

LONDON
SCHOOL of
HYGIENE
& TROPICAL
MEDICINE



**Analysis of *Mycobacterium tuberculosis* 'omics data to inform
on loci linked to drug resistance, pathogenicity and virulence**

Paula Josefina Gómez González

Thesis submitted in accordance with the requirements for the degree of

Doctor of Philosophy

University of London

JANUARY 2022

Department of Infection Biology

Faculty of Infectious and Tropical Diseases

LONDON SCHOOL OF HYGIENE & TROPICAL MEDICINE

Funded by MRC

Research group affiliations: Taane G. Clark & Philip Butcher

I, Paula Josefina Gómez González, 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 the thesis.

Signed  _ Date 31/01/2022

ABSTRACT

Mycobacterium tuberculosis (*Mtb*) is the causative agent of human tuberculosis (TB) which remains one of the deadliest pathogens worldwide. The observed genetic diversity among *Mtb* lineages has been associated with differences in virulence, pathogenicity and drug resistance. However, a better understanding of *Mtb* strain diversity and its implications for *Mtb* biology will inform the development of TB control tools, including diagnostics, drugs, and vaccines. Through the application of 'omics approaches, this thesis presents a comprehensive analysis of whole-genome sequence (WGS) data from *Mtb* clinical isolates to improve the understanding of the pathogen biology and inform on pathogenicity and drug resistance. The integrated analysis of the genome, transcriptome and methylome of ancient and modern lineages of *Mtb* revealed genetic variants and methylation patterns with a potential role in gene expression regulation. Through the analysis of the frequency and distribution of mutations associated with resistance to the new anti-TB drugs (bedaquiline, delamanid and pretomanid) in a large data set (~30k isolates), mutations pre-dating the introduction of these drugs with likely functional effects were observed. This result suggests possible intrinsic or cross-resistance, and potential threats to the effectiveness of MDR-TB treatments. Moreover, by using long-read sequence data, it was possible to characterise the genetic diversity of the 169 *pe/ppp* genes, which are loci traditionally removed from WGS analysis due to their repetitive GC-rich regions. Structural variants in *pe/ppp* genes with lineage-specific patterns were found. Finally, with sequencing technologies gaining traction as diagnostic tools, the use of the MinION portable and long-read platform was assessed. The results support its suitability for epidemiological applications and drug resistance detection, with the potential to characterise *pe/ppp* genes through improved coverage of GC-rich regions. Overall, this thesis demonstrates the potential of sequencing platforms to inform TB control and improve the understanding of *Mtb* biology.

The application of different 'omics provides with a comprehensive analysis of the different *Mtb* lineages showing distinct genomic and transcriptomic profiles that translate into different behaviours, with diagnostic and treatment implications.

ACKNOWLEDGEMENTS

I would like to thank to my supervisors Taane Clark, Jody Phelan and Philip Butcher for all the help, support and guidance received, and the MRC for the funding and for giving me the opportunity to carry out this PhD. I would also like to thank the MRC team, for being always so helpful. And to every member of the Clark & Campino group, including the ones that already left - Ernest, Yaa, Ben, Matt R and Neneh - and everyone that is still around - Dan, Matt H, Gary, Anna, Emilia, Monica, Anton, Ashley, Emma, Julian, Leen, Sophie, Holly, Susana, and especially Amy, for always listening to me. Many thanks to Martin Hibberd, for all the advice received. I have to thank too to everybody that has been involved in my CL3 training and have helped me whenever I needed it in the lab, Teresa, Felipe, Anna, Beth and Andrea. Big thanks to Fernanda, for always budding me, Archie, for being my mentor, and both, for being my friends. During my time at LSHTM I have had the opportunity to travel to amazing places, and most importantly, meet wonderful people that have made these four years unforgettable. Not only I have learnt a lot from my colleagues, but also I have made very good friends, that have managed to make the hardest times less hard. Thanks to all my LSHTM friends that have always supported me, many of them mentioned above, and many others like Rhodri, Becky, Tom, Bronner, Harry and Paco, for making the time at LSHTM worth it.

