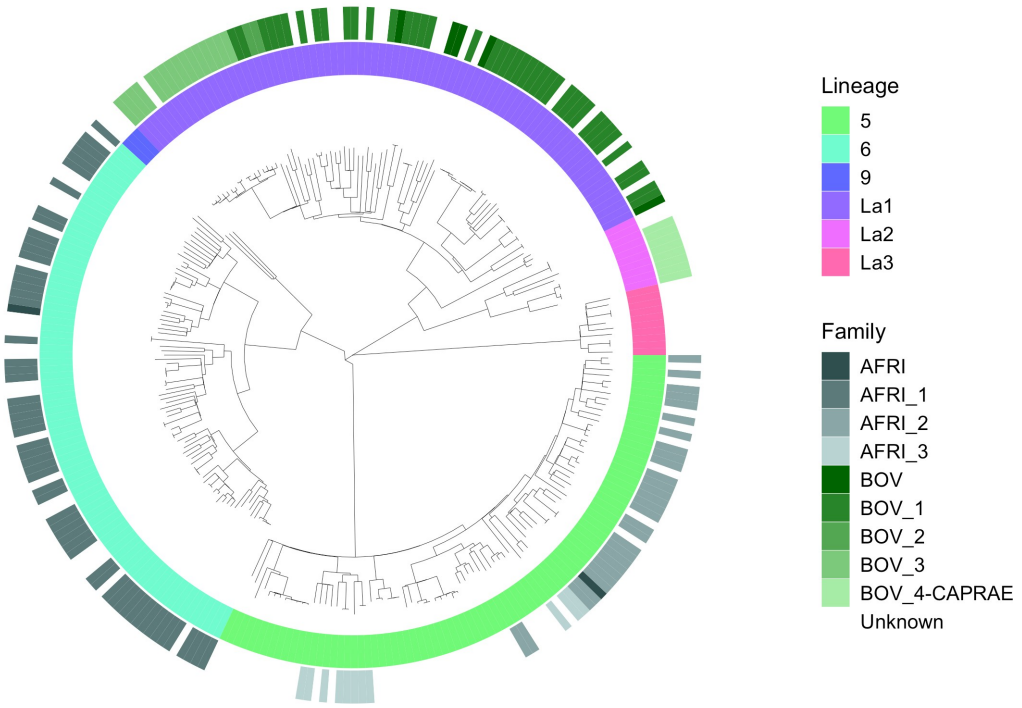

Lineage 2
Lineage
2.1
2.2
Family
Beijing
CAS
Manu_ancestor
Unknown

(d) Lineage 4


Lineage 4
Lineage
4
4.1
4.2
4.3
4.4
4.5
4.6
4.7
4.8
4.9
Family
Cameroon
H
H1
H2
H3
LAM
LAM-RUS
LAM1
LAM11-ZWE
LAM12-Madrid1
LAM2
LAM3
LAM4
LAM5
LAM6
LAM8
LAM9
PINI2
S
T
T-H37Rv
T1
T2
T2-Uganda
T3
T3-ETH
T4
T4-CEU1
T5
T5-Madrid2
Turkey
Unknown
Ural-1
Ural-2
X1
X2
X3

(e) Lineages 5, 6, 9 and La

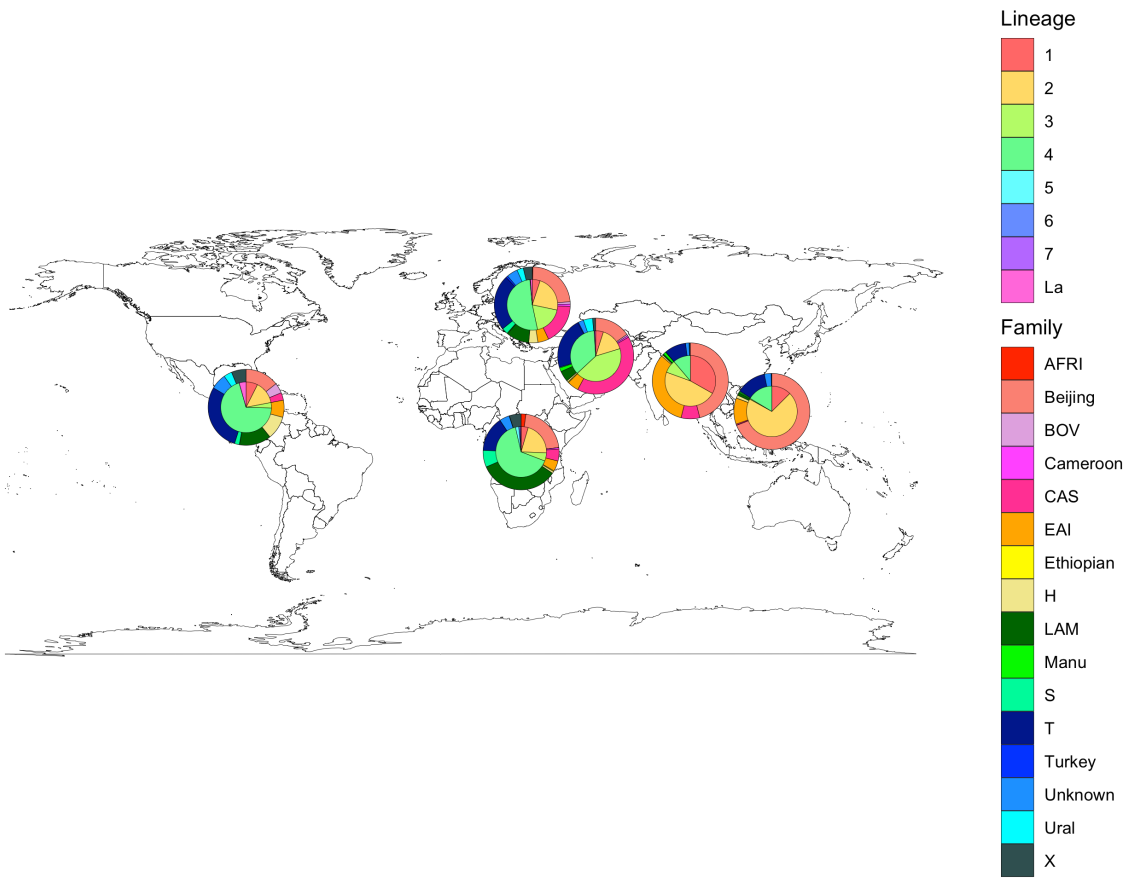
Lineages 5, 6, 9, La



Supplementary figures

Figure S1

Lineage and spoligotype family distribution across World Health Organization regions (n=27,933)



Inner pie chart: Lineage; Outer pie chart: Spoligotype family

Figure S2

Frequencies of spoligotypes at each lineage level (n=27,933)

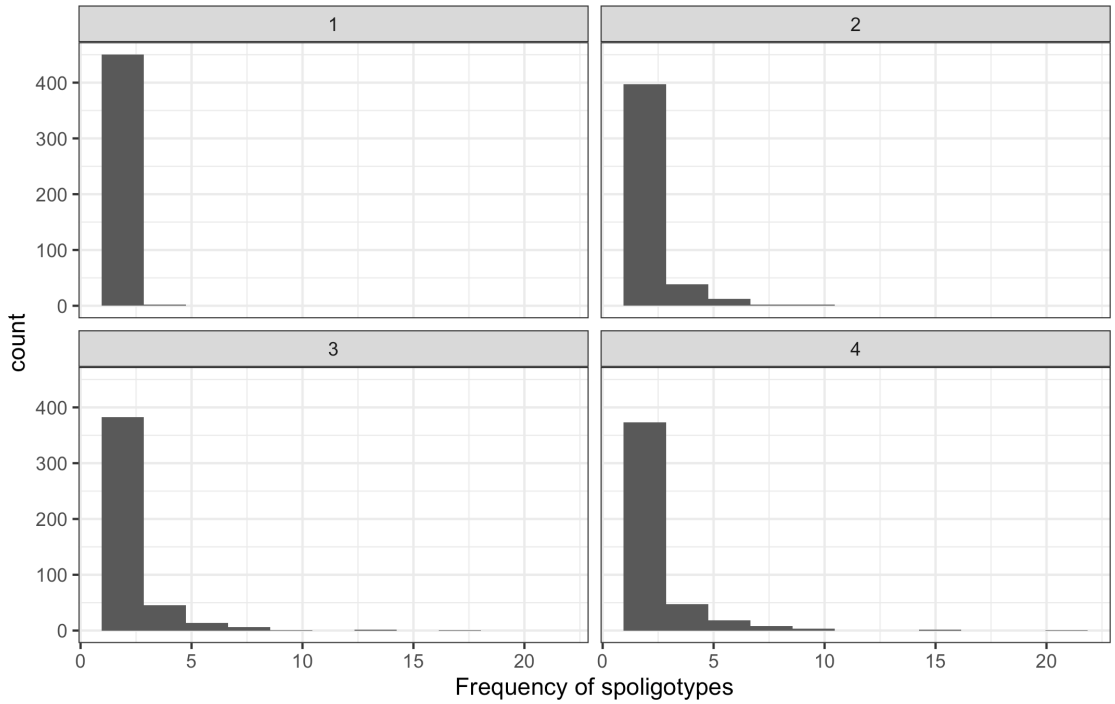
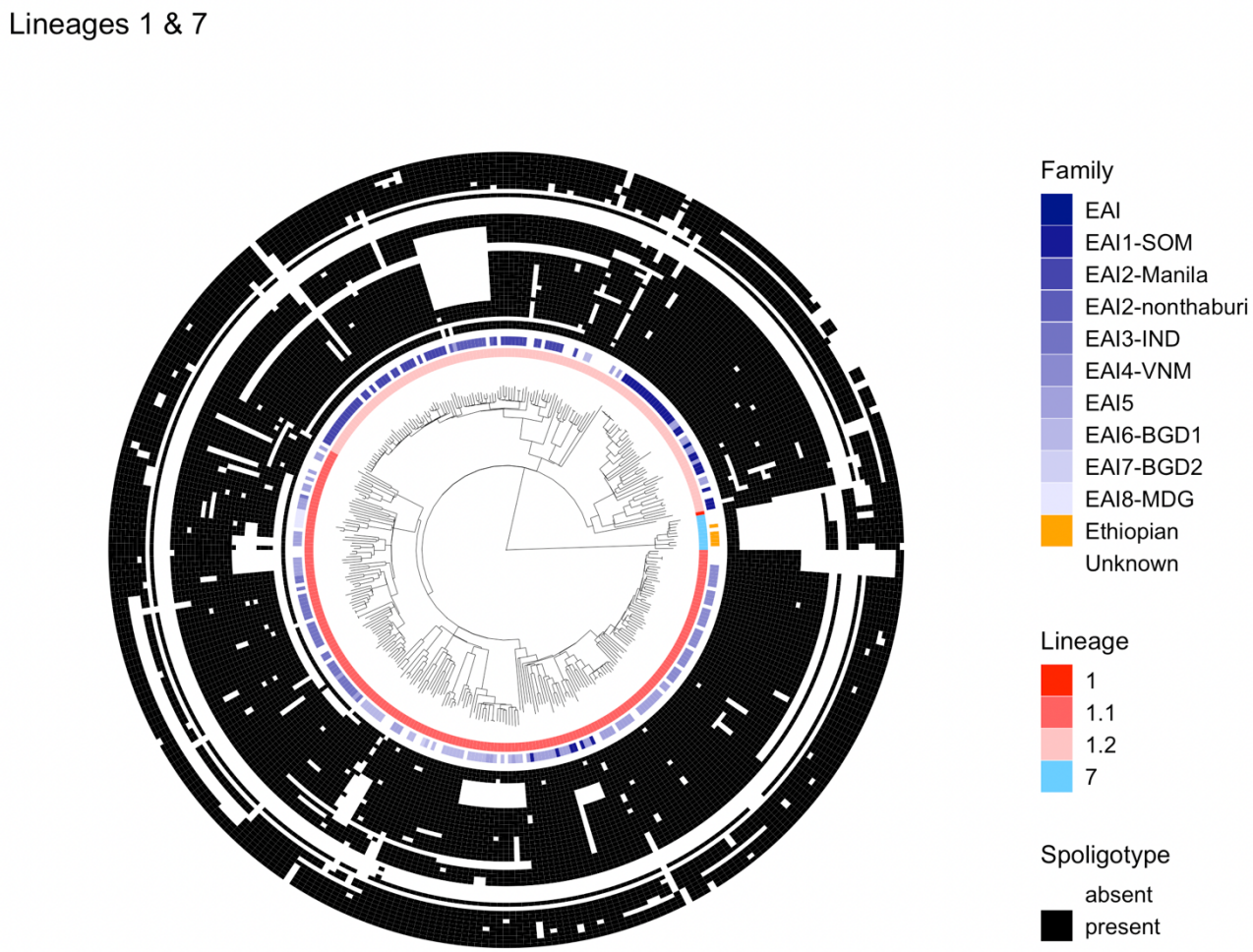
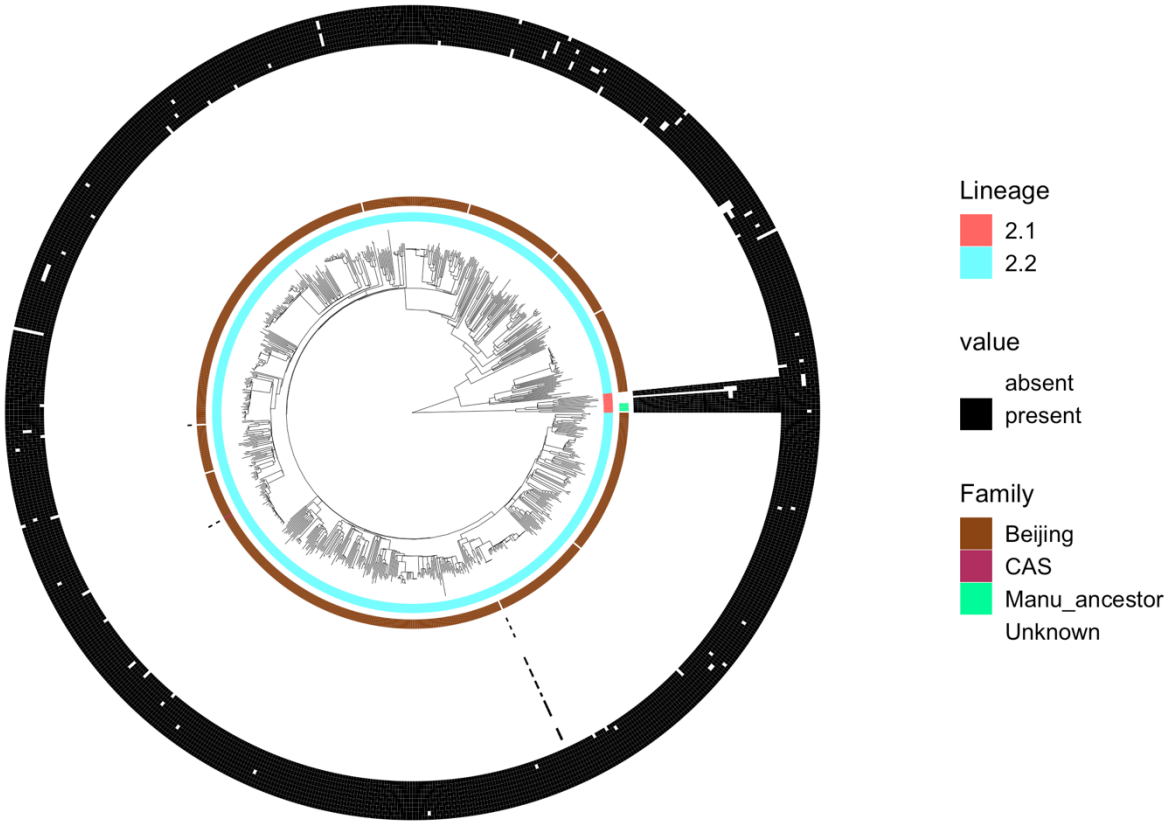


Figure S3

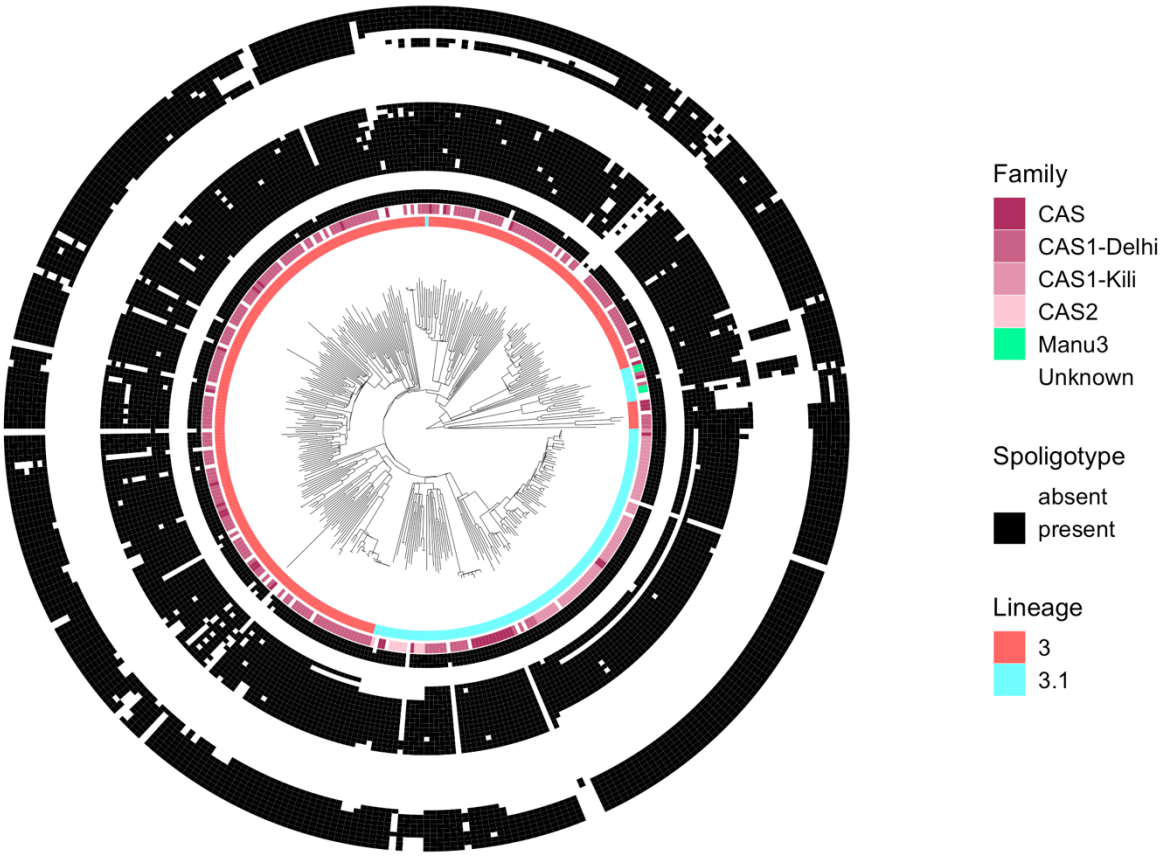
Phylogenetic trees for lineages showing the spoligotype spacer patterns and lineage.



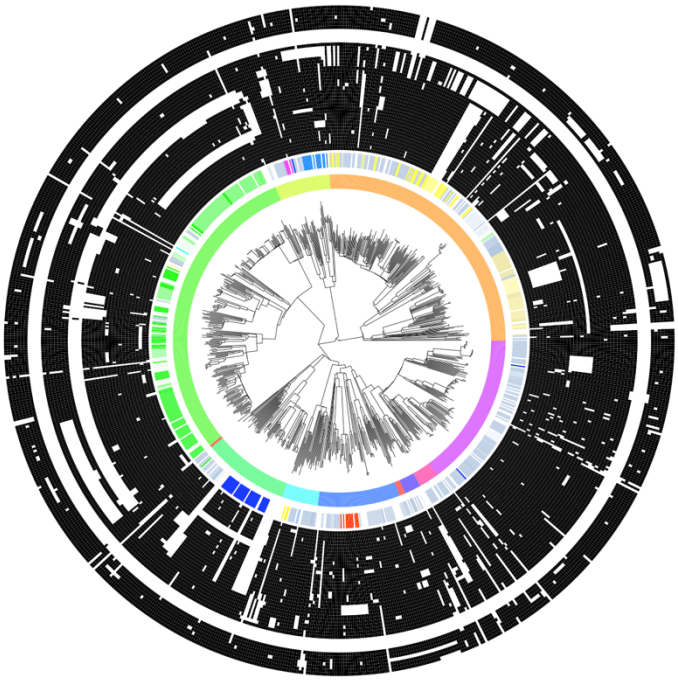
Lineage 2


Lineage
2.1
2.2
value
absent
present
Family
Beijing
CAS
Manu_ancestor
Unknown

Lineage 3


Family
CAS
CAS1-Delhi
CAS1-Kili
CAS2
Manu3
Unknown
Spoligotype
absent
present
Lineage
3
3.1

Lineage 4



Lineage
- 4
- 4.1
- 4.2
- 4.3
- 4.4
- 4.5
- 4.6
- 4.7
- 4.8
- 4.9

Family
- Cameroon
- H
- H1
- H2
- H3
- LAM
- LAM-RUS
- LAM1
- LAM11-ZWE
- LAM12-Madrid1
- LAM2
- LAM3
- LAM4
- LAM5
- LAM6
- LAM8
- LAM9
- PINI2
- S
- T
- T-H37Rv
- T1
- T2
- T2-Uganda
- T3
- T3-ETH
- T4
- T4-CEU1
- T5
- T5-Madrid2
- Turkey
- Unknown
- Ural-1
- Ural-2
- X1
- X2
- X3

Spoligotype
- absent
- present

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

London School of Hygiene & Tropical Medicine
Keppel Street, London WC1E 7HT

T: +44 (0)20 7299 4646
F: +44 (0)20 7299 4656
www.lshtm.ac.uk

RESEARCH PAPER COVER SHEET

Please note that a cover sheet must be completed for each research paper included within a thesis.

SECTION A – Student Details

Student ID Number	1807750	**Title**	Mr
First Name(s)	Gary		
Surname/Family Name	Napier		
Thesis Title	Using whole genome sequencing data to identify strain-types, transmission enhancers and novel drug resistance mutations of *Mycobacterium tuberculosis*		
Primary Supervisor	Prof. Taane Clark		

If the Research Paper has previously been published please complete Section B, if not please move to Section C.

SECTION B – Paper already published

Where was the work published?	Scientific Reports		
When was the work published?	11/05/2022		
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion			
Have you retained the copyright for the work?*	Yes	Was the work subject to academic peer review?	Yes

*If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	
Please list the paper's authors in the intended authorship order:	
Stage of publication	Choose an item.

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	I performed the bioinformatic and statistical analysis, and wrote the first draft of the manuscript. I worked with co-authors on subsequent drafts and finalisation of the paper.

SECTION E

Student Signature	
Date	30/09/2022

Supervisor Signature	
Date	22/09/2022

Chapter 4

Characterisation of drug-resistant *Mycobacterium tuberculosis* mutations and transmission in Pakistan

scientific reports

Check for updates

OPEN

Characterisation of drug-resistant *Mycobacterium tuberculosis* mutations and transmission in Pakistan

Gary Napier[1], Anwar Sheed Khan[2,3], Abdul Jabbar[4], Muhammad Tahir Khan[5], Sajid Ali[6], Muhammad Qasim[2], Noor Mohammad[2,3], Rumina Hasan[1,7], Zahra Hasan[7], Susana Campino[1], Sajjad Ahmad[8], Baharullah Khattak[2], Simon J. Waddell[9], Taj Ali Khan[8]✉, Jody E. Phelan[1]✉ & Taane G. Clark[1,10]✉

Tuberculosis, caused by *Mycobacterium tuberculosis*, is a high-burden disease in Pakistan, with multi-drug (MDR) and extensive-drug (XDR) resistance, complicating infection control. Whole genome sequencing (WGS) of *M. tuberculosis* is being used to infer lineages (strain-types), drug resistance mutations, and transmission patterns—all informing infection control and clinical decision making. Here we analyse WGS data on 535 *M. tuberculosis* isolates sourced across Pakistan between years 2003 and 2020, to understand the circulating strain-types and mutations related to 12 anti-TB drugs, as well as identify transmission clusters. Most isolates belonged to lineage 3 (n = 397; 74.2%) strain-types, and were MDR (n = 328; 61.3%) and (pre-)XDR (n = 113; 21.1%). By inferring close genomic relatedness between isolates (< 10-SNPs difference), there was evidence of *M. tuberculosis* transmission, with 55 clusters formed consisting of a total of 169 isolates. Three clusters consist of *M. tuberculosis* that are similar to isolates found outside of Pakistan. A genome-wide association analysis comparing 'transmitted' and 'non-transmitted' isolate groups, revealed the *nusG* gene as most significantly associated with a potential transmissible phenotype ($P = 5.8 \times 10^{-10}$). Overall, our study provides important insights into *M. tuberculosis* genetic diversity and transmission in Pakistan, including providing information on circulating drug resistance mutations for monitoring activities and clinical decision making.

Tuberculosis disease (TB), caused by bacteria in the *Mycobacterium tuberculosis* complex, is a major global public health problem. Pakistan is a high-burden TB country, being one of eight countries accounting for two-thirds of the estimated 10 million people globally that fell ill with the disease[1]. In 2019, Pakistan had ~ 570,000 TB cases (incidence rate 263 per 100,000) and 43,900 deaths[1], but disease control is being compromised by increasing HIV prevalence and drug resistance. The country has a high burden for rifampacin resistant (RR-TB), as well as multidrug-resistance (MDR-TB), which is the additional resistance to isoniazid treatments. Pre-extensive drug resistance (pre-XDR-TB) is prevalent[1,2], involving *M. tuberculosis* that are MDR-TB and resistant to any fluoroquinolone or at least one of the three second-line injectable drugs (capreomycin, kanamycin, amikacin). XDR-TB requires resistance to any fluoroquinolone and a second-line injectable. In January 2021, WHO updated these definitions of XDR-TB to include other drugs, such as bedaquiline[3]. Here, we adopt the older version of

[1]Faculty of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine, London, UK. [2]Department of Microbiology, Kohat University of Science and Technology, Kohat, Pakistan. [3]Laboratory Hayatabad Medical Complex, Provincial Tuberculosis Reference, Peshawar, Pakistan. [4]Department of Medical Lab Technology, University of Haripur, Haripur, Pakistan. [5]Institute of Molecular Biology and Biotechnology, The University of Lahore, KM Defence Road, Lahore 58810, Pakistan. [6]Institute of Biotechnology and Microbiology, Bacha Khan University, Charsadda, Pakistan. [7]Department of Pathology and Laboratory Medicine, The Aga Khan University, Karachi, Pakistan. [8]Institute of Institute of Pathology and Diagnostic Medicine, Khyber Medical University, Peshawar, Khyber Pakhtunkhwa, Pakistan. [9]Global Health and Infection, Brighton & Sussex Medical School, University of Sussex, Falmer BN1 9PX, UK. [10]Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London, UK. ✉email: tajalikhan.ibms@kmu.edu.pk; jody.phelan@lshtm.ac.uk; taane.clark@lshtm.ac.uk

the definition as the underlying cases were treated within that framework. There were ~ 25,000 cases of MDR-/RR-TB in 2019[1]. The National TB control program aims to reduce by half the prevalence of TB in the general population by 2025, but to achieve this will require the scaling-up of TB detection and clinical care, as well as improved systems for inferring disease transmission, thereby facilitating further targeted interventions.

Whole genome sequencing (WGS) is revolutionizing our understanding of drug resistance and clinical management, as well as transmission patterns, thereby assisting disease control[4]. *M. tuberculosis* drug resistance is linked to genomic variants in drug targets or pro-drug activators, including single nucleotide polymorphisms (SNPs) and small insertions and deletions (indels), some occurring in gene–gene interactions. It is therefore possible to predict resistance genotypically for 19 anti-TB drugs and their groups (e.g. floroquinolines) using curated libraries of > 1000 mutations across > 30 loci[5,6], thereby personalizing treatment. Genotypic predictions are an alternative to bacterial culture-based phenotypic drug susceptibility testing (DST), which can be time-consuming and resource intensive, with reproducibility and inhibitory concentration cut-off challenges for particular drugs[5]. Further, WGS data infers the population structure within the *M. tuberculosis* complex, which is phylo-geographical in nature, with strains falling within distinct (sub-)lineages[7], and potential transmission chains identified through isolates with (near-)identical genomic variation[8]. The identification of highly virulent strain-types or lineages, drug resistance, and transmission clusters will assist the targeting of limited resources for TB control.

There have been recent studies using WGS to characterize *M. tuberculosis* genetic diversity in isolates sourced from Pakistan, where the predominant strains are from the Central Asian (CAS) family, set within lineage 3[2,9–13]. A recent study of TB endemic province of Khyber Pakhtunkhwa (North West Pakistan) found that known mutations in *rpoB* (e.g. S405L), *katG* (e.g. S315T), or *inhA* promoter loci explain the majority of MDR-TB, but there was evidence of complex mixed infections and heteroresistance, which may reflect the high transmission nature of the setting[13]. An earlier study in the same province found similar MDR-TB mutations, but also additional variants in genes conferring resistance to other first and second-line drugs, including in *pncA* (pyrazinamide), *embB* (ethambutol), *gyrA* (fluoroquinolones), *rrs* (aminoglycosides), *rpsL, rrs* and *gid* (streptomycin) loci. Further, acquisition of rifampicin resistance often preceded isoniazid in these isolates, and a high proportion (~ 18%) of pre-MDR isolates had fluoroquinolone resistance markers, being a class of antibiotics that is widely available and used[2]. Eighteen *M. tuberculosis* isolates clustered within eight networks, thereby providing evidence of drug-resistant TB transmission in the Khyber Pakhtunkhwa province[2]. An investigation of XDR-TB isolates sourced across four provinces in Pakistan found similar genes linked to drug resistance as in Khyber Pakhtunkhwa[11], and an increased frequency and expression of novel SNP mutations in efflux pump genes, potentially explaining some drug resistance[11].

Here, we analyse 535 M. *tuberculosis* samples with WGS data, collected between years 2003 and 2020, with phenotypic testing of resistance across 12 drugs (rifampicin, isoniazid, ethambutol, pyrazinamide, streptomycin, ofloxacin, moxifloxacin, amikacin, kanamycin, capreomycin, ciprofloxacin, ethionamide). By identifying ~ 38 k SNPs, and inferring genotypic drug resistance across 19 anti-TB drugs (as well as fluoroquinolone and aminoglycoside classes), we sought to understand the phylogeny of *M. tuberculosis* in the largest Pakistan dataset, identify transmission events, and infer commonly circulating mutations linked to drug resistance. The genetic insights were validated in a large *M. tuberculosis* collection (n = 34 k) with WGS and drug susceptibility test data[7].

Results

Isolates and whole genome sequencing data. A total of 535 M. *tuberculosis* isolates sourced between years 2003 and 2020 from Pakistan with publically available WGS and phenotypic susceptibility testing were analysed[2,9–13]. These isolates covered at least four provinces (Balochistan, Khyber Pakhtunkhwa, Punjab, Sindh), but a high proportion of locations were missing (69.5%), all from one study[12] (Table 1). The majority of samples were from lineage 3 (L3 397, 74.2%; CAS strains), but the other main lineages were represented (L4, 80, 15.0%, including LAM, T and X strains; L2 36, 6.7%, including Beijing; L1 22, 4.1%) (Table 1; S1 Table).

As expected phenotypic drug susceptibility testing (DST) was performed most often for first-line rifampicin (n = 487, 91.0%), isoniazid (n = 487, 91.0%), ethambutol (n = 479, 89.5%), and pyrazinamide (n = 444, 83.0%) (S2 Table). A total of 432 samples (80.7%) were phenotypically resistant to at least one drug (median 3, maximum 10). The number of potential errors on the phenotypic testing appeared modest (218/2430 tests, 9.0%), where established genotypic resistance markers were present in isolates with DST results that implied drug susceptibility. The discordance appeared for nine drugs, but more than half occurred in two drugs (ethambutol 96; pyrazinamide 42) (S2 Table). The majority of isolates were genotypically assessed as MDR-TB (328, 61.3%), with proportions of (pre-) XDR (113, 21.1%) and pan-sensitive (60, 11.2%) (Table 1). There were 31 pre-MDR isolates, and overall there was a high prevalence of rifampicin (460, 86.0%) and isoniazid (435, 81.3%) resistance associated mutations. Resistance to other drugs was also detected, including ethambutol (385, 72.0%), pyrazinamide (258, 48.2%), streptomycin (238, 44.5%), ethionamide (102, 19.1%), any fluoroquinolone (277, 51.8%) or aminoglycoside (75, 14.0%). Very few isolates appeared resistant to bedaquiliine, clofazimine and cycloserine (n < 3; Table 1). Across all lineages, the majority of isolates (> 75%) were at least MDR-TB resistant (S3 Table).

After sequence data alignment, high average coverage was observed across the samples (median 76-fold, range 30—2027 fold). Across the isolates, a total of 37,970 genome-wide SNPs were identified, with 23,741 (62.5%) found in single samples. A phylogenetic tree constructed using the 37,970 genome-wide SNPs revealed the expected clustering by lineage (Fig. 1; S1 Figure).

Evidence of transmission. The median (range) pairwise SNP differences across the 535 isolates was 390 (minimum 0, maximum 1811), with a multi-modal distribution, where modes represent differences within and between lineages (S2 Figure). At a threshold of 10 SNPs, 55 clusters formed consisting of a total of 169 isolates,

Characteristic	Group	N	%
Lineage	1	22	4.1
	2	36	6.7
	3	397	74.2
	4	80	15.0
Drug resistance status[a]	Sensitive	60	11.2
	Pre-MDR	31	5.8
	MDR	328	61.3
	Pre-XDR	47	8.8
	XDR	66	12.3
	Other	3	0.6
Individual drug resistance[a]	Rifampicin	460	86.0
	Isoniazid	435	81.3
	Ethambutol	385	72.0
	Pyrazinamide	258	48.2
	Streptomycin	238	44.5
	Ofloxacin	277	51.8
	Moxifloxacin	277	51.8
	Levofloxacin	277	51.8
	Amikacin	75	14.0
	Kanamycin	79	14.8
	Capreomycin	78	14.6
	Ciprofloxacin	277	51.8
	Ethionamide	102	19.1
	Para aminosalicylic acid	10	1.9
	Cycloserine	2	0.4
	Clofazimine	1	0.2
	Bedaquiline	1	0.2
	Fluoroquinolones	277	51.8
	Aminoglycosides	75	14.0
Collection year	2003—2005	49	9.2
	2015—2017	438	81.9
	2018—2020	48	9.0
Region	Peshawar	77	14.4
	Dera Ismail Khan	25	4.7
	Abbottabad	13	2.4
	Swat	13	2.4
	Rawalpindi	7	1.3
	Hyderabad	5	0.9
	Karachi	5	0.9
	Lahore	5	0.9
	Other	13	2.4
	Missing	372	69.5

Table 1. *Mycobacterium tuberculosis* samples (N = 535). [a]Genotypic prediction using TB-Profiler.

where the median number of isolates in each cluster was 2 (range: 2—22) (S2 Figure). By reducing the cut-off to 5 SNPs, there were only 6 less clusters (total 49) consisting of a total of 33 isolates (overall 136 isolates) (S4 Table). The 169 transmitted isolates (SNP cut-off 10) were found in three of the four provinces recorded (Khyber Pakhtunkhwa 71/169; Punjab 9/169; Sindh 9/169), identified across all lineages (L1 7/169, L2 21/169, L3 98/169, L4 43/169) and in (pre-)XDR (75/169) samples (S3 Figure; S4 Figure). Most clusters had samples with the same drug resistance phenotype (44/55), and there was some evidence of clusters consisting of more than one location (35/55, excluding missing locations) (S3 Figure; S4 Figure). Comparing the 169 "transmitted" isolates in clusters to the others ("non-transmitted"; n = 366), there were overall differences in lineage (Chi-Square, $P < 6 \times 10^{-8}$) and drug resistance (Chi-square $P < 5 \times 10^{-15}$). Specifically, there was marginally weak evidence of an increased risk of transmission in lineage 2 (odds ratio (OR) = 3.00, P = 0.054) and lineage 4 (OR = 2.49, P = 0.073), compared to lineage 1. Signals of increased risk of transmission were stronger among those pre-XDR/XDR (OR = 5.79,

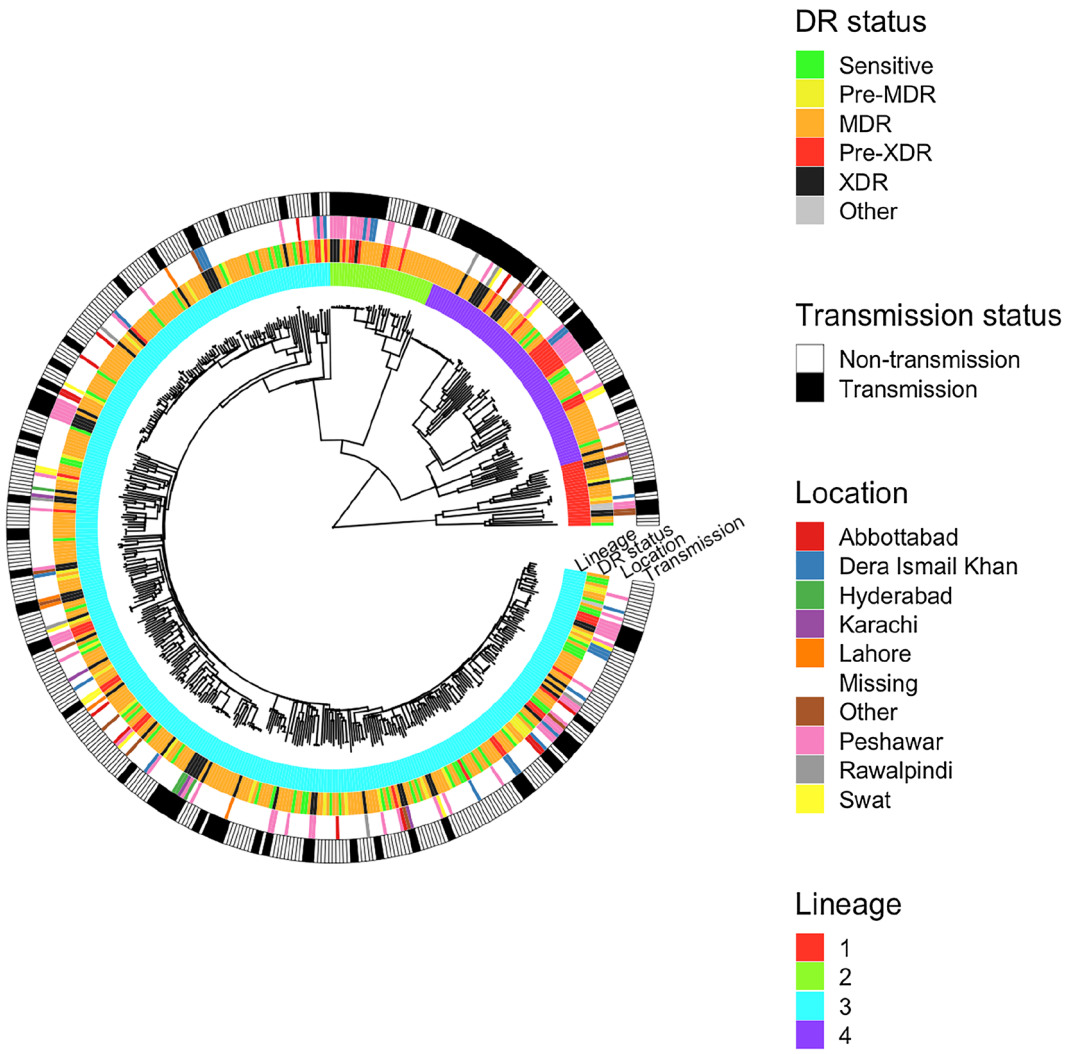


Figure 1. A phylogenetic tree for the 535 M. *tuberculosis* isolates constructed using 37,970 SNPs. The surrounding rings of data for each isolate include: lineage (inner), drug resistance status, location, and transmission status (outer).

$P < 5 \times 10^{-14}$), compared to a less resistant status. There was no association between transmission risk and province (Chi-Square P = 0.64), but there were high levels of missing location data (S5 Table).

A genome-wide association study (GWAS) approach was applied to detect loci potentially linked to transmissibility. It revealed *nusG, Rv2307B, wag31, proX* and *murA* genes to be the most associated with being in a transmission cluster ($P < 10^{-5}$) (S6 Table). *Rv2307* (beta = 0.745, $P = 1.5 \times 10^{-8}$) putatively codes for a glycine rich protein, while *proX* (beta = 0.706, $P = 1.3 \times 10^{-6}$) encodes osmoprotectant binding lipoprotein ProX. There were six mutations found in each of these genes, although no clear pattern relating to either phylogenetic or transmission status could be discerned, with mutations found in both transmission and non-transmission samples, as well as many samples having more than one of these mutations. The *nusG* (beta = 0.791, $P = 5.8 \times 10^{-10}$) encoded protein participates in transcription elongation, termination and anti-termination. There are five key mutations (S206G, E186A, R124L, A161V, F232C). By locating their position on a phylogenetic tree, only R124L was supported by isolates in more than one clade (S5 Figure). The *wag31* gene (beta = 0.912, $P = 3 \times 10^{-7}$) codes for a cell wall synthesis protein, but only one mutation (G67S) was associated with a single small transmission clade (n = 5) (S5 Figure). The *murA* gene codes for a peptidoglycan biosynthesis pathway, and had five mutations (E226K, R247L, D318A, H394Y, E414K), but none were found in more than one clade and only two mutations overlapped with transmission samples (H394Y, E226K) (S5 Figure).

The transmission clusters involved six main sub-lineages (1.1.2, 2.2.1, 3, 3.1.2, 4.5, 4.9), and we looked for similar isolates in other populations within the global 34 k dataset. Using a more relaxed cut-off of 20 SNPs difference to allow for greater time between transmission events, three of the sub-lineages (3, 2.2.1, 4.5) revealed similar isolates collected from other countries (Fig. 2). Lineage 2.2.1 had 19 Pakistan isolates linked to 29 global samples, mostly from countries in Europe and Central Asia. Lineage 3 had 8 Pakistan isolates linked to 5 other samples from the UK, while sub-lineage 4.5 had two Pakistan samples linked to a single isolate from the UK.

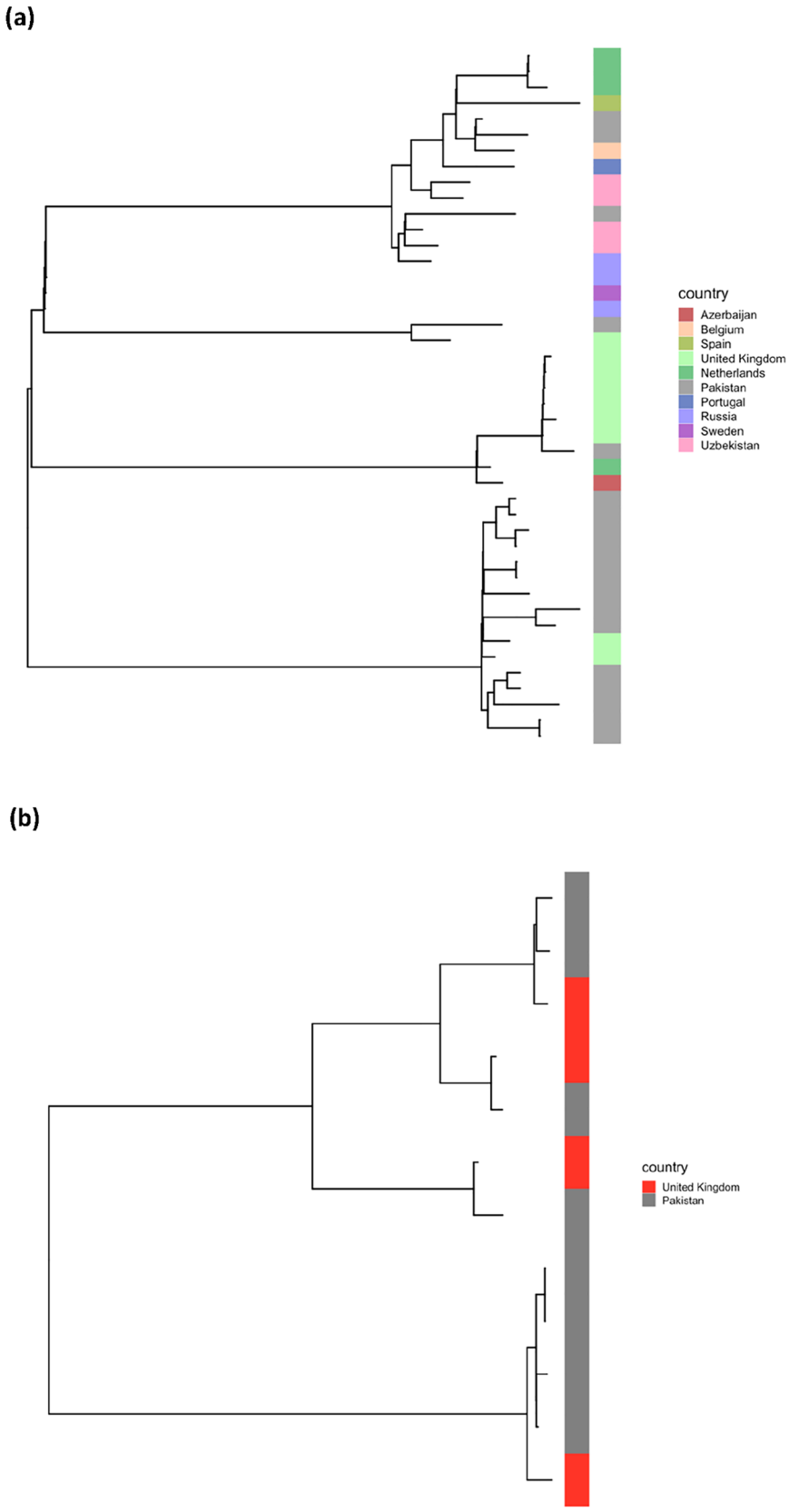


Figure 2. Phylogenetic trees for sub-lineages involving Pakistan samples and closely-related global isolates from previously published datasets. **(a)** Sub-lineage 2.2.1 (19 Pakistan, 25 other). **(b)** Lineage 3 (8 Pakistan, 4 UK).

Drug	N	Gene	Change [N]
Aminoglycosides	129	*rrs*	1401a > g [74], 514a > t [3], 906a > g, [2], 1484g > t [1], 514a > c [47], 905c > g [2], 517c > t [8]
Capreomycin	3	*tlyA*	198_198del [1], N236K [2]
Cycloserine	2	*alr*	M343T [1], L113R [1]
Ethambutol	385	*embA*	−12C > T [19], −16C > G [2], −16C > T [12], −11C > A [5]
		embB	G406A [13], G406S [8], M306I [132], G406D [22], G406C [6], Q497R [20], Q497K [9], Q497P [2], Q853P [2], E405D [1], E504D [2], A313V [1], M306L [20], M306V [127], Y319C [1], Y319S [1], Y334H [2], S347I [1], D354A [7], D1024N [29], D328Y [3]
Ethionamide	54	*ethA*	1200_1201del [1], 1054_1054del [1], 599_599del [1], 1261_1262insCGAGC [1], 1018_1018del [1], 1047_1047del [1], 1300_1301insGT [1], 61_61del [1], 671_671del [1], 1290_1291insC [1], 4326936_4328449del [5], 4326943_4328449del [1], 4326944_4328449del [1], 4327038_4327099del [1], Q269* [4], Q347* [6], L272P [1], L397R [2], T61M [1], 672_673insG [2], 673_674insGC [1], 140_140del [3], 150_150del [1], 299_299del [3], 313_319del [1], 352_365del [2], 382_383insG [4], 392_392del [2], 404_405insAT [1], 703_703del [1], 755_756insGC [2], 825_825del [1]
Fluoroquinolones	277	*gyrA*	G88A [1], G88C [3], D89N [4], A90V [68], S91P [21], D94G [128], D94A [19], D94H [4], D94Y [17], D94N [24]
		gyrB	R446C [1], S447F [5], I486L [1], T500N [3], E501D [4]
Isoniazid	416	*katG*	22_23insA [1], 238_260del [1], 337_337del [1], 679_680insGC [1], 87_87del [1], 974_974del [1], 2148451_2164815del [1], 2149885_2172950del [1], 2151318_2157225del [1], 2152294_2157889del [1], A172V [1], R104Q [1], D259E [1], G297V [1], S140N [2], S315N [8], S315I [1], S315T [365], T275A [2], T380I [3], W191R [2], W328S [1], Y155C [1], Y155S [2], Y337C [1], Y413H [1], V1A [2], 1176_1177insG [1], 1196_1197insGA [1], 1284_1284del [1], 1328_1328del [1], 1486_1487insC [1], 2005_2006insG [2], 58_58del [1], 58_59insCT [1], 596_596del [1], 371_371del [1], 60_61insGT [1]
		ahpC	−54C > T [4], −81C > T [2]
Kanamycin	5	*eis*	−10G > A [1], −14C > T [3], −37G > T [1]
Pyrazinamide	258	*pncA*	−11A > C [4], −11A > G [15], −12 T > C [1], 108_108del [1], 13_14insGA [1], 166_167insG [1], 194_203del [1], 206_207insC [1], 209_210insACC [1], 226_236del [1], 230_231insA [1], 283_283del [1], 314_315insG [2], 346_347insC [2], 377_378insGA [1], 382_383insG [1], 391_392insG [2], 391_392insGG [17], 393_394insC [1], 408_409insT [1], 412_413insCATT [1], 417_418insG [3], 424_425insGA [2], 429_429del [1], 430_431insG [1], 438_439insCG [1], 455_456insATGGCTTGGC [2], 501_502insC [1], 53_53del [1], 61_62insG [1], 7_7del [1], 2285437_2291074del [1], 2288627_2289103del [2], 2288776_2288836del [1], 2288825_2289242del [1], 2289006_2290299del [1], A134V [1], A143D [1], A146V [2], A171T [4], A3E [2], R140G [4], D12A [3], D136Y [1], D49N [1], D49G [1], D63G [1], D63H [4], C14R [1], C72Y [1], Q10* [1], Q10R [5], Q10P [5], Q141P [4], G105D [1], G108R [1], G132S [3], G78S [4], G78V [1], G97S [3], H51Q [1], H57P [1], H57Y [4], H71R [3]. H71Y [3], H82R [1], I133T [2], I31S [1], I5T [2], I6M [1], I6T [1], L156P [1], L159R [2], L19R [2], L27P [1], L35P [1], L4S [2], L4W [1], L85P [1], K96R [2], K96E [1], K96T [2], M175T [2], M1T [1], F58L [1], F94L [1], P54L [23], P62L [2], P62S [2], P69R [3], S104R [1], S164P [1],S67P [4], T100I [1], T135P [4], T142A [1],T142M [1], T160P [3], T47P [3], T61P [1], T76I [2], T76P [6], W119R [3], W68* [1], W68R [2], W68C [3], W68G [2], W68S [1], Y103C [1], Y34S [1], Y41* [1], V128G [1], V139A [4], V139G [1], V180F [8], V45G [1], V7F [2], V9G [2]
Rifampicin	460	*rpoB*	1296_1297insTTC [2], 1306_1308del [3], A286V [2], N437Y [1], D435G [5], D435F [2], D435Y [30], D435V [32], Q429H [2], Q432H [1], Q432L [3], Q432K [4], Q432P [1], H445R [2], H445N [8], H445D [6], H445C [2], H445Q [2], H445L [9], H445P [2], H445Y [11], I480V [1], I491F [1], L430R [3], L430P [8], L452P [12], M434I [11], S428R [1], S428T [1], S441Q [1], S441L [2], S450L [293], S450F [2], S450W [9], S450Y [1], S493L [1], T400A [1], T444I [1], V170F [2]
		rpoC	D747A [1], G332R [6], I491T [13], I885V [1], L527V [5]
Streptomycin	172	*gid*	102_102del [2], 115_115del [3], 351_351del [4], 4407713_4407860del [1], A80P [3], L79S [1]
		rpsL	K43R [126], K88R [21], K88M [1], K88T [2]
INH, Ethionamide	58	*fabG1*	−15C > T [52], −17G > T [1], −8 T > A [1], −8 T > C [4]
		inhA	I194T [4], I21T[1], S94A [4], I21V [1]
PAS	10	*folC*	E153A [1], I43S [5], R49W [1], I43T [1]
		thyX	−16C > T [2]
BDQ, CFZ	1	*mmpR5*	192_193insG [1]

Table 2. Number of samples with known drug resistance-associated mutations. *BDQ* bedaquiline, *CFZ* clofazimine, *INH* isoniazid, *PAS* para aminosalicylic acid. *Premature stop codon.

Drug resistance mutations. The common mutations underlying genotypic drug resistance were in known loci. These included mutations in *rpoB* (D435GFYV 293/460, S450LFWY 308/460) linked to rifampicin, *katG* (S315NIT 374/416) and *fabG1* (−15C > T 52/416) linked to isoniazid, *embB* (G406ASDC 51/385, M306ILV 280/385, Q497RKP 40/385) linked to ethambutol, *gyrA* (A90V 68/277, S91P 22/277, D94GAHYN 195/277) linked to fluoroquinolones, and *pncA* (118 low frequency < 25/258) linked to pyrazinamide (Table 2). A high proportion of mutations detected were present in the global 34 k dataset, including *pncA* 93/118, *katG* 19/38, *rpoB* 37/39, and *embB* 21/21. Nearly half all mutations identified (156/313) were present in single isolates, of which the majority were in the 34 k dataset (101/156) and absent from sensitive strains (S7 Table).

We investigated isolates that had a DST implying resistance, but no established genetic mutations to explain this phenotype. There were 82 isolates (100/2430 tests; S2 Table)) with this discordance across 9 drugs (amikacin (9), capreomycin (2), ciprofloxacin (4), ethambutol (17), isoniazid (25), kanamycin (7)), pyrazinamide (24), rifampicin (6), streptomycin (6)). We identified 68 distinct genetic markers in candidate genes to potentially explain the discordance (Table 3). Twenty-nine (42.6%) mutations had strong evidence of being linked with drug resistance, including from functional consequences, homoplasy or global data information[7,14]. Forty-six (67.6%) mutations were present in the global 34 k dataset, and all of these were absent in sensitive strains (S8 Table), reinforcing them as putatively resistant related.

For rifampicin resistance, we identified three inframe indels in *rpoB* (1291_1292insGCC, 1294_1296del and 1309_1311del) in three isolates. For isoniazid, several nonsense mutations in *katG* were found, with 3 mutations leading to premature stop codons (W438*, W204*, Q36*) and a frameshift mutation (587_588insGGT). For ethambutol resistance, variants in the *embA* promoter region (−42CAT > C, −27TA > T-16C > A, −8C > A) and *embB*

Drug	Gene	Change [N]
Amikacin	*rrs*	−92 T > G [1], 878 g > a [2]
Ciprofloxacin	*gyrA*	A288D [1]
	gyrB	−162C > CG [1], A432V [1]
Ethambutol	*embA*	−16C > A [2], −27TA > T [1], −42CAT > C [1], −8C > A [1], **P455Q** [1], **V534A** [1]
	embB	R524H [1], D328H [1], **D328F** [2], L172R [2], **F330L** [1], T546I [1]
	ubiA	G268D [1], **F238I** [1], **V188L** [1]
Isoniazid	*ahpC*	−52C > A [1], −72C > T [1], −76 T > A [4], −76 T > C [1], **−76 T > G** [1], −93G > A [1]
	kasA	M72I [1], **F402I** [1]
	katG	**587_588insGGT** [1], A122D [1], A348G [1], **R484G** [1], D189Y [1], **Q36*** [1], **G186D** [1], **G299D** [1], I103V [1], **L298S** [1], M105I [1], F408S [1], P100T [1], T271I [1], T475I [1], T625K [1], W204* [1], W438* [1], **Y197D** [1]
Kanamycin	*eis*	L386I [1]
	rrs	−92 T > G [1]
Pyrazinamide	*pncA*	−7 T > G [1], **392_393insGGT** [1], **451_462del** [1], **511_512insTCGCCG** [1], L120R [4], P62T [1], P69T [1], **S18*** [1], V130M [1], V180A [1]
	rpsA	**−98A > T** [1], **Q410R** [2]
Rifampicin	*rpoB*	**1291_1292insGCC** [1], 1294_1296del [1], 1309_1311del [1]
Streptomycin	*gid*	A119D [1], A82P [1], D67G [1], G71* [2]

Table 3. Putative novel drug resistant mutations. *Based on absence in the curated TB-Profiler mutation list; bolded, if not observed in a large TB Global dataset (34 k[7]); underlined, if with multiple levels of evidence for drug resistance (see S8 Table).

were observed. For pyrazinamide resistance, several potentially new mutations were found in *pncA*, including three inframe indels (511_512insTCGCCG, 392_393insGGT and 451_462del), a premature stop codon (S18*), and SNPs in both the coding region (Val180Ala) and the promoter (−7 T > G). For streptomycin resistance, several mutations were found in *gid* including a premature stop codon (G71*), a frameshift (102_102del), and SNPs (A119D, A82P and D67G). These SNPs were found in the 34 k global dataset, and likely acquired as the result of homoplasy. The *gid* A119D mutation was present in 15 isolates (ten different sublineages), of which two had DSTs that reported resistance. The *gid* A82P mutation was present in three isolates from two different sublineages, but no DST data was available for these samples. The *gid* D67G was present in 38 global isolates from five different sublineages. Of these, seven isolates had DST data available with four presenting with resistance.

For second line injectables, the *rrs* 878g > a mutation (seen previously[2]) was observed in four lineage 3 strains with three independent homoplastic acquisitions, indicating it is unlikely to be strain-specifc. Mutations in *rrs* are generally clustered in two regions with the most common mutations involved with streptomycin resistance being located around position 514 and those involved with resistance to amikacin, kanamycin and capreomycin located around 1401. The *rrs* 878g > a falls between the two mutation hotspots, and of the three strains which had DST data (amikacin and kanamycin) in this study, two were resistant to both amikacin and kanamycin and the other was sensitive to both. For fluoroquinolones, the *gyrA* A288D mutation was found in three lineage 3 isolates and was acquired in each sample independently. One isolate tested resistant to ciprofloxacin with no known resistance mutation found in the *gyrA* and *gyrB* genes.

Discussion

The use of whole genome sequencing as a diagnostic is gaining traction in low resource and high TB burden settings, where it has the potential to have greater public health impact[5,7,15]. Portable sequencing platforms and multiplexing of *M. tuberculosis* isolates are making the application of WGS, both timely and cost effective[5]. Our findings in the largest analysis of isolates from Pakistan to date revealed that lineage 2 and 4 strains, which are pre-XDR and XDR-TB, are potentially being transmitted in the country. Evidence of increased transmission among lineages 2 and 4 is consistent with previous characterisations of these clades as more transmissible[7], and therefore their strain-types should be monitored more closely despite greater prevalence of lineage 3. It is surprising that pre-XDR and XDR-TB samples were found to be clustered more than expected compared to MDR-TB isolates given the usual fitness cost of drug resistance. This observation suggests that compensatory mutations ought to be investigated in future work. Similarly, the finding that mutations in *nusG, Rv2307B, wag31, proX* and *murA* genes maybe associated with transmission should be followed-up experimentally, where those with variants appearing in more than one clade could be priortised. Advances in the characterisation of transmission events[16], GWAS[9,17] and machine learning methods[18,19] could enhance the ability to detect mutations linked to transmissiblility. However, host factors and host–pathogen genetic interactions are also likely to be important. More broadly, the routine collection, processing and WGS of *M. tuberculosis* DNA across Pakistan will provide robust insights into mutations underlying drug resistance and geo-temporal dynamics.

Whilst our study uses a convenience sample that is not necessarily representative of the proportions of MDR-TB in the wider Pakistan population, it is enriched by the presence of many mutations that lead to drug resistance. The enrichment of drug resistant isolates from endemic TB regions with high transmission will reveal important resistance mutations, including potential novel variants. To investigate the underlying mechanisms

of drug resistance, we compared susceptibility profiles from phenotypic methods and genotypic prediction. This analysis led to the identification of a number of potential new drug resistance mutations, including in genes causing resistance to rifampicin, isoniazid, ethambutol and pyrazinamide. Three inframe deletions were found in the rifampicin resistance determining region of *rpoB*. Inframe deletions have not been widely reported as a major mechanism of resistance to rifampicin and it is surprising to see a relatively high number of these mutations in our dataset. Previously unreported nonsense mutations were also found in the *katG* gene, a locus responsible for resistance to isoniazid. A novel nonsense mutation, frameshift and inframe indels were found in the *pncA* gene, which codes for the activator of pyrazinamide. Mutations in the promoter region of the *pncA* gene lead to changes in the expression of PncA and resistance[20]. The identified −7 T > G promoter mutation is thus likely to cause resistance. However the functional effects of SNPs found in the coding region of *pncA* are more difficult to predict[20]. The *pncA* V180A mutation has been reported previously to be associated with pyrazinamide resistance[20]. For streptomycin, we observed several point mutations and a premature stop codon in the *gid* gene. The *gid* D67G mutation was found in 38 isolates in the 34 k global dataset[7], of which 57% of those were phenotypically resistant to streptomycin. The incomplete penetrance of the streptomycin-associated *gid* D67G mutation could be explained by the relative low-level resistance conferred by mutations in *gid,* which could be below established critical cut-offs of minimum inhibitory concentration for susceptibility phenotyping, but still elevated with respect to wild-type.

Overall, our work reinforces that the adoption of WGS platforms as a diagnostic tool, combined with mutational databases of drug resistance markers, will inform clinical decision making. The ability to perform WGS for genomic investigations across time and geography will improve the understanding of transmission dynamics, and inform control programmes to reduce disease burden. The benefits will be greatest in high prevalence TB settings, typically low and middle income countries, such as Pakistan. Although WGS is not currently at a viable level of affordability, it is anticipated that amplicon and whole genome approaches using (portable) next generation platforms will shortly become simple, affordable and accessible rapid diagnostics compared to traditional laboratory-based methods that currently require specialist training, equipment and long culture times. Importantly, there is evidence that WGS is more detailed and accurate in its profiling of drug resistance than traditional DST, thereby likely to improve treatment and mortality outcomes in drug-resistant TB in high-burden countries[21].

Methods

Sequence data and processing. WGS were sourced across six studies[2,9–13] (ENA accessions: PRJEB7798, PRJEB10385, PRJEB25972, PRJEB32684, PRJEB43284), where contributing isolates belong to a single patient. Phenotypic DSTs were conducted using WHO endorsed methods, as specified in descriptions of the original studies[2,9–13]. Raw reads were trimmed to remove low-quality sequences in Trimmomatic (v0.39)[22], and aligned to the H37Rv reference genome (AL123456) with BWA mem (v0.7.17)[23]. SNPs and indels called by samtools software[24] were processed using gatk GenotypeGVCFs (v4.1.3.0) (gatk.broadinstitute.org). Monomorphic SNPs and variants in non-unique regions of the genome (e.g. *pe/ppe* genes) were excluded. A multi-FASTA format file was created from the filtered SNP file and H37Rv reference fasta using bedtools makewindows (v2.28.0)[25]. This multiple alignment was used to construct a phylogenetic tree with IQ-TREE (v1.6.12), involving a general time reversible model with rate heterogeneity set to a discrete Gamma model and an ascertainment bias correction (parameters −m GTR + G + ASC), with 1000 bootstrap samples[26]. Pairwise distance matrices were calculated in Plink software (v1.90b4)[27]. Drug resistance and lineages were predicted in silico from raw sequence data using TB-Profiler (v2.4)[5]. The Pakistan analysis results were compared to a global collection of 34 k *M. tuberculosis* with WGS and DST data[7].

A cut-off of 10 SNPs difference was established to define transmission clades, and label samples as "transmitted" or "non-transmitted". A sensitivity analysis was performed to assess the impact of changing the cut-off. Linear mixed models were used perform a GWAS of transmissibility using SNPs, location, drug resistance and adjusting for *M. tuberculosis* (sub-)lineage and outbreak-based population structure, being implemented in GEMMA (v.1.1.2) (http://www.xzlab.org/software.html). We report association p-values less than a Bonferroni cut-off based on testing 4,000 genes ($P < 1.25 \times 10^{-5}$). To identify if samples involved in transmission clades (> 10 samples) were similar to others (< 20 SNPs) in the global dataset (n = 34 k)[7], we constructed phylogenetic trees using FastTree for the relevant sub-lineages (1.1.2, 2.2.1, 3, 3.1.2, 4.5, 4.9). The likelihoods of ancestral locations were inferred with the ape (v5.0) and phytools packages in R.

Received: 3 November 2021; Accepted: 5 April 2022
Published online: 11 May 2022

References

1. World Health Organization (WHO). *Global Tuberculosis Report 2021*. (2021).
2. Jabbar, A. *et al.* Whole genome sequencing of drug resistant Mycobacterium tuberculosis isolates from a high burden tuberculosis region of North West Pakistan. *Sci. Rep.* **9**, 1–9 (2019).
3. World Health Organisation. *Meeting Report of the WHO Expert Consultation on Drug-Resistant Tuberculosis Treatment Outcome Definitions, 17–19 November 2020*. (2020).
4. Phelan, J. *et al.* The variability and reproducibility of whole genome sequencing technology for detecting resistance to anti-tuberculous drugs. *Genome Med.* **8**, 172 (2016).
5. Phelan, J. E. *et al.* Integrating informatics tools and portable sequencing technology for rapid detection of resistance to anti-tuberculous drugs. *Genome Med.* **11**, 1–7 (2019).
6. Coll, F. *et al.* Rapid determination of anti-tuberculosis drug resistance from whole-genome sequences. *Genome Med.* **7**, 51 (2015).

7. Napier, G. *et al.* Robust barcoding and identification of *Mycobacterium tuberculosis* lineages for epidemiological and clinical studies. *Genome Med.* **12**, 114 (2020).
8. Glynn, J. R. *et al.* Whole genome sequencing shows a low proportion of tuberculosis disease is attributable to known close contacts in rural Malawi. *PLoS ONE* **10**, 132840 (2015).
9. Coll, F. *et al.* Genome-wide analysis of multi- and extensively drug-resistant *Mycobacterium tuberculosis*. *Nat. Genet.* **50**, 307–316 (2018).
10. Kanji, A. *et al.* Alternate efflux pump mechanism may contribute to drug resistance in extensively drug-resistant isolates of *Mycobacterium tuberculosis*. *Int. J. Mycobacteriol.* **5**, 97 (2016).
11. Ali, A. *et al.* Whole genome sequencing based characterization of extensively drug-resistant mycobacterium tuberculosis isolates from pakistan. *PLoS ONE* **10**, 117771 (2015).
12. Cryptic-Consortium. Prediction of susceptibility to first-line tuberculosis drugs by DNA sequencing. *N. Engl. J. Med.* **379**, 1403–1415 (2018).
13. Khan, A. S. *et al.* Characterization of rifampicin-resistant *Mycobacterium tuberculosis* in Khyber Pakhtunkhwa, Pakistan. *Sci. Rep.* **11**, 1–10 (2021).
14. Phelan, J. E. *et al.* Integrating informatics tools and portable sequencing technology for rapid detection of resistance to anti-tuberculous drugs. *Genome Med.* **11**, 41 (2019).
15. Phelan, J. *et al. Mycobacterium tuberculosis* whole genome sequencing and protein structure modelling provides insights into anti-tuberculosis drug resistance. *BMC Med.* **14**, 307 (2016).
16. Sobkowiak, B. *et al.* Bayesian reconstruction of mycobacterium tuberculosis transmission networks in a high incidence area over two decades in Malawi reveals associated risk factors and genomic variants. *Microb. Genomics* **6**, 4 (2020).
17. Oppong, Y. E. A. *et al.* Genome-wide analysis of *Mycobacterium tuberculosis* polymorphisms reveals lineage-specific associations with drug resistance. *BMC Genomics* **20**, 1–15 (2019).
18. Libiseller-Egger, J., Phelan, J., Campino, S., Mohareb, F. & Clark, T. G. Robust detection of point mutations involved in multidrug-resistant Mycobacterium tuberculosis in the presence of co-occurrent resistance markers. *PLoS Comput. Biol.* **16**, 1008518 (2020).
19. Deelder, W. *et al.* Machine learning predicts accurately *Mycobacterium tuberculosis* drug resistance from whole genome sequencing data. *Front. Genet.* **10**, 922 (2019).
20. Tunstall, T., Phelan, J., Eccleston, C., Clark, T. G. & Furnham, N. Structural and genomic insights into pyrazinamide resistance in *Mycobacterium tuberculosis* underlie differences between ancient and modern lineages. *Front. Mol. Biosci.* **8** (2021).
21. Zürcher, K. *et al.* Mortality from drug-resistant tuberculosis in high-burden countries comparing routine drug susceptibility testing with whole-genome sequencing: A multicentre cohort study. *Lancet Microbe* **2**, e320–e330 (2021).
22. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
23. Li, H. *Aligning Sequence Reads, Clone Sequences and Assembly Contigs with BWA-MEM*. (2013).
24. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993 (2011).
25. Quinlan, A. R. & Hall, I. M. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
26. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).
27. Purcell, S. *et al.* PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).

Acknowledgements

GN is funded by an BBSRC-LiDO PhD studentship. JEP is funded by a Newton Institutional Links Grant (British Council, no. 261868591). TGC is funded by the Medical Research Council UK (Grant no. MR/M01360X/1, MR/N010469/1, MR/R025576/1, and MR/R020973/1). SC is funded by Medical Research Council UK grants (ref. MR/M01360X/1, MR/R025576/1, and MR/R020973/1). The authors declare no conflicts of interest.

Author contributions

J.E.P. and T.G.C. conceived and directed the project. A.S.K., A.J., M.T.K., S.A., M.Q., N.M., R.H., Z.H., S.C., S.A., B.K., S.J.W. and T.A.K. contributed data. G.N. performed bioinformatic and statistical analyses under the supervision of S.C., J.E.P. and T.G.C. G.N., S.C., J.E.P. and T.G.C. interpreted results. G.N. wrote the first draft of the manuscript with inputs from J.E.P. and T.G.C. All authors commented and edited on various versions of the draft manuscript and approved the final version. G.N., S.C., J.E.P., and T.G.C. compiled the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-022-11795-4.

Correspondence and requests for materials should be addressed to T.A.K., J.E.P. or T.G.C.

S1 Table: Strain-types

Lineage	Large sequence polymorphisms lineage	Spoligotype family	Region of difference no.	N	%
3	East-African-Indian	CAS	750	363	67.9
2.2.1	East-Asian (Beijing)	Beijing-RD181	105;207;181	29	5.4
4.5	Euro-American	H;T	122	23	4.3
4.9	Euro-American (H37Rv-like)	T1	None	22	4.1
1.1.2	Indo-Oceanic	EAI3;EAI5	239	14	2.6
3.1.2	East-African-Indian	CAS;CAS2	750	11	2.1
3.1.2.1	East-African-Indian	CAS2	750	10	1.9
3.1.3	East-African-Indian	CAS	750	9	1.7
4.8	Euro-American (mainly T)	T1;T2;T3;T5	219	9	1.7
1.2.2.2	Indo-Oceanic	NA	239	7	1.3
4.2.2.2	Euro-American (Ural)	T;LAM7-TUR	None	7	1.3
2.2.1.1	East-Asian (Beijing)	Beijing-RD150	105;207;181;150	5	0.9
3.1	East-African-Indian	Non-CAS1-Delhi	750	4	0.7
2.2.1.2	East-Asian (Beijing)	Beijing-RD142	105;207;181;142	2	0.4
4	Euro-American	LAM;T;S;X;H	None	2	0.4
4.1.1.1	Euro-American (X-type)	X2	183	2	0.4
4.1.1.3	Euro-American (X-type)	X1;X3	193	2	0.4
4.2.1	Euro-American (TUR)	H3;H4	None	2	0.4
4.2.2	Euro-American (Ural)	T;LAM7-TUR	None	2	0.4
4.6	Euro-American	T;LAM	None	2	0.4
4.6.5	Euro-American	T;LAM	None	2	0.4
1.1.3.3	Indo-Oceanic	NA	239	1	0.2
4.1.1.2	Euro-American (X-type)	X1	None	1	0.2
4.1.2.1	Euro-American (Haarlem)	T1;H1	182	1	0.2
4.6.2	Euro-American	T;LAM	726	1	0.2
4.6.2.1	Euro-American	T3	726	1	0.2
4.6.2.2	Euro-American (Cameroon)	LAM10-CAM	726	1	0.2

S2 Table: Drug-resistant samples according to drug susceptibility tests (DSTs) and genotypic predictions

Drug	DST N	DST resistant N	DST resistant %	Genotypic resistant* N	Genotypic resistant %	DST Susceptible Genotypic resistant	DST resistant Genotypic non-resistant
Rifampicin	487	417	85.6	460	86.0	6	6
Isoniazid	487	411	84.4	435	81.3	7	25
Ethambutol	479	265	55.3	385	72.0	96	17
Pyrazinamide	444	189	42.6	258	48.2	42	24
Streptomycin	43	24	55.8	238	44.5	4	6
Ofloxacin	85	46	54.1	277	51.8	5	0
Moxifloxacin	52	4	7.7	277	51.8	29	0
Levofloxacin	0	-	-	277	51.8	0	0
Amikacin	110	42	38.2	75	14.0	0	9
Kanamycin	112	44	39.3	79	14.8	0	7
Capreomycin	57	15	26.3	78	14.6	18	2
Ciprofloxacin	37	37	100	277	51.8	0	4
Ethionamide	37	6	16.2	102	19.1	11	0
PAS	0	-	-	10	1.9	0	0
Cycloserine	0	-	-	2	0.4	0	0
Clofazimine	0	-	-	1	0.2	0	0
Bedaquiline	0	-	-	1	0.2	0	0
Linezolid	0	-	-	0	0.0	0	0
Delamanid	0	-	-	0	0.0	0	0
Fluoroquinolones	174	87	50.0	277	51.8	-	-
Aminoglycosides	322	125	38.8	75	14.0	-	-

* from TB-Profiler; PAS para aminosalicylic acid; DST drug susceptibility test

S3 Table: Drug resistance (DR) categories by Lineage (L)

DR status	L1 N	L1 %	L2 N	L2 %	L3 N	L3 %	L4 N	L4 %	Total N	Total %
Sensitive	1	4.5	0	0.0	53	13.4	6	7.5	60	11.2
Pre-MDR	2	9.1	0	0.0	28	7.1	1	1.3	31	5.8
MDR	12	54.5	25	69.4	242	61.0	49	61.3	328	61.3
Pre-XDR	0	0.0	7	19.4	25	6.3	15	18.8	47	8.8
XDR	5	22.7	4	11.1	48	12.1	9	11.3	66	12.3
Other	2	9.1	0	0.0	1	0.3	0	0.0	3	0.6
Total	22	100	36	100	397	100	80	100	535	100.0

MDR multidrug resistant, XDR extensively drug resistant

S4 Table: Sensitivity analysis of the clustering by SNP distance

SNP distance	No. Clusters	N	Median (Maximum)	L1	L2	L3	L4	Sensitive	Pre MDR	MDR	Pre XDR	XDR	Other DR
0	28	60	2 (3)	2	6	35	17	2	2	17	19	20	0
1	28	60	2 (3)	2	6	35	17	2	2	17	19	20	0
5	49	136	2 (17)	7	16	77	36	2	3	60	29	40	2
10	55	169	2 (22)	7	21	98	43	2	3	87	31	44	2
15	54	176	2 (22)	8	21	103	44	2	4	90	32	46	2
20	63	200	2 (22)	8	24	121	47	2	5	106	35	49	3
25	68	213	2 (22)	8	24	131	50	2	5	118	35	50	3
30	71	220	2 (22)	8	25	137	50	2	5	124	35	51	3

L Lineage, MDR multidrug resistant, XDR extensively drug resistant, DR drug resistance

S*5* Table: Characteristics of 169 *M. tuberculosis* isolates in 55 clusters with a SNP distance of 10 compared to others

Characteristic		Trans. (169)	%	Non-trans. (366)	%	Odds ratio	95% CI	P-value
Lineage	1	7	4.1	15	4.1	1.00	-	-
	2	21	12.4	15	4.1	3.00	0.98 - 9.15	0.054
	3	98	58.0	299	81.7	0.70	0.28 - 1.77	0.454
	4	43	25.4	37	10.1	2.49	0.92 - 6.76	0.073
DR status	Sensitive/MDR	94	55.6	328	89.6	1.00	-	-
	Pre-XDR/XDR	75	44.4	38	10.4	5.79	3.67 - 9.14	$4.6x10^{-14}$
Location	Dera Ismail Khan	13	7.7	12	3.3	1.00	-	-
	Peshawar	46	27.2	31	8.5	1.37	0.55 - 3.39	0.497
	Other	30	17.8	31	8.5	0.89	0.35 - 2.27	0.812

Trans. transmitted

S6 Table: Genome-wide association analysis of transmission

Gene	Function	beta	95% CI	P-value
nusG	Transcription termination protein NusG	0.791	0.545 - 1.036	$5.8x10^{-10}$
Rv2307B	Hypothetical glycine rich protein	0.745	0.491 - 0.999	$1.5x10^{-8}$
wag31	Cell wall synthesis protein Wag31	0.912	0.567 - 1.256	$3.1x10^{-7}$
proX	Possible osmoprotectant binding lipoprotein ProX	0.706	0.423 - 0.988	$1.3x10^{-6}$
murA	Peptidoglycan biosynthesis pathway	0.660	0.380 - 0.939	$4.7x10^{-6}$

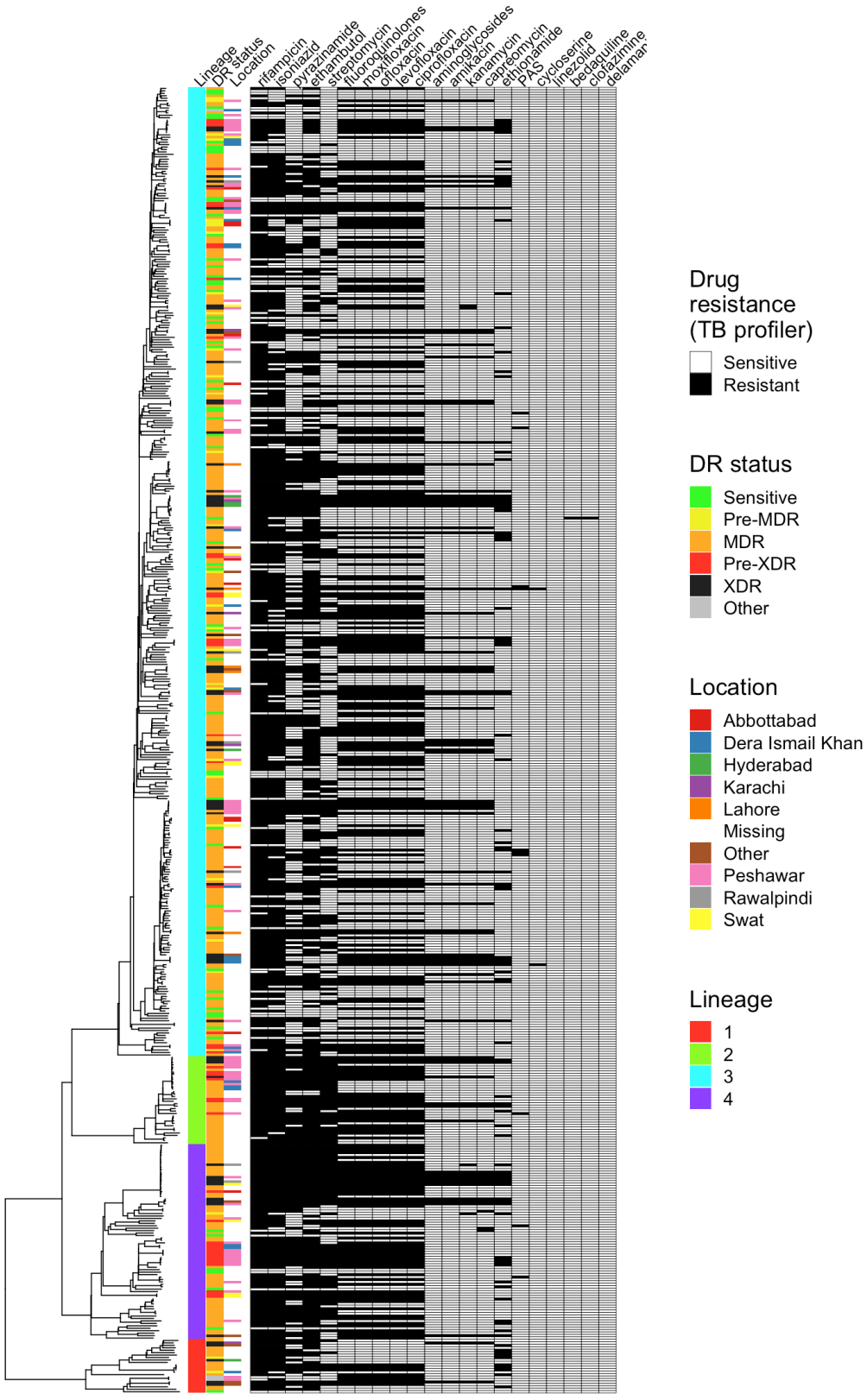
S7 Table: Number of samples with known drug resistance-associated mutations

S7_table_known_mutations.xls

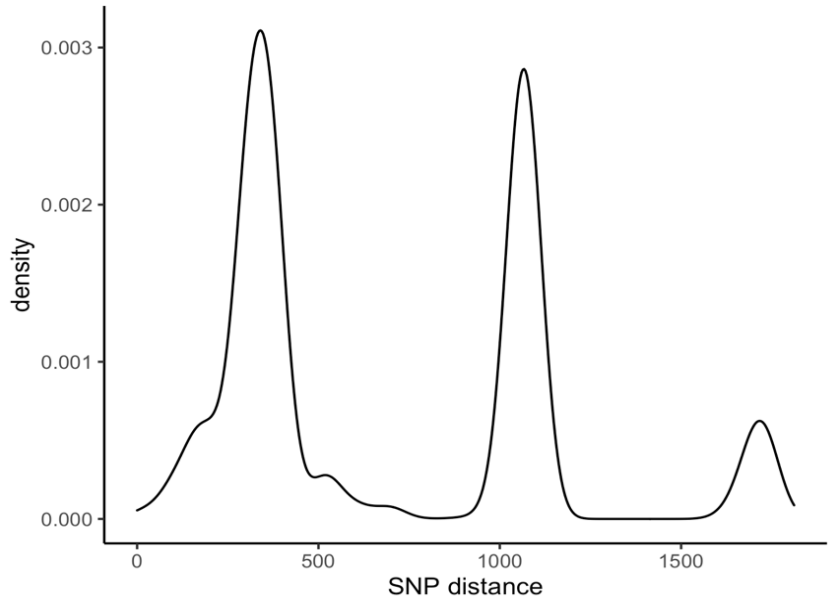
S8 Table: Phenotypically resistance samples (n=82) with variants previously unknown to be associated with drug resistance.

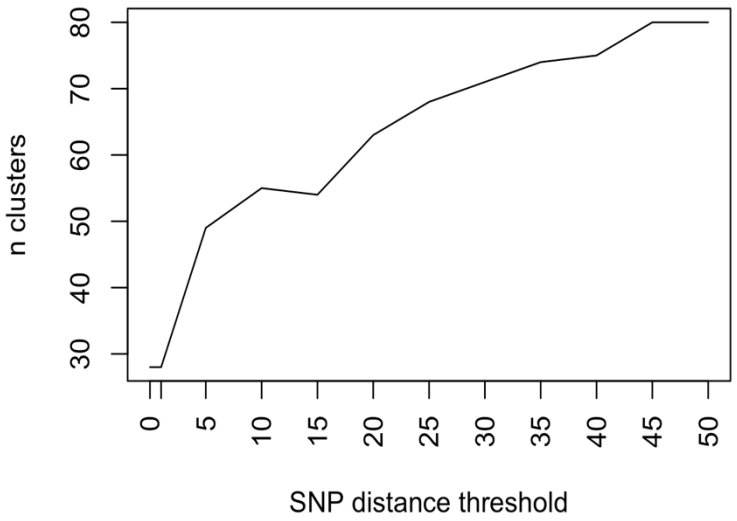
S8_table_novel_mutations.xlsx

S1 Figure: Phylogenetic tree for the 535 *M. tuberculosis* isolates with individual genomic drug resistance predictions.

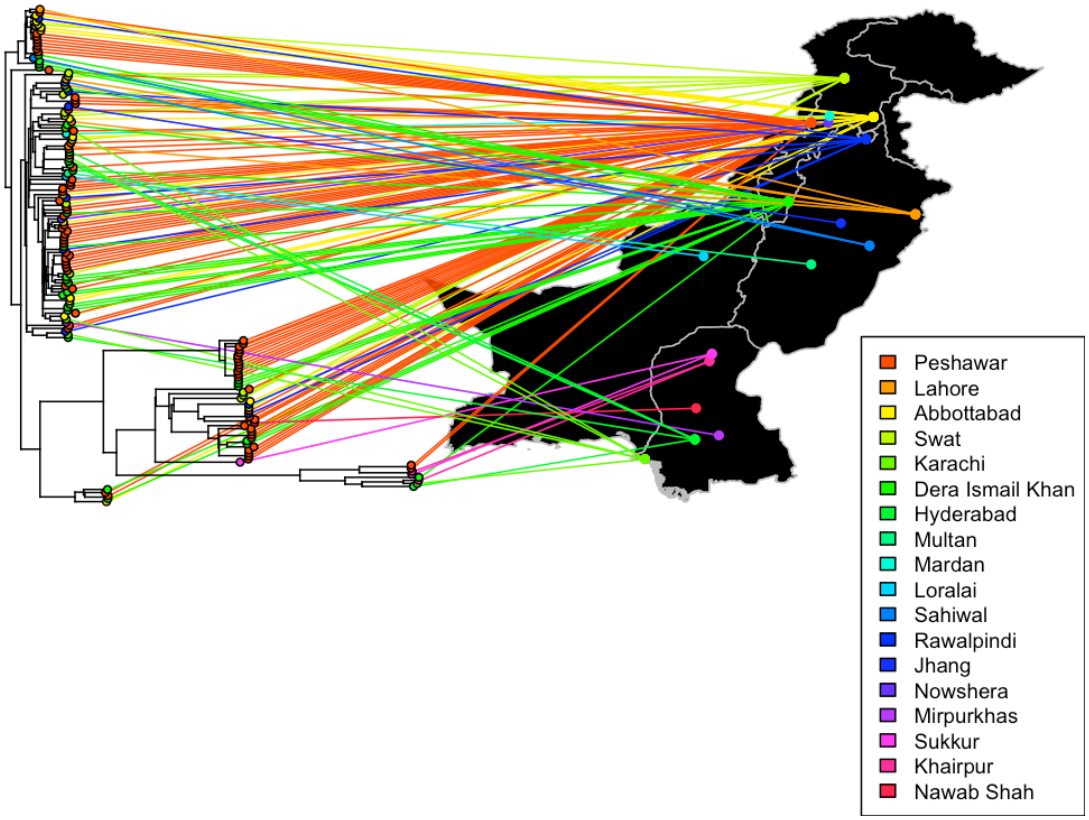


S2 Figure: SNP distance analyses and clusters (n=535). (top) Density of pairwise SNP differences for all samples; (bottom) number of clustering samples at minimum pairwise SNP difference thresholds

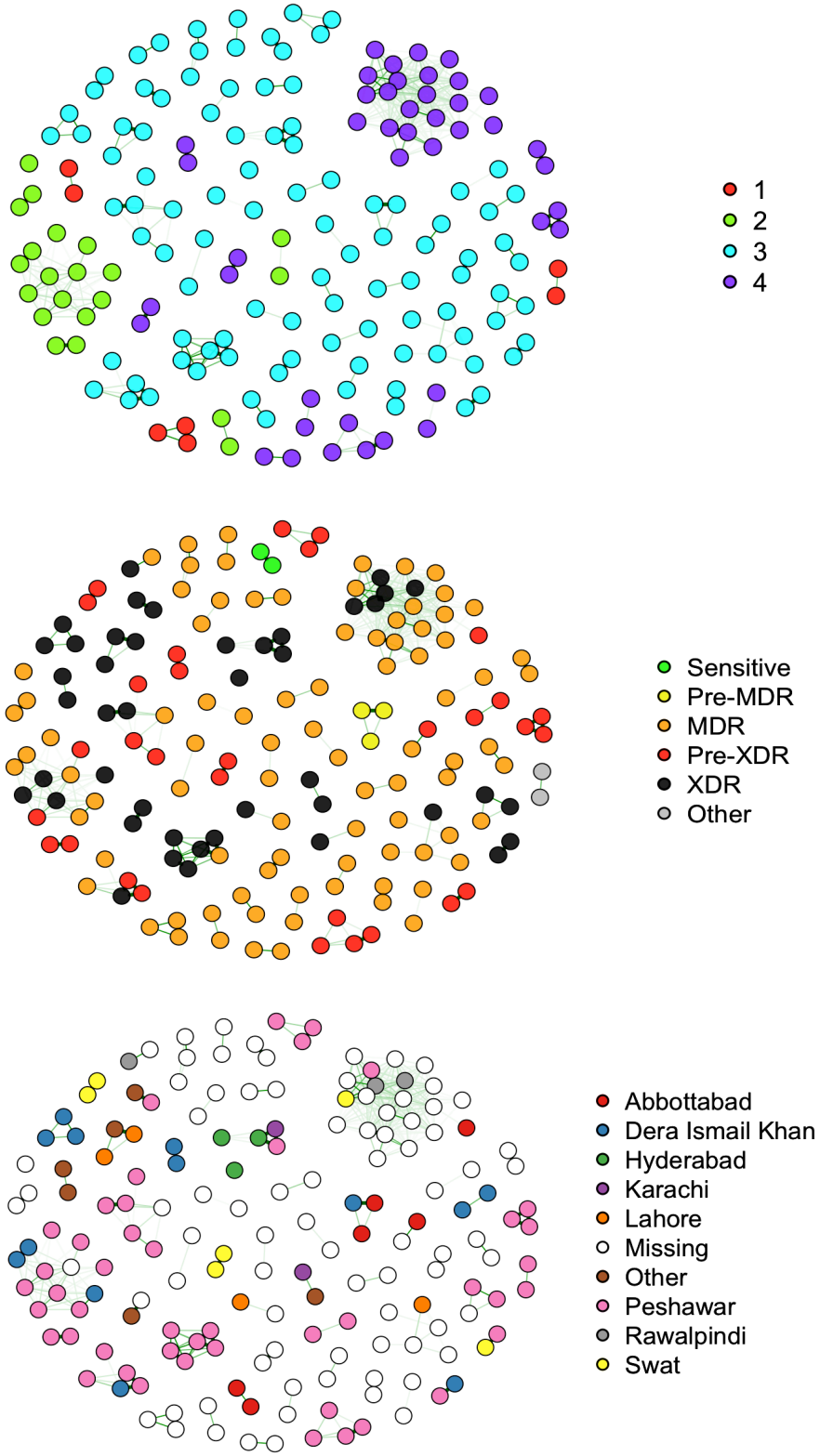




S3 Figure: Locations of samples in the transmission chains (n = 169)

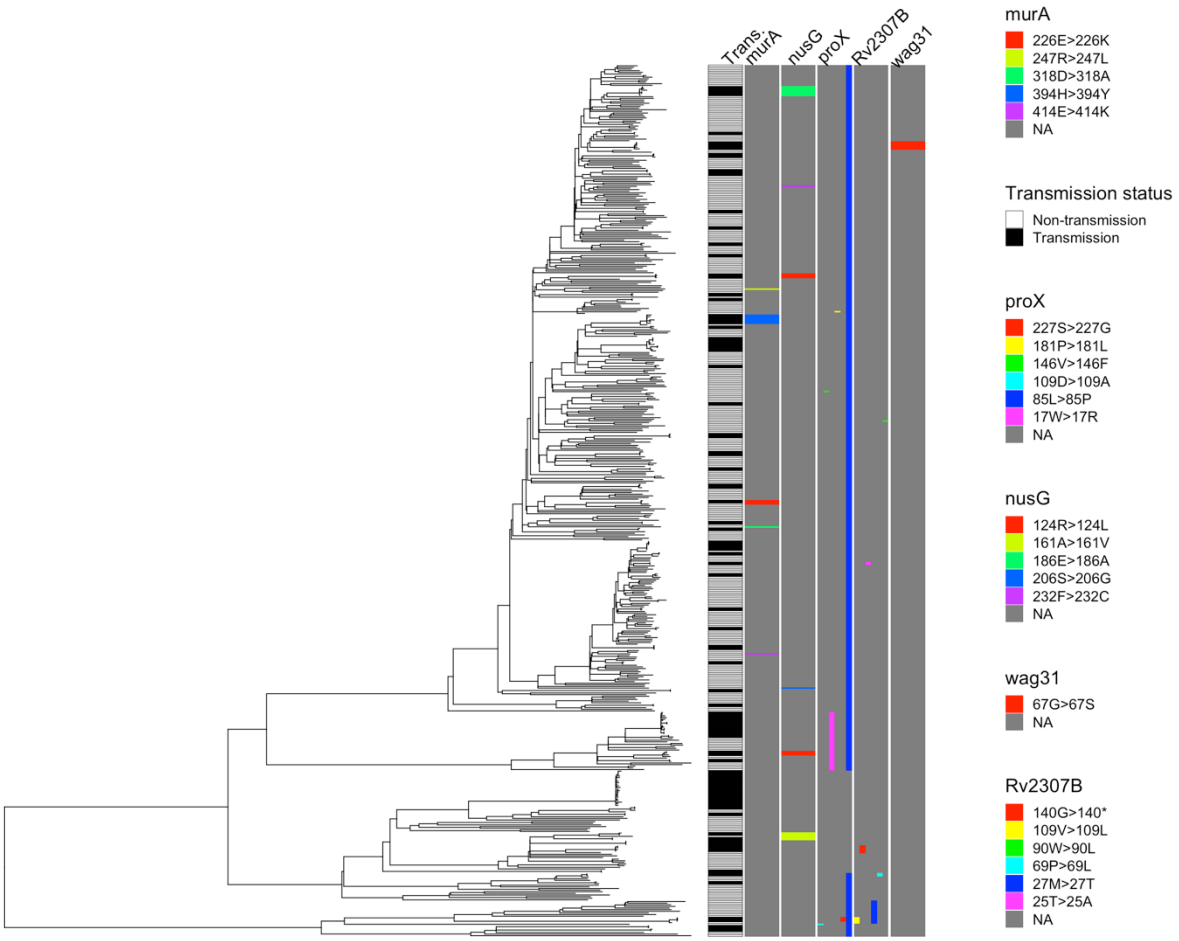


S4 Figure: The clusters with isolates at <= 10 SNP distance (n = 169), by lineage (top), drug resistance status (middle), and location (bottom).



S5 Figure

Phylogenetic location of mutations in genes compared with location of transmission samples.



LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

London School of Hygiene & Tropical Medicine
Keppel Street, London WC1E 7HT

T: +44 (0)20 7299 4646
F: +44 (0)20 7299 4656
www.lshtm.ac.uk

RESEARCH PAPER COVER SHEET

Please note that a cover sheet must be completed for each research paper included within a thesis.

SECTION A – Student Details

Student ID Number	1807750	**Title**	Mr
First Name(s)	Gary		
Surname/Family Name	Napier		
Thesis Title	Using whole genome sequencing data to identify strain-types, transmission enhancers and novel drug resistance mutations of *Mycobacterium tuberculosis*		
Primary Supervisor	Prof. Taane G. Clark		

If the Research Paper has previously been published please complete Section B, if not please move to Section C.

SECTION B – Paper already published

Where was the work published?	Scientific Reports		
When was the work published?	12/01/2023		
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion			
Have you retained the copyright for the work?*	Yes	Was the work subject to academic peer review?	Yes

*If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	
Please list the paper's authors in the intended authorship order:	
Stage of publication	Choose an item.

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	I performed the bioinformatic and statistical analysis, and wrote the first draft of the manuscript. I worked with co-authors on subsequent drafts and finalisation of the paper for submission.

SECTION E

Student Signature	
Date	18/01/2023

Supervisor Signature	
Date	18/01/2023

Chapter 5

Large-scale genomic analysis of *Mycobacterium tuberculosis* reveals extent of target and compensatory mutations linked to isoniazid, rifampicin and multi-drug resistance

scientific reports

Check for updates

OPEN

Large-scale genomic analysis of *Mycobacterium tuberculosis* reveals extent of target and compensatory mutations linked to multi-drug resistant tuberculosis

Gary Napier[1], Susana Campino[1], Jody E. Phelan[1,3]✉ & Taane G. Clark[1,2,3]✉

Resistance to isoniazid (INH) and rifampicin (RIF) first-line drugs in *Mycobacterium tuberculosis* (Mtb), together called multi-drug resistance, threatens tuberculosis control. Resistance mutations in *katG* (for INH) and *rpoB* (RIF) genes often come with fitness costs. To overcome these costs, Mtb compensatory mutations have arisen in *rpoC*/*rpoA* (RIF) and *ahpC* (INH) loci. By leveraging the presence of known compensatory mutations, we aimed to detect novel resistance mutations occurring in INH and RIF target genes. Across ~ 32 k Mtb isolates with whole genome sequencing (WGS) data, there were 6262 (35.7%) with INH and 5435 (30.7%) with RIF phenotypic resistance. Known mutations in *katG* and *rpoB* explained ~ 99% of resistance. However, 188 (0.6%) isolates had *ahpC* compensatory mutations with no known resistance mutations in *katG*, leading to the identification of 31 putative resistance mutations in *katG*, each observed in at least 3 isolates. These putative *katG* mutations can co-occur with other INH variants (e.g., *katG*-Ser315Thr, *fabG1* mutations). For RIF, there were no isolates with *rpoC*/*rpoA* compensatory mutations and unknown resistance mutations. Overall, using WGS data we identified putative resistance markers for INH that could be used for genotypic drug-resistance profiling. Establishing the complete repertoire of Mtb resistance mutations will assist the clinical management of tuberculosis.

Tuberculosis (TB), caused by *Mycobacterium tuberculosis* (Mtb) bacteria, is a major global public health problem. TB control is complicated by drug resistance, especially to first-line rifampicin (RIF) and isoniazid (INH), together called multi-drug resistance (MDR-TB). To acquire resistance to anti-TB drugs, Mtb drug targets or activating proteins are often mutated[1]. As a consequence, the biological function of these proteins is impaired or sometimes completely lost[2], causing the bacterium to incur a fitness cost. These costs can manifest as a phenotypic difference, such as reduced virulence or transmissibility. For example, the *katG* gene codes for the KatG enzyme, a catalase-peroxidase that protects the bacterium from reactive oxygen species damage and is used to detoxify hydrogen peroxide[3], improving survival within macrophages and the host immune response. The enzyme also activates the pro-drug INH, converting it to an active form[4].

Mutations in the *katG* gene that disrupt INH binding to KatG often leave Mtb drug resistant and a protein with impaired enzymatic function. In some cases, mutations can confer drug resistance without a punitive fitness cost. For example, the *katG* Ser315Thr mutation confers resistance but minimally affects fitness, hence is highly prevalent among (pre-)MDR-TB strains[5,6]. For RIF, the target is the β' subunit of RNA polymerase, coded by the *rpoB* gene. Mutations in *rpoB* prevent RIF from binding, but incur a high fitness cost since the intricate machinery of RNA polymerase is intolerant to large structural changes[7]. One notable exception is *rpoB* Ser450Leu, which is highly prevalent in RIF-resistant strains[8]. Indeed, so restrictive are changes to the β subunit, more than 95%

[1]Department of Infection Biology, Faculty of Infectious and Tropical Diseases, London School of Hygiene & Tropical Medicine, London WC1E 7HT, UK. [2]Faculty of Epidemiology and Population Health, London School of Hygiene & Tropical Medicine, London WC1E 7HT, UK. [3]These authors contributed equally: Jody E. Phelan and Taane G. Clark. ✉email: jody.phelan@lshtm.ac.uk; taane.clark@lshtm.ac.uk

of drug resistance mutations occur in the RIF resistance determining region (RRDR), an 81 base-pair section of the *rpoB* gene[9].

To overcome these fitness costs, secondary mutations can arise that improve or promote either the target protein itself or an alternative with a similar function. In the case of INH/*katG*, the expression of *ahpC*, which codes for a protein with similar enzymatic function, is often increased via mutations in the promoter of the *ahpC* gene[10,11]. RIF compensatory mutations occur in RNA polymerase subunits α (*rpoA*), β' (*rpoC*) or even within the β subunit (*rpoB*) itself. These mutations are thought to occur at the interfaces of the subunits, helping to restore overall RNA polymerase function, while maintaining an altered binding site in the β subunit[12].

The TB-Profiler platform[6] uses 2,300 mutations across 35 loci to profile Mtb resistance for 21 anti-TB drugs, including RIF and INH. However, the full repertoire of resistance mutations, including for MDR-TB is not fully characterised. The accompanying TB-Profiler database consists of ~ 32 k isolates with whole genome sequence and drug susceptibility test (DST) phenotypic data, with inferred genotypic profiles. Here, by investigating those isolates with compensatory mutations but no known resistance mutations, we aim to identify the presence of novel mutations linked to genes for INH, RIF, and therefore MDR-TB. Further, we attempt to understand the patterns of co-existence between resistance and compensatory mutations in relation to INH and RIF drug resistance.

Results

Isolate data. A total of 32,669 Mtb isolates with whole genome sequencing and DST data were analysed, and encompassed all major lineages (L4 51.1%, L2 25.3%, L3 11.5%, L1 9.7%) (Table 1). Across the 17,524 samples with DST data, 6262 (35.7%), 5435 (30.7%) and 5011 (28.6%) were phenotypically resistant to INH, RIF, and MDR-TB, respectively. Genotypic resistance prediction using TB-Profiler software inferred that 9546 (/32,669; 29.2%) and 7974 (24.4%), 5385 (16.5%) were resistant to INH, RIF, and MDR-TB, respectively (Table 1). The most common mutations underlying INH resistance were *katG* Ser315Thr (n = 7165; 21.9%), *fabG1* -15C>T (n = 1989; 6.1%), and *inhA* -154G>A (n = 332; 1.0%). Similarly, for the RIF resistance, the most frequent *rpoB* mutations were Ser450Leu (15.2%), Asp435Val (1.8%) and His445Tyr (1.3%) (Table S1).

To characterise putative novel resistance mutations, we considered samples that had a compensatory mutation, but no known resistance mutation. A manually curated list of established compensatory mutations (n = 33) (Table S2) covered *ahpC* (n = 18; e.g., -47_-46ins, -48G>A, -51G>A, -52C > A, -52C>T, -81C>T), *rpoC* (n = 13; e.g., Asn698Ser, Asp485Asn, Ile491Thr, Ile491Val, Leu516Pro, Trp484Gly, Val483Ala, Val483Gly), and *rpoA* (n = 2; e.g., Thr187Ala) loci. The number of occurrences of individual compensatory mutations within the 32 k isolates varied for *rpoC/A* (RIF, range: 5 – 427 isolates) and *ahpC* (INH, range: 3–97 isolates) genes. No isolate had more than one compensatory mutation for RIF or INH, and across MDR-TB.

Putative novel resistance mutations. Using the *rpoA* and *rpoC* compensatory mutations, there were no RIF resistant isolates without a known *rpoB* resistance mutation (Figure S1). For INH, there were 561 samples with a compensatory mutation, of which 188 (33.5%) had no known *katG* resistance mutation (Figure S1). Within the 188 samples we looked for mutations in *katG* that could potentially explain the emergence of the compensatory mutation. In total, 782 unique non-synonymous mutations were found in the *katG* gene. Only 31 (4.0%) of these *katG* mutations occurred in at least three isolates, and had > 50% of isolates with a resistant DST and genotypic resistance to at least one other drug. These 31 high-quality *katG* mutations were present in 171 isolates, including 64 and 107 with and without compensatory mutations, respectively (Table 2; Figure S1). Of the 188 isolates that had a compensatory mutation, 124 (66.0%) did not have any of the 31 highly quality *katG* mutations, but 86 (/124; 69.3%) were found to have rare *katG* mutations that did not pass the minimum frequency cut-off (> = 3) used to define putative resistance mutations (Table S3). These rare *katG* mutations could also potentially explain the acquisition of a compensatory mutation but were not analysed further.

Characteristic	–	N	%
Lineage	1	3154	9.7
	2	8257	25.3
	3	3745	11.5
	4	16,684	51.1
	Other	829	2.5
Genotypic status	Sensitive	19,587	60.0
	Rifampicin resistant	7974	24.4
	Isoniazid resistant	9546	29.2
	MDR-TB	5385	16.5
	Pre-XDR-TB	2085	6.4
	XDR-TB	16	0.1
	Other drug resistance	2558	7.8

Table 1. *Mycobacterium tuberculosis* isolates analysed (n = 32,669). MDR-TB = multi-drug resistant; XDR-TB = extensively drug resistant.

Change	Freq	Proportion Co-occurring with a resistance mutation	Proportion Co-occurring with a compensatory mutation	Distance from heme-binding site	Predicted stability change (ΔΔG)
Trp191Gly	25	0.920	0.160	27.912	−3.366
Ala109Thr	13	1	0.154	12.219	−1.546
Tyr98Cys	11	0	0.091	14.835	−2.197
Asp142Gly	9	0.222	0.222	16.149	−1.321
Gln439His	8	1	1	25.545	−0.839
Tyr413Cys	8	0.750	0.375	16.805	−1.399
Gly169Ser	7	1	0.286	12.856	−1.655
Gly299Ser	7	0.429	0.286	20.045	−1.463
Pro232Ser	7	0.714	0.429	8.781	−1.038
Thr677Pro	7	0.714	0.286	49.969	−0.712
Asp189Gly	6	0.500	0.333	25.272	−0.757
Asp419Tyr	6	0.500	0.167	21.015	−0.688
Arg484His	5	0.200	0.200	29.488	−2.107
Asn655Asp	5	1	0.600	55.642	−1.589
Phe183Leu	5	0.800	0.800	21.224	−0.833
Trp161Cys	5	0.600	1	21.583	−2.140
Ala312Glu	4	1	1	16.287	−1.663
Arg78Pro	4	0.500	0.500	25.94	−0.100
Asp189Asn	4	0.500	0.250	25.272	−0.899
Asp675Tyr	4	0.750	0.250	49.174	0.082
Glu233Gly	4	0.750	0.250	12.377	−1.163
Gly124Ser	4	1	1	21.677	−1.038
Ala122Asp	3	0.333	0.333	16.621	−1.207
Arg385Pro	3	0.667	0.667	16.767	−0.936
Gln88Pro	3	1	0.333	22.655	0.038
Gly182Arg	3	0.333	1	21.573	−0.851
Leu132Arg	3	1	0.333	16.532	−1.618
Met257Val	3	0.667	0.333	15.827	−1.681
Thr271Ile	3	0	0.333	9.640	−1.259
Thr326Pro	3	0.667	0.667	14.505	−0.341
Trp90Arg	3	0.667	0.667	22.822	−2.682

Table 2. List of 31 high-quality potential resistance mutations for isoniazid in the *katG* gene (171 samples).

Resistance and co-occurrence with other resistance mutations. The 31 putative INH-*katG* resistance mutations occurred in multiple lineages (L1-L5) with many showing evidence of convergent evolution (Fig. 1). These putative mutations occur in similar numbers of sub-lineages and at similar *katG* gene positions compared to known resistance mutations, indicating that they show comparable phylogenetic and gene location characteristics (Fig. 2). Due to the multi-drug regimens used for TB treatment, resistance often develops to multiple drugs in a stepwise manner[13]. The co-occurrence of the 31 *katG* mutations with other resistance mutations was analysed to characterise the isolate profiles in which putative resistance mutations occur. The 31 *katG* mutations were most frequently found in isolates characterised as MDR-TB (35.1%), but also common in pre-MDR-TB (26.3%) and pre-XDR-TB (29.2%) samples. Interestingly, around half (83/171; 48.5%) of isolates with any of the 31 *katG* mutations had co-occurrence with others linked to INH, with the *fabG1* -15C>T promoter mutation being the most frequent (60/171; 35.1%) (Table S4). This observation is in stark contrast to the *katG* Ser315Thr mutation, the most prevalent resistance mutation in INH resistant isolates and known to confer a high level of resistance, which only co-occurs with other INH resistance mutations in 16.1% of isolates. Of the 171 isolates with a putative resistance mutation (Figure S1), 107 (62.6%) had an available DST result for INH, with 99 reporting a resistant phenotype leading to a highly significant association between the putative drug resistance mutations and DST phenotype (Chi-squared $P < 1.4 \times 10^{-18}$). Of those with a resistant DST (n = 99), 53 (53.5%) had no other known mutations that could explain resistance.

Isolates with mutations conferring a high level of drug resistance tend to have low numbers of co-occurring resistance mutations linked to that resistance. As a proxy for measuring resistance level, we calculated the proportion of known and putative resistance mutation samples with co-occurring non-*katG* (*fabG1*, *inhA*, *kasA*) resistance mutations. Mutations at the *katG* 315 codon position, which are known to confer high resistance[14], had a relatively low proportion of isolates with co-occurring non-*katG* resistance mutations; four out of the five known codon 315 mutations have < 20% of isolates with co-occurring resistance mutations. There was no major difference in the number of co-occurring non-*katG* mutations between the putative (n = 31) and known resistance

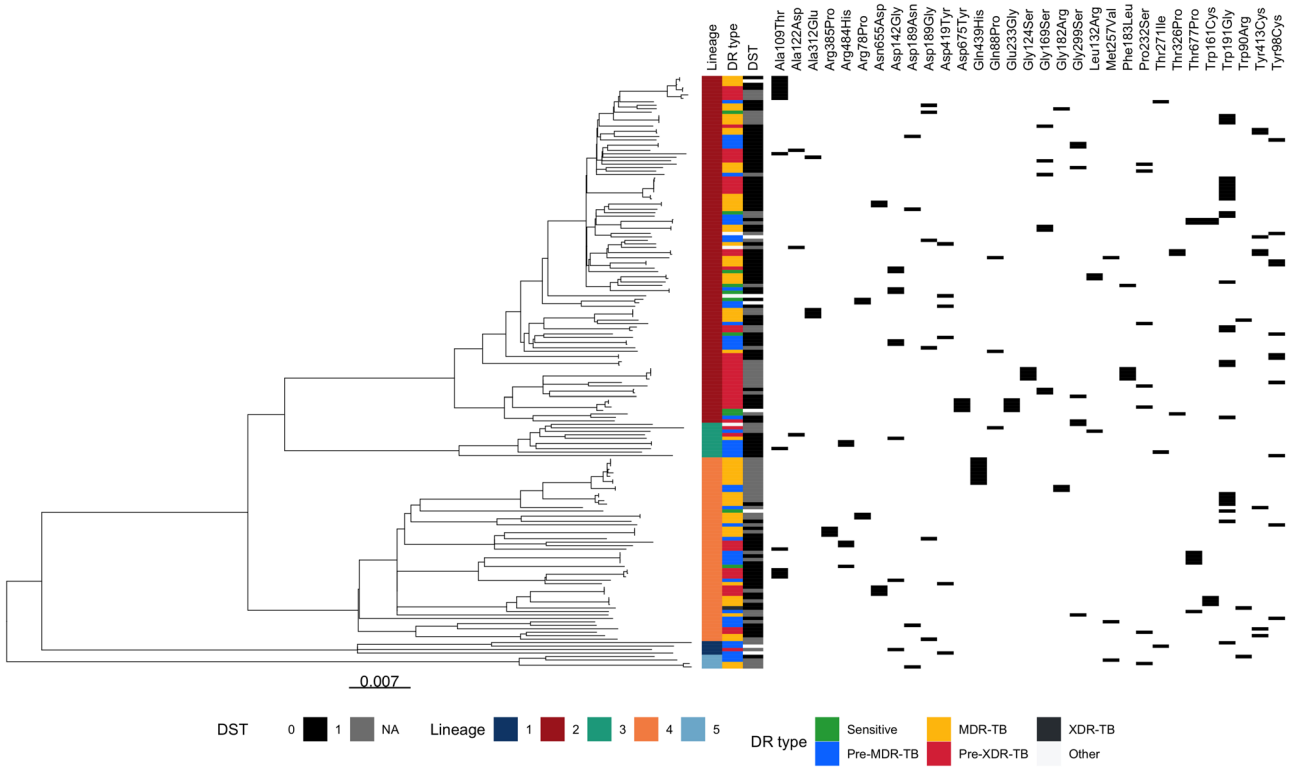


Figure 1. Phylogenetic tree of isolates (n = 171) with 31 putative novel *katG* gene mutations for Isoniazid resistance, with lineage, drug resistance (DR) status, and phenotypic drug susceptibility test (DST) data.

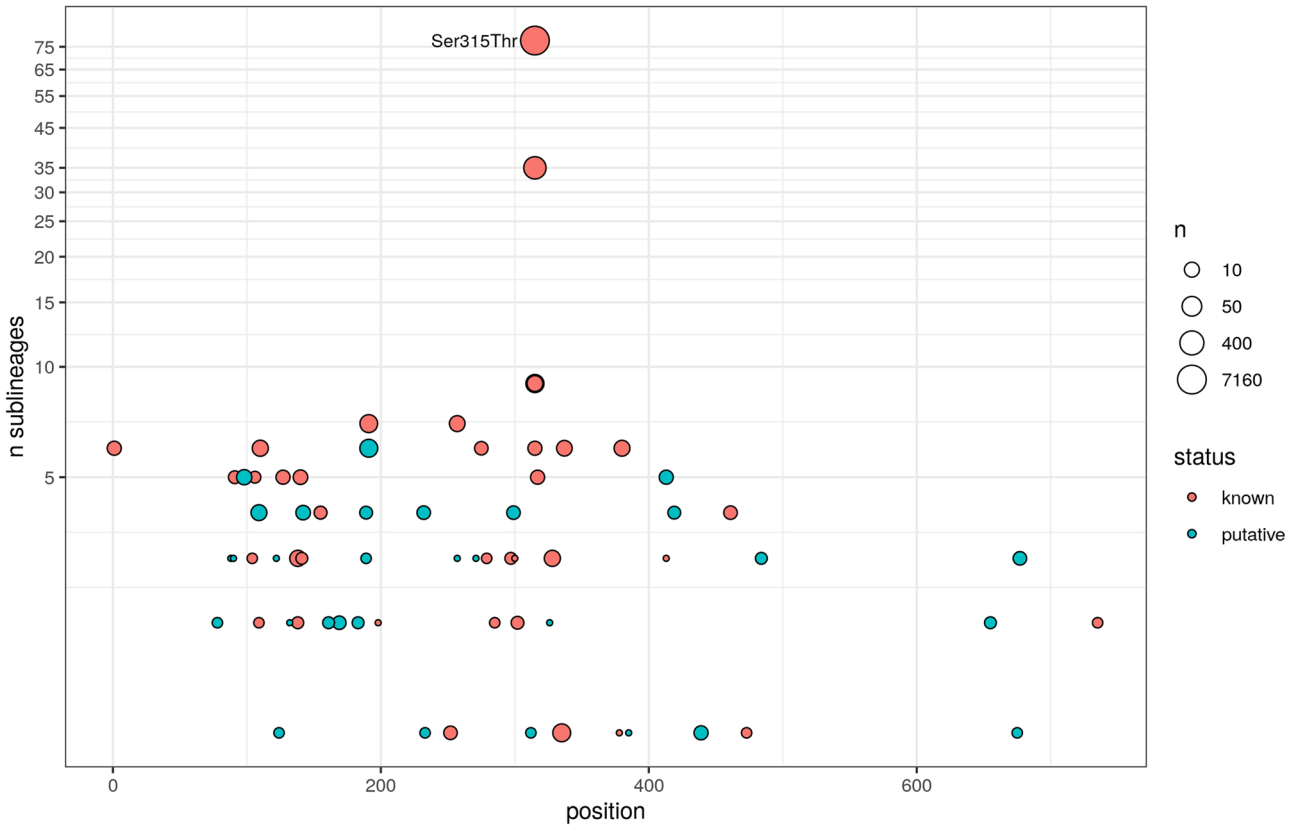


Figure 2. Homoplasy among 40 known *katG* and 31 putative resistance *katG* mutations. The common *katG* Ser315Thr mutation is highlighted. Mutations occurring in < 3 isolates and non-protein coding regions are omitted.

katG substitutions (n = 40; all > 2 isolates; Table S5) (mean resistance co-occurrence proportion: known 0.274 vs. putative mutations 0.413; T-test $P = 0.15$).

Mutation fitness. Compensatory mutations are linked to mutations with high fitness costs (e.g., *katG* loss of function (LOF)). To estimate the fitness impact of the putative resistance mutations, the frequency of co-occurrence with a compensatory mutation was calculated. As a proof of principle, this relationship was tested by comparing the frequency of compensatory mutations in samples containing LOF mutations against those that have SNP-based resistance mutations. Having a LOF mutation is associated with an increased risk of having a compensatory mutation (odds ratio = 13.86, Chi-squared $P < 0.0001$). In general, rarer mutations were observed to co-occur more frequently with compensatory mutations (Fig. 2). The proxy fitness cost was discretised into 'low,' 'medium' and 'high' categories based on tertiles (see Methods). The *katG* Ser315Thr is known to confer a low fitness cost[15], and was classified into the 'low' category, with only ~ 3% of samples containing the mutation co-occurring with a compensatory mutation. In fact, mutations at the codon 315 position appear to have low fitness cost (Fig. 2), where four out of the five known codon 315 resistance mutations were classified into the 'low' category.

Overall, compensatory mutations seem to occur in a higher proportion in isolates with the putative *katG* resistance mutations (0.388; 64/165) compared to Ser315Thr (0.031; 185/6010) (Chi-squared $P < 10^{-16}$), suggesting that on average they incur a greater fitness cost compared to this high frequency global mutation. Similarly, comparing to the 40 known resistance *katG* mutations from above (Table S5), there was a higher proportion of isolates with a compensatory mutation in those with the putative mutations (proportion of isolates with compensation mutation: known 0.026 vs. putative 0.389; Chi-squared $P = 6 \times 10^{-5}$). This difference remained statistically significant even when excluding the codon 315 positions (Chi-squared $P = 4 \times 10^{-4}$). Further, there appears to be little association between known non-*katG* resistance mutation co-occurrence (resistance level) and proxy fitness cost in both the 40 known and the 31 putative *katG* resistance mutations (Linear model $P = 0.073$) (Table S5). Also, across each of the three fitness cost categories (high, medium, and low), there were no strong differences in resistance mutation co-occurrence (resistance level) between isolates with the known and putative *katG* resistance mutations (minimum $P = 0.144$; Fig. 3). No strong differences in the co-variation between resistance level and fitness cost across known and putative mutations supports the veracity of our putative resistance variants. Interestingly, it has previously been observed that RIF-associated compensatory mutations in *rpoC* co-occur most frequently with *rpoB* Ser450Leu, which is the most common RIF resistance mutation and is thought to have a low fitness impact. This observation was also confirmed in our analysis, where 24.5% of the 4970 samples with the *rpoB* Ser450Leu mutation had a compensatory mutation. This was followed by Gln432Lys (19.2%), Val170Phe (16.4%), Gln432Leu (14.3%) and Pro454His (14.3%) (Table S6). Only Gln432Pro had a higher percentage co-occurrence, with 41.9% of 31 samples with this mutation also having a compensatory mutation.

Protein structure modelling. To explore the functional effects of the 31 putative resistance mutations, in silico predictions of their effects on the *katG* target protein were assessed (Fig. 3, Table 2). The estimable distances from the *katG* heme-binding site, thought to be close to the active INH binding site and crucial to enzymatic activity[16], did not differ significantly between known and putative substitution resistance mutations (Table S5) (mean distance: known 26.626 A^o vs. putative 22.058 A^o; Wilcoxon $P = 0.06$). There was no significant difference in protein stability change between the known resistance and putative *katG* mutations (mean $\Delta\Delta$ G: known − 1.078 vs. putative − 1.257; Wilcoxon $P = 0.24$).

Discussion

Our goal was to identify putative novel mutations underlying resistance to RIF and INH by finding isolates with established compensatory mutations. No novel *rpoB* gene mutations potentially linked to resistance to RIF were found, but this may be expected since there are limited ways in which the precise machinery of RNA polymerase can change without a loss of function. In contrast, many changes in the KatG protein can leave the bacteria largely unaffected. Our methodology flagged 31 mutations in *katG* that were analysed further. Evidence from available phenotypic DST data strongly suggests that the 31 *katG* mutations identified confer resistance. These mutations occur in multiple sub-lineages and independently in the phylogeny, a pattern of convergent evolution that is well established in known *katG* resistance mutations. Due to the relative rare occurrence of these mutations, they are either not present in the WHO catalogue or they have been designated as uncertain significance. However two of the mutations (Gly169Ser[17] and Asp142Gly[17,18]) were previously designated as likely to explain resistance in clinical isolates. Whilst our analysis focused on 31 high quality and frequent putative mutations in *katG*, less common mutations identified in one or two isolates may be of interest, including for functional evaluation and surveillance applications.

No significant differences were found in the proportion of isolates with (non-*katG*) known resistance mutation co-occurrence (our proxy for resistance level) between the filtered known (n = 40) and putative (n = 31) resistance mutations. In showing a similar pattern of resistance mutation co-occurrence we infer that the putative resistance mutations confer on average a similar level of resistance to known mutations, and this further supports their causal role with resistance. There was, however, a difference in the fitness cost between the known and putative resistance mutations, measured using compensatory mutation co-occurrence, with the latter appearing to have on average a higher cost. This observation is in agreement with previous studies, which report higher co-occurrence of *ahpC* promoter mutations with non-315 *katG* mutations compared to codon 315 mutations[19]. Whilst known resistance mutations are likely to converge on the most stable protein configurations and hence proliferate, the putative mutations are rarer and less likely to have been previously associated with drug resistance. The Interpretation of fitness cost and its relationship to compensatory mutations is less clear for *rpoB*/*C*/*A*

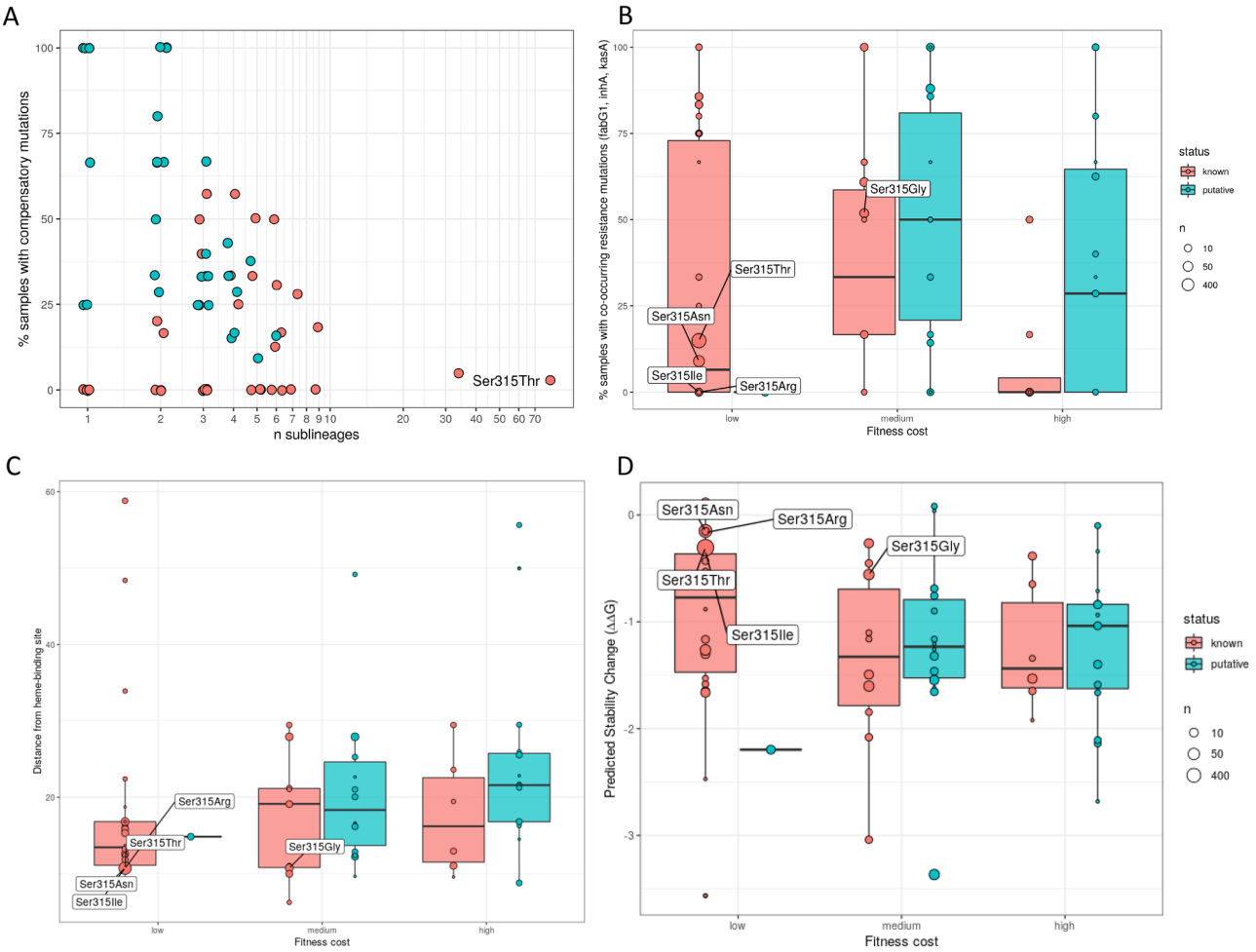


Figure 3. Comparison of mutation characteristics between putative and known resistance mutations. A) For each resistance mutation, the percentage of samples with a co-occurring compensatory mutation is plotted against the total number of sub-lineages it occurs in. B) Boxplot showing the percentage co-occurrence with other resistance mutations grouped by the discretised fitness categories (see Methods). Bottom boxplots show C) distance from INH heme binding site, and D) stability change distributions, grouped by the fitness categories.

(RIF) compared to *katG*/*ahpC* (INH). For example, *rpoB* Ser450Leu is thought to incur a minor fitness cost, yet compensatory mutations are found most frequently with this mutation. Conversely, *rpoB* Asp435Gly is described as having a 'severe' fitness cost[20], yet in our data none of the 90 samples with this mutation have compensatory *rpoC* mutations. Interestingly, three of these five mutations occur at position Gln432, indicating that mutations at this codon are heavily associated with having a compensatory mutation. There was no relationship between resistance level (using co-occurrence with other resistance mutations as a proxy) and fitness cost in either the 40 known or 31 putative filtered resistance mutations. Again, this similar pattern of variation indicates the veracity of the putative resistance mutations. Further, the functional impact of the 31 putative *katG* mutations is supported by in silico protein modelling, with distances to the functionally important *katG* heme active binding site similar to those of known variants, indicating that they are likely confer a similar pattern of resistance. In contrast to the differences between known and putative mutations in their percentages of isolates with compensatory mutations, surprisingly, there was no difference in the *in-silico* ΔΔG measure predictions. However, the ΔΔG measure is an indicator of protein stability, and therefore only an indirect indication of fitness cost.

There is the opportunity to apply a similar approach to other forms of Mtb drug resistance with a compensatory-resistance dynamic. This is especially true for non-essential targets that can exhibit multiple resistance mutations without a loss of function, similar to *katG*. For example, compensatory mutations for streptomycin are purported to restore translational accuracy of the ribosome, the target of the anti-TB drug[21]. Similarly, compensatory mutations have been found to act upon structures intolerant to change, including DNA gyrase subunit A (*gyrA* gene) for fluoroquinolones, and 16S rRNA of the 30S ribosome subunit (*rrs* gene) for aminoglycosides (e.g., capreomycin[20]). Ultimately, through identifying the full repertoire of resistance and compensatory mutations for anti-TB drugs, there will be improvements in clinical management and surveillance decision making using whole genome and amplicon sequencing data.

Conclusions

We have presented an approach to identify potential resistance mutations to monitor the development of resistance mechanisms to important first-line isoniazid and rifampicin anti-TB drugs, and therefore MDR-TB. The list of putative resistance mutations can inform functional studies of resistance, and after validation, be incorporated into genotypic drug resistance prediction, thereby informing clinical and infection control activities.

Material and methods

Input data and processing. The main input data consists of a database of 32 k isolates with DST and sequence data has been described previously[22]. Sequences were aligned to the H37Rv reference genome[23] (AL123456) with BWA mem (v0.7.17) software[24]. Joint SNP and indel calling was carried out in gatk GenotypeGVCFs (v4.1.3.0) software[25]. Monomorphic SNP/indel variants and those in non-unique regions of the genome (e.g., *ppe* genes) were excluded. Multi-FASTA alignments were created from the filtered variant and H37Rv reference fasta files using bedtools makewindows (v2.28.0)[26] and custom python scripts. Phylogenetic trees were constructed using IQ-TREE (v1.6.12) software, applying a general time reversible model with rate heterogeneity set to a discrete gamma model and an ascertainment bias correction (parameters – m GTR + G + ASC), with 1000 bootstrap samples[27]. Drug resistance types and lineages were predicted in-silico with TB-Profiler (v4.3.0) software[6,28]. TB-Profiler software was also used to identify all known drug resistance, compensatory and putative novel resistance mutations. Resistance patterns of samples were determined using phenotypic DSTs (available for 54% of samples) and predictions from TB-Profiler software (available for all samples). These resistance patterns were used to filter mutations (as described below). Known resistance mutations were defined based on the manually curated TBDB database (version commit: 4,738,132) which contains all WHO-endorsed mutations and additional ones reported in the literature.

Finding putative resistance markers using compensatory mutations. To improve the power of the analysis, novel compensatory mutations in *ahpC* were first characterised, as they are less well established than those in *rpoC/A*. From the sequence database (n = 32 k), all non-synonymous mutations present in at least three samples were found in *ahpC*. Although compensatory mutations do not cause resistance, they are strongly associated. Therefore, all mutations were filtered with requirements that > 50% samples were predicted resistant to INH by TB-Profiler and > 50% of samples were INH DST resistant. As there were many potential *ahpC* mutations, further filtering criteria were applied to these. Specifically, mutations were retained if they were associated with a loss of function in *katG* mutations, occurred in the same position as known *ahpC* mutations, and if they appeared in multiple lineages (convergent evolution). Only one of these criteria needed to be met to be considered a potential compensatory *ahpC* mutation. The full list of compensatory mutations consisted of 31 mutations (Table S1). A proxy for fitness cost was based on tertiles of the percentage of samples with compensatory mutations (low: < = 17%, medium: > 17% and < = 40%, high: > 40% and < = 100%). To find putative resistance mutations, all non-synonymous mutations in the relevant resistance genes were extracted from the TB-Profiler database (*katG* for INH, *rpoB* for RIF). Some variants known not to be associated with INH resistance[29] (e.g., *katG*-Arg463Leu) were excluded.

For each drug, mutations were found in samples where a compensatory mutation was present and known resistance mutations were absent in the relevant genes, but a non-resistance-associated mutation was present in the relevant target genes. These mutations were then filtered to exclude known drug-resistance-associated variants, and subjected to the same criteria as the putative compensatory mutations i.e., present in three or more samples, < 50% samples were predicted sensitive by TB-Profiler and < 50% of samples were DST sensitive. Mutations occurring in promoter regions of *rpoB* were excluded as candidate potential resistance mutations, as there is no known mechanism of resistance that could result from increased expression of the RNA polymerase beta subunit. Using this list of potential resistance mutations, all TB-Profiler database isolates (n = 32 k) were then searched for their presence, regardless of compensatory mutation status. It should be noted that therefore not all samples with potential resistance mutations necessarily have compensatory mutations, and vice versa.

Protein structural modelling. The open source software ChimeraX[30] was used to model distances from the INH heme binding site. Effects of mutations on protein stability were predicted using *in-silico* changes in Gibbs free energy ($\Delta\Delta G$) by mCSM[31] software.

Data availability

All genomic data is available on the short read archive (https://www.ebi.ac.uk/ena/browser/). Code and accessions used in the study can be found at https://github.com/GaryNapier/comp_mut. Analysis scripts are available at https://github.com/AntonS-bio.

Received: 26 August 2022; Accepted: 3 January 2023
Published online: 12 January 2023

References

1. Phelan, J. *et al. Mycobacterium tuberculosis* whole genome sequencing and protein structure modelling provides insights into anti-tuberculosis drug resistance. *BMC Med.* **14**, (2016).
2. Gomes, L. C., Campino, S., Marinho, C. R. F., Clark, T. G. & Phelan, J. E. Whole genome sequencing reveals large deletions and other loss of function mutations in Mycobacterium tuberculosis drug resistance genes. *Microb. Genomics* **7**, 000724 (2021).
3. Trivedi, A., Singh, N., Bhat, S. A., Gupta, P. & Kumar, A. Redox biology of tuberculosis pathogenesis. *Adv. Microb. Physiol.* **60**, 263–324 (2012).

4. Wang, J. Y., Burger, R. M. & Drlica, K. Role of superoxide in catalase-peroxidase-mediated isoniazid action against mycobacteria. *Antimicrob. Agents Chemother.* **42**, 709–711 (1998).
5. Hazbón, M. H. *et al.* Population Genetics Study of Isoniazid Resistance Mutations and Evolution of Multidrug-Resistant Mycobacterium tuberculosis. *Antimicrob. Agents Chemother.* **50**, 2640 (2006).
6. Phelan, J. E. *et al.* Integrating informatics tools and portable sequencing technology for rapid detection of resistance to anti-tuberculous drugs. *Genome Med.* **11**, 41 (2019).
7. Song, T. *et al.* Fitness costs of rifampicin-resistance in Mycobacterium tuberculosis are amplified under conditions of nutrient starvation and compensated by mutation in the β' subunit of RNA polymerase. *Mol. Microbiol.* **91**, 1106 (2014).
8. Heep, M. *et al.* Frequency of rpoB mutations inside and outside the cluster I region in rifampin-resistant clinical Mycobacterium tuberculosis isolates. *J. Clin. Microbiol.* **39**, 107–110 (2001).
9. Cao, Y. *et al.* Automatic Identification of Individual rpoB gene mutations responsible for rifampin resistance in *Mycobacterium tuberculosis* by use of melting temperature signatures generated by the Xpert MTB/RIF ultra assay. *J. Clin. Microbiol.* **58**, e00907 (2019).
10. Sherman, D. R. *et al.* Compensatory ahpC gene expression in isoniazid-resistant *Mycobacterium tuberculosis*. *Science* **272**, 1641–1643 (1996).
11. Sherman, D. R., Mdluli, K., Hickey, M. J., Barry, C. E. & Stover, C. K. AhpC, oxidative stress and drug resistance in *Mycobacterium tuberculosis*. *BioFactors* **10**, 211–217 (1999).
12. Comas, I. *et al.* Whole-genome sequencing of rifampicin-resistant *Mycobacterium tuberculosis* strains identifies compensatory mutations in RNA polymerase genes. *Nat. Genet.* **44**, 106–110 (2012).
13. Libiseller-Egger, J., Phelan, J., Campino, S., Mohareb, F. & Clark, T. G. Robust detection of point mutations involved in multidrug-resistant *Mycobacterium tuberculosis* in the presence of co-occurrent resistance markers. *PLOS Comput. Biol.* **16**, e1008518 (2020).
14. Coll, F. *et al.* Genome-wide analysis of multi- and extensively drug-resistant *Mycobacterium tuberculosis*. *Nat. Genet.* **2018**(50), 307–316 (2018).
15. Pym, A. S., Saint-Joanis, B. & Cole, S. T. Effect of katG mutations on the virulence of *Mycobacterium tuberculosis* and the implication for transmission in humans. *Infect. Immun.* **70**, 4955–4960 (2002).
16. Munir, A. *et al.* Using cryo-EM to understand antimycobacterial resistance in the catalase-peroxidase (KatG) from *Mycobacterium tuberculosis*. *Structure* **29**, 899-912.e4 (2021).
17. Kandler, J. L. *et al.* Validation of novel *Mycobacterium tuberculosis* isoniazid resistance mutations not detectable by common molecular tests. *Antimicrob. Agents Chemother.* **62**, e00974 (2018).
18. Torres, J. N. *et al.* Novel katG mutations causing isoniazid resistance in clinical *M. tuberculosis* isolates. *Emerg. Microbes Infect.* **4**, e42 (2015).
19. Liu, L. *et al.* The impact of combined gene mutations in inhA and ahpC genes on high levels of isoniazid resistance amongst katG non-315 in multidrug-resistant tuberculosis isolates from China. https://doi.org/10.1038/s41426-018-0184-0 (2018).
20. Alame Emane, A. K., Guo, X., Takiff, H. E. & Liu, S. Drug resistance, fitness and compensatory mutations in *Mycobacterium tuberculosis*. *Tuberculosis (Edinb.)* **129**, 102091 (2021).
21. Björkman, J., Hughes, D. & Andersson, D. I. Virulence of antibiotic-resistant *Salmonella typhimurium*. *Proc. Natl. Acad. Sci. U. S. A.* **95**, 3949–3953 (1998).
22. Napier, G. *et al.* Robust barcoding and identification of Mycobacterium tuberculosis lineages for epidemiological and clinical studies. *Genome Med.* https://doi.org/10.1186/s13073-020-00817-3 (2020).
23. Cole, S. T. *et al.* Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* **393**, 537–544 (1998).
24. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
25. Depristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **2011**(43), 491–498 (2011).
26. Quinlan, A. R. & Hall, I. M. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
27. Minh, B. Q. *et al.* IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* https://doi.org/10.1093/molbev/msaa015 (2020).
28. Coll, F. *et al.* Rapid determination of anti-tuberculosis drug resistance from whole-genome sequences. *Genome Med.* **7**, 51 (2015).
29. Van Doorn, H. R. *et al.* The susceptibility of *Mycobacterium tuberculosis* to isoniazid and the Arg–>Leu mutation at codon 463 of katG are not associated. *J. Clin. Microbiol.* **39**, 1591–1594 (2001).
30. Pettersen, E. F. *et al.* UCSF ChimeraX: Structure visualization for researchers, educators, and developers. *Protein Sci.* **30**, 70–82 (2021).
31. Pires, D. E. V., Ascher, D. B. & Blundell, T. L. mCSM: Predicting the effects of mutations in proteins using graph-based signatures. *Bioinformatics* **30**, 335–342 (2014).

Acknowledgements

GN is funded by an BBSRC-LiDO PhD studentship. JEP is funded by a Newton Institutional Links Grant (British Council, no. 261868591). TGC is funded by the Medical Research Council UK (Grant no. MR/M01360X/1, MR/N010469/1, MR/R025576/1, MR/R020973/1, and MR/X005895/1). SC is funded by Medical Research Council UK grants (ref. MR/M01360X/1, MR/R025576/1, and MR/R020973/1). The authors declare no conflicts of interest. The funders had no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript should be declared.

Author contributions

J.E.P. and T.G.C. conceived and directed the project. G.N. performed bioinformatic and statistical analyses under the supervision of S.C., J.E.P. and T.G.C. G.N., S.C., J.E.P. and T.G.C. interpreted the results. G.N. wrote the first draft of the manuscript with inputs from J.E.P. and T.G.C. All authors commented and edited on various versions of the draft manuscript and approved the final version. G.N., J.E.P., and T.G.C. compiled the final manuscript. All authors have consented to the publication of this manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-023-27516-4.

Correspondence and requests for materials should be addressed to J.E.P. or T.G.C.

SUPPLEMENTARY TABLES

Table S1

Common established resistance mutations for INH and RIF*. Mutations occurring in <5 isolates in the 32k dataset are omitted.

Drug	Gene	Mutation [frequency]
isoniazid	*katG*	Ser315Thr [7165], Ser315Asn [167], Ser315Gly [27], Ile335Val [25], Trp191Arg [25], Al [14], Asn138His [14], Trp328Leu [14], Ser315Ile [13], Thr380Ile [13], 371del [12], Met Tyr337Cys [12], 1284del [9], Ser140Asn [9], Gln127Pro [8], Ile317Val [8], Ser315Arg [8 [8], Gln461Pro [7], Phe252Leu [7], Thr275Ala [7], Tyr155Cys [7], Ser302Arg [6], Trp91 Tyr155Ser [6], 18dup [5], Ala106Val [5], Asn138Ser [5], Gly297Val [5], Leu141Phe [5]
	fabG1	-15C>T [1989], -17G>T [173], -8T>A [159], -8T>C [154], -8T>G [11]
	inhA	-154G>A [332], Ser94Ala [234], Ile194Thr [191], Ile21Thr [103], Ile21Val [94]
	ahpC	Asp73His [50], Glu76Lys [14]
rifampicin	*rpoB*	Ser450Leu [4970], Asp435Val [592], His445Tyr [410], His445Asp [293], Asp435Tyr [277], Leu452Pro [251], Glu761Asp [227], Leu430Pro [140], Ser450Trp [130], Ile491Phe [126], His445Leu [101], His445Arg [98], Asp435Gly [95], His445Asn [95], Val170Phe [73], Ser450Phe [46], Asp435Phe [41], Ser441Leu [39], His445Cys [36], Ala286Val [33], Gln432Pro [31], Ile480Val [29], Gln432Lys [26], Met434Ile [25], 1297_1299dup [24], Gln432Leu [21], Ser441Gln [21], Thr400Ala [21], His445Gln [20], Leu430Arg [18], Asp435Ala [14], 1296_1304del [12], Asn437Asp [12], Gln432Glu [10], 1329_1331dup [9], Gln429His [9], His445Gly [9], Ser431Gly [9], Ser450Gln [9], His445Ser [8], Ser428Arg [8], 1312_1314del [7], Asp435Glu [7], Phe424Val [7], Pro454His [7], Pro454Leu [7], Ser493Leu [7], Ala451Val [6], Glu460Gly [6], Met434Val [6], Phe424Leu [6], 1287_1295del [5], 1300_1305del [5], Gln429Leu [5], Ser428Gly [5]

* from TB-Profiler

Table S2

Mutations in compensatory genes and their frequencies in 32k samples. All mutations occur in >1 lineage.

Drug	Compensatory locus	Mutation [frequency]
isoniazid	*ahpC*	-47_-46ins [49*], -48G>A [92], -51G>A [44*], -52C>A [49*], -52C>T [97], -54C>T [33], -57C>T [26], -72C>T [37*], -74G>A [9*], -75T>G [3*], -76T>A [25*], -76T>G [5*], -77del [9*], -77T>A [6*], -77T>G [10*], -81C>T [57], -88_-87ins [4*], -90G>A [8*]
rifampicin	*rpoA*	Thr187Ala [59], Thr187Pro [6]
	rpoC	Asn698His [10], Asn698Lys [17], Asn698Ser [125], Asp485Asn [48], Asp485His [5], Ile491Thr [107], Ile491Val [161], Leu516Pro [66], Phe452Leu [25], Pro434Arg [11], Trp484Gly [53], Val483Ala [135], Val483Gly [427]

[frequencies]; ins = insertion, del = deletion; * novel markers with strong evidence for compensatory effects through convergent evolution, co-occurrence with loss of function mutations in *katG* as well as association with INH resistant isolates.

Table S3

Less frequent mutations (n=84) in *katG* (<3 isolates) in 86 isolates with no known isoniazid resistance mutations, with compensatory mutations but no potential resistance mutation**

Change	Frequency	# Co-occurring with a resistance mutation	# Co-occurring with a compensatory mutation	Distance from heme-binding site	Predicted Stability Change (ΔΔG)
Leu43Arg	2	0	2	38.063	-1.520
Leu48Arg	1	0	1	35.410	-1.260
Leu76Pro	1	0	1	24.637	-1.373
Thr86Pro	2	0	2	22.091	-0.490
Ala93Thr	2	0	2	17.033	-1.007
His97Pro	1	0	1	15.693	-0.074
Ile103Val	2	0	1	7.7860	-1.467
Gly111Asp	1	0	1	11.404	-1.512
Ala122Val	1	0	1	18.105	-0.689
Gly124Ala	1	0	1	21.029	-0.723
Phe129Ser	1	0	1	20.644	-2.619
Trp135*	1	0	1	-	-
Pro136Leu	2	2	2	11.891	-0.267
Leu141Val	1	1	1	13.825	-1.872
Asp142Asn	1	0	1	16.634	-1.387
Lys143Asn	1	0	1	16.183	-1.704
Arg145Ser	1	1	1	18.166	-2.084
Gly156Asp	1	0	1	29.539	-1.595
Ala162Val	3	1	1	19.634	-0.804
Asp163Asn	1	0	1	19.888	-0.609
Asp163Ala	1	0	1	19.888	-0.788
Ile165Thr	1	0	1	15.343	-2.971
Phe167Ser	1	0	1	17.090	-3.291
Leu173Arg	2	0	2	15.564	-1.872
Gly184Asp	1	0	1	23.169	-2.176
Gly186Ser	2	0	2	25.211	-1.465
Gly186Asp	1	0	1	25.211	-1.924
Met225Ile	1	1	1	16.258	-0.154
Thr251Lys	1	1	1	13.673	-0.467
Arg253Trp	2	2	2	17.457	-0.390
Thr262Pro	2	0	2	11.530	-0.237
Ala264Val	2	2	1	10.974	0.230

Gly273Arg	4	0	1	8.9660	-0.981
His276Gln	1	0	1	12.199	-0.662
Glu289Ala	2	2	1	24.324	-0.982
Gly299Asp	1	0	1	19.291	-1.975
Ala312Val	1	0	1	15.602	-0.590
Thr324Leu	1	0	1	14.819	-0.392
Pro325Ser	2	0	1	12.972	-2.410
Trp328Arg	1	0	1	16.309	-2.336
Asp329Ala	2	1	1	16.298	-0.523
Asp329Glu	2	0	2	16.298	-0.604
Glu342Gly	1	0	1	19.479	-1.412
Thr344Ser	1	0	1	18.705	-1.089
Ser383*	1	1	1	-	-
Thr394Pro	1	0	1	18.440	-0.410
His400Pro	1	0	1	21.679	0.426
Phe408Ser	1	0	1	14.601	-2.775
Ala411Asp	1	0	1	14.011	-2.547
Tyr413Ser	1	0	1	16.985	-3.198
Asp419Val	1	0	1	21.844	0.183
Pro422Leu	1	0	1	28.127	-0.371
Tyr426*	1	0	1	-	-
Leu458His	1	0	0	55.573	-2.819
Ile462Ser	1	1	1	56.503	-3.388
Ala476Glu	1	0	1	43.138	-2.446
Ala478Arg	1	0	1	39.741	-0.866
Ala480Gln	2	0	2	37.360	-1.471
Phe483Leu	2	0	1	31.498	-1.654
Lys488Glu	2	0	2	27.512	-0.682
Gly490Asp	2	0	1	32.196	-0.930
Gly494Ala	2	0	2	37.761	-0.959
Gly495Ser	8	0	1	39.032	-1.750
Gly495Cys	2	0	2	39.032	-1.578
Pro501Ser	2	2	2	35.629	-2.376
Leu521Pro	1	0	1	50.801	-1.520
Gly560Arg	1	0	1	62.465	-0.240
Thr568Pro	2	2	1	44.809	-0.224
Pro569Leu	1	0	1	44.191	-0.400
Asp612Gly	1	0	1	32.396	-0.696
Ala621Asp	1	0	1	42.314	-2.436
Thr625Lys	2	0	2	46.002	0.112
Leu627Pro	1	0	1	45.368	-0.904

Gly630Arg	1	0	1	49.604	-0.760
Gly644Asp	2	1	1	52.172	-1.423
Asp663Tyr	2	1	1	59.875	-0.018
Gln679Tyr	1	0	1	57.006	-0.175
Ser700Phe	1	1	1	48.170	-0.948
Arg705Trp	1	0	1	50.045	-1.636
Val708Asp	1	0	1	52.230	-2.899
Tyr711Asp	2	0	2	54.722	-3.780
Asp723Asn	1	0	1	48.405	-1.170
Asp735Tyr	1	0	1	34.481	-0.253
Arg736Lys	1	0	1	35.676	-1.420

* stop codon; ** see Methods for definition

Table S4

Known resistance and other mutations co-occurring in isolates with the 31 putative drug resistance mutations

Putative resistance mutations (*katG*)	Known resistance mutations	Other mutations	n
Ala109Thr	fabG1-15C>T	-	7
Ala109Thr	fabG1-17G>T	-	1
Ala109Thr	fabG1-8T>A	katG-Val697Ala	3
Ala109Thr	inhA-Ile194Thr; fabG1-15C>T	-	2
Ala122Asp	inhA-154G>A	-	1
Ala122Asp	-	-	2
Ala312Glu	-	kasA-Val142Ile	1
Ala312Glu	-	-	3
Arg385Pro	-	-	3
Arg484His	-	kasA-Gly269Ser	1
Arg484His	-	-	4
Arg78Pro	-	-	4
Asn655Asp	fabG1-15C>T	-	2
Asn655Asp	-	-	3
Asp142Gly	katG-Ser140Asn	-	1
Asp142Gly	-	-	8
Asp189Asn	fabG1-15C>T	-	2
Asp189Asn	-	-	2
Asp189Gly	fabG1-15C>T	-	1
Asp189Gly	-	katG-Ser446Asn	1
Asp189Gly	-	-	4
Asp419Tyr	fabG1-15C>T	-	1
Asp419Tyr	fabG1-8T>C	-	1
Asp419Tyr	-	kasA-Gly312Ser	1
Asp419Tyr	-	-	3
Asp675Tyr; Glu233Gly	fabG1-15C>T	katG-Thr380Ala	2
Asp675Tyr; Glu233Gly	-	-	1
Asp675Tyr; Pro232Ser; Glu233Gly	-	-	1
Gln439His	fabG1-15C>T	-	8
Gln88Pro	fabG1-15C>T	katG-Lys600Gln	1

Gln88Pro	inhA-154G>A	-	1
Gln88Pro; Met257Val	-	kasA-His253Tyr	1
Gly169Ser	fabG1-15C>T	-	6
Gly169Ser	-	-	1
Gly182Arg	-	-	3
Gly299Ser	fabG1-15C>T	-	1
Gly299Ser	-	katG-Gln525Leu	1
Gly299Ser	-	-	5
Leu132Arg	fabG1-15C>T	katG-Val246Gly	1
Leu132Arg	fabG1-15C>T	-	2
Met257Val	katG-Gln461Pro	-	1
Met257Val	-	katG-Tyr28Leu	1
Phe183Leu	-	-	1
Phe183Leu; Gly124Ser	fabG1-15C>T	-	4
Pro232Ser	fabG1-15C>T	-	2
Pro232Ser	-	katG-Asp419Gly	1
Pro232Ser	-	katG-Gln295Glu	1
Pro232Ser	-	-	2
Thr271Ile	-	-	3
Thr326Pro	-	-	1
Thr326Pro; Tyr413Cys	fabG1-8T>C	-	2
Thr677Pro	inhA-154G>A	-	1
Thr677Pro	katG-Ser315Thr	-	4
Thr677Pro; Trp161Cys	-	-	2
Trp161Cys	-	-	3
Trp191Gly	fabG1-15C>T	katG-Val320Ala	1
Trp191Gly	fabG1-15C>T	-	17
Trp191Gly	fabG1-8T>C	-	2
Trp191Gly	inhA-154G>A	katG-Thr625Ala	2
Trp191Gly	-	-	3
Trp90Arg	-	-	3
Tyr413Cys	fabG1-15C>T	-	2
Tyr413Cys	inhA-Ser94Ala; fabG1-15C>T	-	1
Tyr413Cys	-	katG-Trp438Gly	1
Tyr413Cys	-	-	2
Tyr98Cys	-	katG-Leu378Met	1
Tyr98Cys	-	-	10

Table S5

Proportions of co-occurring known resistance and compensatory mutations, distance from heme binding sit and predicted stability change for the known *katG* resistance mutations. Co-occurring known resistance mutations are in *fabG1*, *inhA*, and *kasA* genes.

Change	Freq	Proportion co-occurring with a resistance mutation	Proportion co-occurring with a compensatory mutation	Distance from heme-binding site	Predicted Stability Change (ΔΔG)
Ser315Thr	7163	0.16	0.02	-0.306	11.693
Ser315Asn	167	0.10	0.05	-0.150	11.693
Ser315Gly	27	0.52	0.11	-0.558	11.693
Trp191Arg	25	0.56	0.28	-1.602	27.925
Ile335Val	25	0.04	0.00	-1.263	16.565
Ala110Val	14	0.86	0.00	-0.619	11.992
Asn138His	14	0.00	0.57	-1.532	11.675
Trp328Leu	14	0.00	0.00	-1.662	16.309
Thr380Ile	13	1.00	0.31	-0.265	9.721
Ser315Ile	13	0.00	0.00	-0.340	11.693
Met257Ile	12	0.83	0.00	-1.301	16.356
Tyr337Cys	12	0.17	0.17	-1.494	20.164
Ser140Asn	9	0.44	0.44	-0.384	12.600
Val1Ala	8	0.00	0.38	-	-
Ile317Val	8	0.00	0.00	-1.272	12.165
Gln127Pro	8	0.75	0.00	0.121	16.793
Tyr155Cys	7	0.29	0.57	-1.648	29.829
Ser315Arg	7	0.00	0.00	-0.165	11.693
Thr275Ala	7	0.00	0.00	-1.165	10.011
Gln461Pro	7	0.29	0.00	-0.153	57.763
Phe252Leu	7	1.00	0.00	-1.645	14.614
Ser302Arg	6	0.17	0.17	-0.453	22.561
Trp91Arg	6	0.17	0.33	-2.081	19.039
Tyr155Ser	6	0.67	0.33	-3.040	29.829
Leu141Phe	5	0.00	0.00	-1.583	13.825
Ala106Val	5	0.80	0.00	-0.433	10.162
Asn138Ser	5	0.00	0.20	-1.846	11.675
Gly297Val	5	0.00	0.40	-0.648	22.251
Ala109Val	4	0.00	0.00	-0.574	12.583
Asp419His	4	0.50	0.00	-1.161	21.844
Arg104Gln	4	0.00	0.25	-1.103	7.187

Gly279Asp	4	0.00	0.00	-1.342	19.236
Val473Phe	4	0.00	0.00	-1.529	47.209
Gly285Asp	4	0.75	0.00	-0.526	21.774
Asp735Ala	4	0.25	0.00	-0.663	34.481
Tyr413His	3	0.67	0.00	-2.472	16.985
Trp300Gly	3	0.00	0.00	-3.564	19.266
Trp198*	3	0.00	0.67	-	-
Gly234Arg	3	0.00	0.00	-0.883	14.333
Leu378Pro	3	0.00	1.00	-1.921	10.44
Trp204*	2	0.00	0.00	-	-
Trp668*	2	0.00	0.50	-	-
Arg249Cys	2	0.00	0.00	-1.457	15.705
Met126Ile	2	1.00	0.00	-0.859	18.132
Ala139Pro	2	0.00	0.00	-0.599	11.833
Glu588*	2	0.00	1.00	-	-
Trp412*	2	0.00	1.00	-	-
Ala172Val	2	0.50	1.00	-0.474	14.531
Ala424Gly	2	0.00	0.00	-0.676	30.436
Ala264Thr	2	0.00	0.00	-1.175	10.974
Trp505*	2	0.00	0.50	-	-
Ser700Pro	2	0.00	1.00	-0.074	48.17
Asn138Asp	2	0.50	0.00	-2.008	11.675
Gly299Cys	2	0.00	1.00	-0.808	19.291
Glu195Lys	2	0.00	0.00	-0.258	24.749
Met257Thr	2	0.00	1.00	-2.068	16.356
Thr85Pro	2	0.00	0.00	-0.315	22.097
Ala424Val	2	0.00	0.00	-0.169	30.436
Trp149*	2	0.00	1.00	-	-
Trp438*	2	0.00	0.50	-	-
Gln525Pro	2	0.50	0.00	0.101	52.663
Trp90*	2	0.00	0.50	-	-
Ala65Thr	1	0.00	0.00	-0.913	32.704
Asp357His	1	0.00	0.00	0.433	26.139
Trp321*	1	0.00	0.00	-	-
Thr324Pro	1	0.00	1.00	-0.467	14.819
Arg463Trp	1	1.00	0.00	-0.593	59.168
Trp328Ser	1	0.00	1.00	-3.368	16.309
Leu384Arg	1	0.00	0.00	-2.321	15.103
Thr308Pro	1	0.00	1.00	-0.769	19.476
Arg418*	1	0.00	0.00	-	-
Phe567Ser	1	1.00	0.00	-2.448	47.080

Ala550Asp	1	0.00	0.00	-2.005	53.786
Gln295Pro	1	0.00	0.00	0.021	25.611
Thr394Ala	1	0.00	0.00	-1.272	18.44
Tyr28*	1	0.00	0.00	-	-
Gln88*	1	0.00	0.00	-	-
Thr275Pro	1	0.00	1.00	-0.475	10.011
Asp311Gly	1	0.00	1.00	-0.970	18.332
Ala162Thr	1	1.00	0.00	-1.914	19.634
Ala379Val	1	0.00	1.00	-0.495	11.387
Trp328Cys	1	0.00	0.00	-2.115	16.309
Ser175*	1	0.00	0.00	-	-
Ser671*	1	0.00	0.00	-	-
Gly299Ala	1	0.00	1.00	-0.794	19.291
Ser17Asn	1	0.00	0.00	-	-
Ala574Val	1	0.00	0.00	-0.672	37.811
Asp695Ala	1	0.00	1.00	1.655	48.128
Gly307Arg	1	1.00	0.00	-0.395	20.350
Thr180Lys	1	0.00	0.00	-0.580	19.544
Trp477*	1	0.00	1.00	-	-
Leu587Pro	1	1.00	0.00	-1.021	30.651
Trp351*	1	0.00	1.00	-	-
Asp259Glu	1	0.00	0.00	-0.679	16.389
Gln36*	1	0.00	1.00	-	-
Thr326Met	1	1.00	0.00	0.099	15.246
Gln352*	1	0.00	1.00	-	-
Ser140Gly	1	0.00	0.00	-1.084	12.600
Trp300*	1	1.00	0.00	-	-

Table S6
Proportions of co-occurring compensatory mutations with *rpoB* rifampicin resistance mutations.

rpoB mutation	Frequency	Proportion co-occurring with a compensatory mutation
Ser450Leu	4970	0.245
Leu430Pro	140	0.036
Ser450Trp	130	0.008
Val170Phe	73	0.164
Gln432Pro	31	0.419
Gln432Lys	26	0.192
Gln432Leu	21	0.143
Pro454His	7	0.143

SUPPLEMENTARY FIGURES

Figure S1

Analysis strategy and numbers of falling into each category

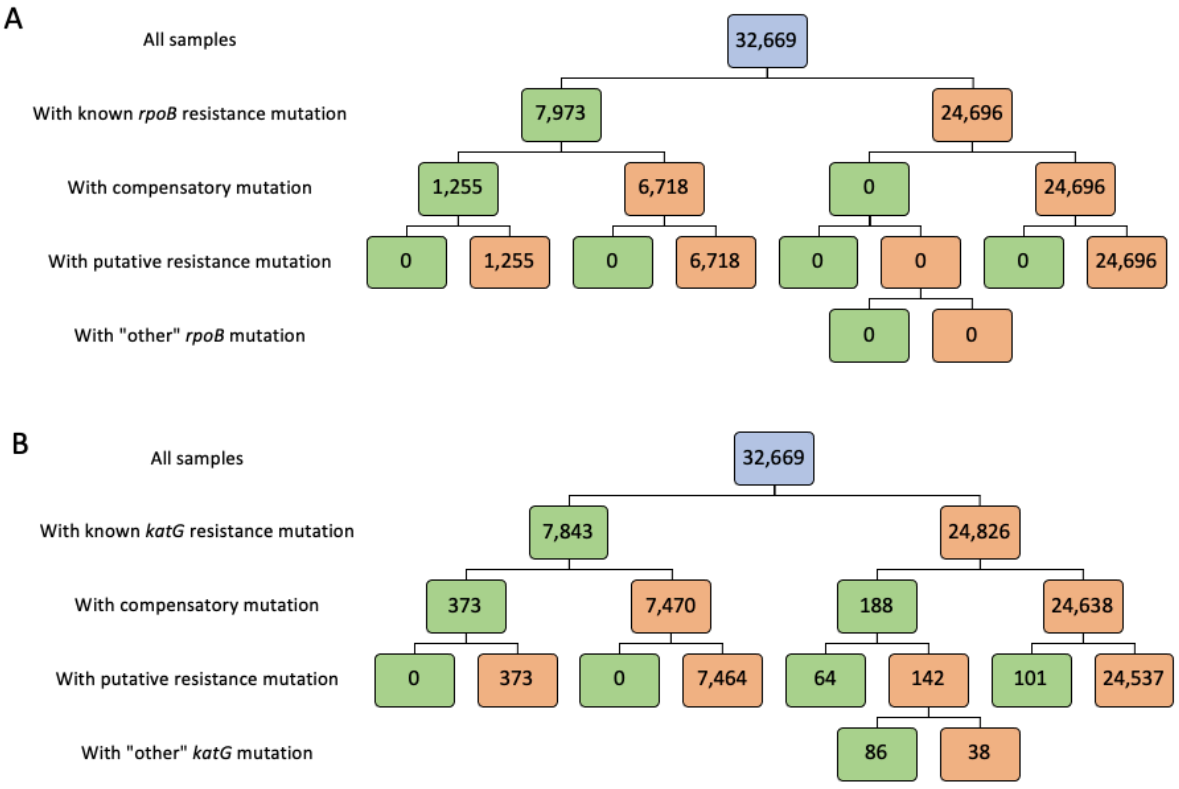
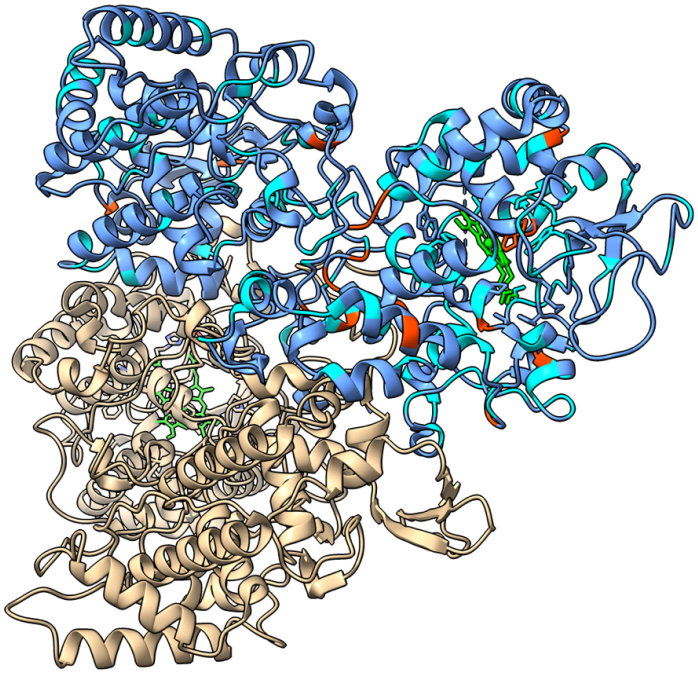


Figure S2

Protein structural model of KatG (7ag8) with known (cyan) and putative (red) mutations highlighted on chain A (blue). The vast majority of putative mutations cluster around the bound heme (green).



Chapter 6
Discussion and conclusion

Discussion

This thesis has focused on the three aspects of MTBC genomics (strain-typing, phylogenetic and transmission clustering, and drug resistance profiling), with insights that could improve diagnostics, clinical decision-making, and epidemiological applications, including tracking of outbreaks.

The work in **Chapter 2** established an updated SNP barcode for the rapid and accurate determination of an *Mtb* isolate's sub-lineage. Detection of these SNPs in individual samples can inform on strain-types, add valued data to *Mtb* databases and inform associated complementary software tools, including TB-Profiler [1]. The high accuracy of the new barcode was demonstrated with 'training' and 'test' sets. By introducing 30 new (sub-)lineages there will be greater resolution of strain-types within the MTBC phylogeny. With increasing whole genome sequencing of MTBC isolates, the number of sub-lineages is likely to increase, and the same core methods (e.g., F_{ST} analysis, training and test validation framework) used will increase and improve the resolution of the phylogeny, leading to a modified barcode. Indeed, the present improvement on the old barcode (65 SNPs) resulted from a 22-fold increase in samples (1,601 in [2], to 35,298 analysed here [3]). A strength of the barcode is that SNPs were chosen that led to synonymous changes, avoiding repetitive, highly changeable regions (i.e., non-PE/PPE regions), and not in drug resistance target genes, thus are under minimal selective pressure to undergo change. Moreover, MTBC does not undergo horizontal gene transfer with a low mutation rate (0.04–2.2 SNPs per-genome-per-year, notwithstanding lineage differences [4]), so the barcoding markers should be reasonably stable.

The new sub-lineage identification relied on statistical analysis of ratios of clade branch lengths to those of their descendants to recapitulate the phylogenetic patterns of pre-existing (sub-

)lineages. Future barcoding projects could establish novel (sub-)lineages in more principled ways. One way could be to associate sub-lineages more closely with phenotype or metadata, for example, reflecting isolates or strain-types that are more predisposed to being drug-resistant or exhibit increased virulence. In some cases, a (sub-)lineage prediction could be a proxy for phenotypic prediction, hence further informing clinical and epidemiological decisions. This approach would be challenging because, although phenotypic lineage differences have been somewhat well-established, these are at the macro-scale of the main lineages rather than the fine-grained level of sub-lineages. For example, phenotypic differences between lineages 2 and 4 and the other main lineages (e.g., ancient) have been found [5]. However, how these phenotypic-genotypic differences map to disease outcome is complicated by the interactions between host, pathogen, and environment [6].

Another enhancement to the adopted approach could be to apply statistical clustering methods to the phylogenetic topology, building on the presented branch length analysis. Such clustering algorithms could assist in identifying distinct sub-lineages within the larger tree. For example, the software package fastbaps ("Fast hierarchical Bayesian analysis of population structure") seeks to identify clusters within a larger population (i.e., trees) using a model-based Bayesian approach [7]. While helpful, it seems unlikely that an unsupervised algorithm would find all robust clusters coinciding with either those intuitively identified from the tree or pre-established clades based on biological information (e.g., strain-types). Indeed, this was the case in early stages when identifying new sub-lineages. The automated process could not satisfactorily match established sub-lineages nor identify clear new ones, leading to the semi-automated approach with statistical backing. Given the fractal nature of any tree, whether a cluster should be considered a *bona fide* sub-lineage is an important issue, but the value of these clades is in being able to identify their unique SNPs for epidemiological applications. This value is amplified for

transmission clusters. My aim was to produce a barcode that has a high phylogenetic resolution to rapidly position an isolate on the MTBC tree without having to reconstruct the tree, which is computationally expensive for many isolates. Further work could develop an informatic platform to assist this task, and identify closely related isolates, which would have benefits for tracking transmission outbreaks.

Taking advantage of the high-resolution strain-typing system developed in **Chapter 2**, the work in **Chapter 3** sought to compare it to the older spoligotyping system. There are many more spoligotypes profiled in the literature (~3k), and even after filtering out spoligotypes appearing in fewer than five samples, four times as many spoligotypes remained (~400) compared to characterised (sub-)lineages (~100). This system would therefore seem to provide even higher resolution than that of **Chapter 2**. However, the results revealed that much of this variation can be attributed to noise. As a first approximation, the (supplementary) figures of phylogenetic trees in **Chapter 3** clearly reveal large variation in individual spoligotypes, despite some overall patterning in the major lineages (1-7; lineage 2 is particularly clear in its overall pattern). In much the same manner as the barcoding SNPs in **Chapter 2**, a score was developed in which 1 indicated a spoligotype was exclusive to a lineage.

The proportion of spoligotypes exclusive to lineages greatly decreased at each lower lineage level, and even among those spoligotypes with a score of 1, they frequently occurred in a low proportion of that lineage's sample (spoligotype 11011111111111111110011111110000101111111111 (SIT 19) for example appeared exclusively in lineage 1 but only in ~17% of its samples). Only at the top lineage level did the spoligotypes show overall good predictive power, where over 96% of spoligotypes were exclusive to a main lineage (1-7). Overall, as expected, it seems that the utility of spoligotypes is inferior to the

strain-typing system of **Chapter 2**, which offers greater resolution and less noise. This is perhaps not surprising given that spoligotypes profile an unstable repeat region which is highly variable [8]. In contrast, the updated lineage system deliberately uses synonymous SNPs in stable regions, which are thereby not subject to natural selection and less likely affected by stochastic processes. One advantage spoligotypes could have over the lineage/barcode system is that their nomenclature reflects historical phylogeographical distribution ('spoligotype family'). For example, the 'Beijing' clade and subclades reflect this strain having originated in East Asia. The samples were collected on a global scale and the place of collection does not necessarily correspond with the phylogeographical family name, so a Beijing strain for example could have been collected anywhere in the world at a certain point in time. This is useful in tracking the historical spread of MTBC lineages on large scales and provides potential new information as to the differential spread of families, indicating possible important genotype-phenotype differences such as virulence or drug resistance. There is evidence suggesting such associations [9][10], and having two strain-typing systems working together should improve epidemiological insights. As the lineage system of **Chapter 2** is updated with more samples and more SNPs, providing an increasingly accurate barcode, the correspondence of the lineages to spoligotypes may improve.

Work in **Chapter 4** offered insights into *Mtb* transmission and drug resistance in Pakistan, a country with a high TB burden. Transmission clades were identified through isolates that are genetically closely related, using an arguably *ad hoc* threshold of <10 SNPs difference. Using a GWAS approach of transmitted and non-transmitted isolates, the *nusG* gene (coding for transcription factor NusG) was most closely associated with these transmission samples. While it is possible to assign a fixed SNP cut-off to establish transmission, typically estimated by considering the identity by state distribution within a study, more advanced approaches using

transmission models and phylogenetic trees are gaining traction [11]. Stimson (2019) [12] uses the SNP differences along with data indicating how long those differences have taken to accumulate. This temporal dimension is inferred from previous evidence about case timing, the *Mtb* molecular clock, and transmission processes, and implemented in *Transcluster* R package. Another alternative approach [13] (implemented in *TransPhylo*, another R package) and uses a phylogenetic tree, assumes not all transmission events may have been captured, and infers missing isolates that were not sampled.

Irrespective of approach, there are complications in determining TB transmission because *Mtb* has low variation, the molecular clock is unstable, and therefore transmission distances are variable, as are the latent and infectious periods. There are also different transmission bottlenecks, with variation in the number and diversity of bacteria transmitted. In short, these complications distort transmission inference if just looking at SNP distances and a phylogenetic tree. This variability is encoded as distributions of parameters in the TransPhylo model, such as 'Generation time distribution', 'Sample time distribution', 'Sampling density', and others, again processed using Bayesian methods. These methods have been applied in low burden with immigration [14] and high burden settings [11]. Whilst these approaches advance models of transmission beyond the simple SNP threshold and are less *ad hoc*, they require substantial metadata in order to satisfy the model parameters, and without these metadata spurious inferences are likely. For example TransPhylo requires a high sampling density - a requirement that is unlikely in poor, high burden countries such as Pakistan. Often, required metadata were simply missing or not sufficient in the Pakistan study. Such sophisticated transmission models are perhaps best applied when incorporated into study design - i.e. suitable transmission data ought to be collected on the basis that these models will will be applied prospectively, rather

than in a retrospective way. In lieu of the appropriate data for these models, a SNP threshold is likely adequate to account for many transmission events.

Nevertheless, because there are gaps in the data such as information about sampling density, dividing the samples into 'transmitted' and 'non-transmitted' as phenotypes for the purpose of association analysis using GWAS could be questioned. In other words, there could be transitive samples linking those between 'non-transmission' and 'transmission' clusters, thus bringing some samples in to 'transmission' status. The misclassification of non-transmitted strains that are being transmitted is likely to reduce the statistical power of the analysis. In a perfect dataset capturing every transmission event however, there will be some variation in SNP distances, and clusters of smaller SNP distances can be said to be those 'transmitted', representing a proxy for transmissibility. Samples having greater transmissibility will be transmitted in less time and hence have smaller SNP distances between them when looked at cross-sectionally. That the study is incompletely sampled is a weakness, but capturing samples with small SNP distances ought to reflect more transmissible samples.

The biology underlying the association between the transcription factor *nusG* and transmissibility needs to be elucidiated. Transcription factors have complex relationships between genotype and phenotype given that they code for proteins that regulate the expression of other proteins and are in turn influenced by other genes and their possible transcription factors. However, a future investigation could combine the outlined modelling approaches above with GWAS to build a more accurate genotype-transmissibility picture. Should the same gene appear to still be closely associated then further investigations into its transmissibility role would be justified. Further, it may be possible to look at the effects on *in vitro* phenotypes (e.g., growth) after CRISPR-mediated gene editing of the *nusG* locus.

The drug resistance analysis revealed unknown mutations in several anti-TB drug gene targets. There was evidence to suggest their role in resistance, including functional consequences, conversion, and presence in the wider global dataset (n=32k). Also, many were present globally, but were absent in sensitive samples. Although, as with the transmission analysis, the dataset used is not comprehensive, it is nevertheless enriched with drug resistance isolates, thus providing opportunities to reveal important new mutations. These potential resistance mutations require further investigation with experimental work and phenotypic testing to confirm their causative role in resistance. In high burden TB countries such as Pakistan, drug resistance can be a result of drug misuse or counterfeit drugs [15], and detecting new resistance mutations is valuable in tracking the effects of these harmful practices and importantly, improving clinical decision-making.

Within the Pakistan dataset, several novel mutations were found in in genes that play a compensatory role, together with known compensatory mutations. Interestingly, several isolates were found with a co-ocurrence of a compensatory mutation without having a known resistance mutation. Upon further investigation many of these isolates were found to have one of the novel potential resistance mutations identified in this study. Therefore, in **Chapter 5**, a comprehensive search for isolates with the same pattern was conducted in the global dataset (n=32k) to find new resistance mutations. The two drugs were selected for this analysis were isoniazid (INH) and rifampicin (RIF), as compensatory-resistance mutations and mechanisms are known. For INH, the mutations in the *ahpC* gene can be compensate for the fitness cost of mutations in resistance linked *katG*. For RIF, compensatory mutations occur in *rpoA* and *rpoC*, and linked to resistance mutations in *rpoB*, particularly those in the RRDR [16].

The strategy was to curate a set of known compensatory mutations, then identify isolates with these mutations, but no known resistance mutations in linked target resistance genes. For RIF, the novel putative mutations in *rpoB* occurred in one or two samples ('rare mutations'), and closer inspection suggests they are most likely to be lineage-specific. The lack of putative resistance mutations, or even robust 'rare' mutations in *rpoB* is not surprising, as the RNA polymerase is sensitive to small changes given its complexity and essential biological role, thereby limiting potential mutations to the RRDR. The lack of novel mutations in *rpoB* demonstrates that this method does not output very many false positive associations and can be seen as a negative control for the method.

For INH, there were 31 novel mutations in *katG* which were present in at least three isolates. Many rare mutations were found in *katG* and can be considered good potential candidates for *bona fide* INH resistance mutations, as they were only excluded due to sample size and not phylogenetic reasons. While it has been shown that the putative *katG* mutations may have effects on KatG protein, and are proximal to the INH heme binding site, they should be followed up with laboratory functional work (e.g., *in vitro* CRISPR manipulation of *katG* with phenotypic DST). More generally, if the mutations are spurious, then basing clinical decisions such as drug regimens on false positives could result in mismanagement and indeed further drug resistance.

In curating a list of potential compensatory mutations or potential resistance mutations, the criterion for inclusion was the presence of an association with genotypic and phenotypic resistance to other drugs. Resistance to other drugs is relevant in detecting potential compensatory or resistance mutations in a specific drug. INH and RIF resistance are closely associated, with 78% of reported RIF-resistant cases in 2019 also being resistant to INH [17]. Therefore, classifying sequenced samples as genotypically or phenotypically MDR-TB is

informative for establishing the compensatory or resistance mechanisms of mutations. Further, the approach could be enhanced by identifying potential compensatory mutations thorough statistical association methods, such as Direct Coupling Analysis (DCA), which aims to measure the strength of direct relationships between mutations by excluding effects from other loci. Future work could apply this approach to unknown *ahpC* mutations occurring with known resistance mutations to find more accurate cause-and-effect compensatory-resistance relationships. Similarly, such mutations could be tested experimentally in the laboratory using *in vitro* methods, as mentioned earlier.

Although the study is limited in its ability to declare the list of *katG* mutations definitively causing drug resistance, it demonstrates that leveraging the presence of compensatory mutations to detect novel resistance mutations is useful. This approach offers a valuable shortcut in tracking and determining drug resistant samples which would otherwise have been false negatives. Widening the scope of application, the methods could be applied to any combination of genes, in any pathogenic organism, which have a similar compensatory-resistance dynamic. Non-essential targets such as *katG* can exhibit multiple resistance mutations without loss of function, and so would be especially suited to the approach presented.

Unfortunately in *Mtb*, other known compensatory-resistance mechanisms seem to act upon essential structures and so are limited in their ability to mutate, much like the *rpoB* case. Translational accuracy of the ribosome, which is impacted by streptomycin (STM) resistance mutations, is restored in compensatory mechanisms for STM resistance [18]. Purported compensatory mutations have been found to interact with structures intolerant to change - DNA gyrase (*gyrA* gene) for flouroquinolones, and the ribosome (*rrs* gene) for other aminoglycasides such as capreomycin [19].

This thesis has demonstrated the value of applying pipelines to large numbers of WGS samples to answer a variety of clinically and epidemiologically relevant questions. There is concern however that the benefits of these technologies will not be sufficiently streamlined or cost-effective to be available in parts of the world most effected by TB, and so its real-time clinical utility is low [20]. To overcome this, sequencing is moving beyond standard pipelines and analyses, with advances in portable and cost-effective technologies, as well as the economic use of existing ones through the sequencing of targeted amplicons across many samples and loci.

One persistent problem in scaling sequencing analyses has been the necessity for culture, requiring specialist labs and necessitating a timescale of weeks before bioinformatic analysis and results are possible. Attempts have therefore been made to sequence directly from samples using either innovative new ways of doing WGS (e.g. genome enrichment), or sequencing just the parts of genome relevant to, say, drug resistance (e.g., amplicon sequencing). Even though these innovations have advanced sequencing towards being useful on much shorter time scales and for more personalised treatment, they still have trade-offs in terms of real-world application.

WGS directly from sample has been successfully achieved by enrichment of key parts of the *Mtb* genome, using oligonucleotide enrichment technology SureSelectXT [21]. While dramatically reducing time to results from weeks to days, and with high quality data, the method is prohibitively expensive for underfunded laboratories and has specialist technical requirements. A lower cost, even quicker WGS study direct from a collected sputum sample was achieved with a low-cost DNA extraction method [22] which saw same-day results. This study also made use of the Oxford Nanopore MinION technology, an inexpensive and portable sequencing platform. While the MinION technology has a high error rate, this is offset by continuation of sequencing

until high coverage is achieved. Furthermore, its very long reads can overcome difficulty mapping repeat regions to the reference genome [23]. Although fast, cheap and accurate results were achieved, the direct-from-sample method is complicated due to DNA from other bacterial and human cells being present in the sample. This gap is however being addressed, with bioinformatics pipelines assisting this crucial filtering step [24], well as adaption of the technology to read and sequence only DNA from the genome of interest. The drawback to MinION is its small scale, where a maximum of 12 samples can be sequenced in a single run [20], but this may be cost-effective for clinical applications in drug resistance settings and could be increased in the future.

Amplicon sequencing can also bypass culturing by amplifying relevant genetic loci directly from samples, making it more economical than WGS. Sequencing results can be obtained quickly and accurately by detecting known resistance SNPs in multiple genes simultaneously, and because only relevant genes are selectively amplified, the amount of required DNA is reduced, as is the interference of unrelated DNA sequences [25].

These promising avenues are varied in their methods, but all are converging on lower cost and speedier results. As technology becomes more rapid and cheaper, it ought to become more available to the poorest countries and those with the highest-burden of TB. This will hopefully enable rapid diagnosis, detailed epidemiological studies, and accurate drug resistance profiling on a global scale. Although real-time and cheaper results are valuable in the clinical and epidemiological settings, the technological innovations are only useful to the extent that analysis can take place in the wider context of *Mtb* research. For example, a rapid test determining drug resistant SNPs is only likely to be meaningful if they are well-established as such by prior research, which includes large cohorts of WGS data. Examples of such

contributions in this thesis include the comprehensive barcode of **Chapter 2**, which would not have been possible without a large high-quality global WGS dataset. Similarly, detection of novel resistance mutations in **Chapter 5** required a large-scale sweep of the global dataset; these mutations in time will become part of the database of known resistance mutations which will be the target of investigation in any clinical sequencing analysis.

It seems likely therefore that high-quality WGS from culture in laboratories will continue to be of value to *Mtb* research, and complement those data generated using innovative rapid-cheap technologies in clinics and field settings. Such molecular technological improvements combined with those in bioinformatic approaches and information systems are likely to lead to reductions in the burden of TB disease, including through more timely informed interventions before transmission, greater personalization of drug regimens, and discovery of insights into TB biology for drug and vaccine development.

Conclusion

This thesis has contributed to the development of three central kinds of analyses at the heart of *Mtb* research, namely strain-typing, phylogenetic clustering or transmission inference, and drug resistance profiling, through refinement of pre-existing frameworks and the addition of new data or with novel empirical findings. **Chapter 2** improved upon an existing lineage system with a greatly expanded dataset, introducing new MTBC lineages and strengthening the identification of old ones. Building on this work, **Chapter 3** clarified the relationship between this lineage system and the well-established and widely used strain-typing system of spoligotypes by assessing their ability to predict lineage. This chapter concluded that the older spoligotyping system is noisy and predicts poorly at the lower lineage levels. Assessing drug resistance and transmission in Pakistan (**Chapter 4**), a phylogenetic approach was used to uncover potentially

important associations between a single gene and transmission, as well as potential novel drug resistance mutations. A comprehensive search for the presence of potential novel resistance markers combined with the presence of known compensatory mutations (**Chapter 5**), revealed new and putative resistance mutations in *katG*, the activator of the pro-drug isoniazid.

Set in an environment where advances in sequencing technologies are leading to more timely and affordable data generation, this thesis covers the application of analytical approaches suitable for clinical and epidemiological domains. Such insights from this and other studies are needed to assist efforts to reduce the high burden of TB.

References

1. Phelan JE, O'Sullivan DM, Machado D, Ramos J, Oppong YE, Campino S, et al. Integrating informatics tools and portable sequencing technology for rapid detection of resistance to anti-tuberculous drugs. Genome Medicine. 2019;11:41.

2. Coll F, Preston M, Guerra-Assunção JA, Hill-Cawthorn G, Harris D, Perdigão J, et al. PolyTB: a genomic variation map for Mycobacterium tuberculosis. Tuberculosis (Edinburgh, Scotland). 2014;94:346–54.

3. Napier G, Campino S, Merid Y, Abebe M, Woldeamanuel Y, Aseffa A, et al. Robust barcoding and identification of Mycobacterium tuberculosis lineages for epidemiological and clinical studies. Genome Medicine. 2020;12:114.

4. Menardo F, Duchêne S, Brites D, Gagneux S. The molecular clock of Mycobacterium tuberculosis. PLOS Pathogens. 2019;15:e1008067.

5. Coscolla M, Gagneux S. Consequences of genomic diversity in mycobacterium tuberculosis. 2014;26:431–44.

6. Coscolla M, Gagneux S. Does M. tuberculosis genomic diversity explain disease diversity? 2010;7:e43.

7. Tonkin-Hill G, Lees JA, Bentley SD, Frost SD, Corander J. Fast hierarchical Bayesian analysis of population structure. Nucleic Acids Research. 2019;47:5539.

8. 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. Journal of Clinical Microbiology. 1997;35:907.

9. Forrellad MA, Klepp LI, Gioffré A, García JS, Morbidoni HR, de la Paz Santangelo M, et al. Virulence factors of the mycobacterium tuberculosis complex. Virulence. 2013;4:3–66.

10. Ribeiro SC, Gomes LL, Amaral EP, Andrade MR, Almeida FM, Rezende AL, et al. Mycobacterium tuberculosis strains of the modern sublineage of the beijing family are more likely to display increased virulence than strains of the ancient sublineage. Journal of Clinical Microbiology. 2014;52:2615–24.

11. Sobkowiak B, Banda L, Mzembe T, Crampin AC, Glynn JR, Clark TG. Bayesian reconstruction of mycobacterium tuberculosis transmission networks in a high incidence area over two decades in Malawi reveals associated risk factors and genomic variants. Microbial Genomics. 2020;6:e000361.

12. Stimson J, Gardy J, Mathema B, Crudu V, Cohen T, Colijn C. Beyond the SNP Threshold: Identifying Outbreak Clusters Using Inferred Transmissions. Molecular Biology and Evolution. 2019;36:587–603.

13. Didelot X, Fraser C, Gardy J, Colijn C, Malik H. Genomic infectious disease epidemiology in partially sampled and ongoing outbreaks. Molecular Biology and Evolution. 2017;34:997–1007.

14. Ayabina D, Ronning JO, Alfsnes K, Debech N, Brynildsrud OB, Arnesen T, et al. Genome-based transmission modelling separates imported tuberculosis from recent transmission within an immigrant population. Microbial genomics. 2018;4.

15. Bate R, Jensen P, Hess K, Mooney L, Milligan J. Substandard and falsified anti-tuberculosis drugs: A preliminary field analysis. International Journal of Tuberculosis and Lung Disease. 2013;17:308–11.

16. Zaw MT, Emran NA, Lin Z. Mutations inside rifampicin-resistance determining region of rpoB gene associated with rifampicin-resistance in Mycobacterium tuberculosis. Journal of Infection and Public Health. 2018;11:605–10.

17. W.H.O. Global Tuberculosis Report 2020. W.H.O.; 2020.

18. Björkman J, Hughes D, Andersson DI. Virulence of antibiotic-resistant Salmonella typhimurium. Proceedings of the National Academy of Sciences of the United States of America. 1998;95:3949.

19. Alame Emane AK, Guo X, Takiff HE, Liu S. Drug resistance, fitness and compensatory mutations in Mycobacterium tuberculosis. Tuberculosis. 2021;129:102091.

20. Lee RS, Pai M. Real-Time Sequencing of Mycobacterium tuberculosis: Are We There Yet? Journal of Clinical Microbiology. 2017;55:1249.

21. Brown AC, Bryant JM, Einer-Jensen K, Holdstock J, Houniet DT, Chan JZ, et al. Rapid Whole-Genome Sequencing of Mycobacterium tuberculosis Isolates Directly from Clinical Samples. Journal of Clinical Microbiology. 2015;53:2230.

22. Votintseva AA, Bradley P, Pankhurst L, Del Ojo Elias C, Loose M, Nilgiriwala K, et al. Same-Day Diagnostic and Surveillance Data for Tuberculosis via Whole-Genome Sequencing of Direct Respiratory Samples. Journal of Clinical Microbiology. 2017;55:1285.

23. Gómez-González PJ, Campino S, Phelan JE, Clark TG. Portable sequencing of Mycobacterium tuberculosis for clinical and epidemiological applications. Briefings in Bioinformatics. 2022. https://doi.org/10.1093/BIB/BBAC256.

24. Cuevas-Córdoba B, Fresno C, Haase-Hernández JI, Barbosa-Amezcua M, Mata-Rocha M, Muñoz-Torrico M, et al. A bioinformatics pipeline for Mycobacterium tuberculosis sequencing that cleans contaminant reads from sputum samples. PLoS ONE. 2021;16.

25. Jouet A, Gaudin C, Badalato N, Allix-Béguec C, Duthoy S, Ferré A, et al. Deep amplicon sequencing for culture-free prediction of susceptibility or resistance to 13 anti-tuberculous drugs. European Respiratory Journal. 2021;57.