To everybody that has supported me, helped me, and have made me enjoy this time in London, without whom it would not have been the same. Thanks to my SOP friends Nicole, Vipul, Sahar, Satinder, Jonny, Francisco, Nicola and many others, always ready to have fun; Marina, Dani and Arturo, my climbing and paella crew; Marta Mojarrieta, Marjorie and Edu, for being the best flatmates I could have; Daria and my basketball teammates; and especially,

to those that have been here since the very beginning, Marta, Alvaro, Miguel and Francesca, that made coming back to London extremely easy and, without whom, this time would not have been the same.

To my parents, my sister and my friends in Spain, for being always there. To the TB crew, for making working during the lockdown more bearable, and especially Jody, for his constant help.

I would happily repeat the experience if I could.

“TB, on the other hand, is not a drama queen. It kills silently and slowly.

Nevertheless, it’s an extremely effective killer.”

Dr. Aaron Motsoaledi

Table of contents

Abbreviations and Acronyms	8
1 Introduction	11
1.1 Global burden of tuberculosis disease	12
1.1.1 Tuberculosis and COVID-19	12
1.2 Disease aetiology, risk factors and host susceptibility	13
1.3 Diagnosis	16
1.4 Treatment and vaccines	18
1.5 Drug resistance	20
1.6 <i>Mycobacterium tuberculosis</i>	22
1.6.1 <i>Mycobacterium tuberculosis</i> complex and strain diversity	22
1.6.2 Genomic diversity	23
1.6.3 Transcriptomics	26
1.6.4 Epigenetics: DNA methylation	27
1.7 The <i>pe</i> and <i>ppe</i> genes	29
1.8 Whole-genome sequencing and ‘omics	32
1.8.1 Next-generation sequencing: short- and long-read sequencing technologies	32
1.8.2 Application in ‘omics	34
1.8.3 Analysis of NGS data	35
References	36

2	Objectives and Structure of the Thesis	56
2.1	Objectives	57
2.2	Structure of the Thesis	58
3	An integrated whole-genome analysis of <i>Mycobacterium tuberculosis</i> reveals insights into relationship between its genome, transcriptome and methylome . . .	61
4	Genetic diversity of candidate loci linked to <i>Mycobacterium tuberculosis</i> resistance to bedaquiline, delamanid and pretomanid	95
5	Functional genetic variation in <i>pe/ppe</i> genes contributes to diversity in <i>Mycobacterium tuberculosis</i> lineages and potential interaction with the human host . . .	158
6	Portable sequencing of <i>Mycobacterium tuberculosis</i> for clinical and epidemiological applications	240
7	Discussion	287
7.1	General discussion	288
7.2	Conclusions	295
7.3	The future of TB ‘Omics	296
	References	299

Abbreviations and Acronyms

4mC	C ⁴ -methyl-cytosine
5mC	C ⁵ -methyl-cytosine
6mA	N ⁶ -methyl-adenine
AIDS	Acquired immunodeficiency syndrome
BAM	Binary alignment map
BCG	Bacilli Calmette-Guérin
BDQ	Bedaquiline
BWA	Burrow-Wheeler Aligner
CD4	Cluster of differentiation 4
CFZ	Clofazimine
COVID-19	Coronavirus Disease 2019
CRISPR	Clustered regularly interspaced short palindromic repeats
DLM	Delamanid
<i>dN</i>	Number of non-synonymous substitutions per non-synonymous site
DNA	Desoxyribonucleic acid
DR	Direct repeat
<i>dS</i>	Number of synonymous substitutions per synonymous site
DST	Drug susceptibility testing
EMB	Ethambutol
eQTL	Expression quantitative trait loci
ESAT-6	6 kDa early secretory antigenic target
GWAS	Genome-wide association studies
HGAP	Hierarchical genome assembly process
HIV	Human immunodeficiency virus
HLA	Human leukocyte antigen
HTSeq	High-throughput sequencing

IGRAs	Interferon- γ release assays
indels	Insertions and deletions
INH	Isoniazid
IPD	Inter-pulse duration
L1-9	Lineage 1-9
LAMP	Loop mediated isothermal amplification
LED	Light-emitting diode
LoF	Loss of function
LPAs	Line probe assays
LZD	Linezolid
MDR-TB	Multidrug-resistant TB
MIC	Minimum inhibitory concentration
MPTR	Major polymorphic tandem repeat
MTases	Methyltransferases
<i>Mtb</i>	<i>Mycobacterium tuberculosis</i>
MTBC	<i>Mycobacterium tuberculosis</i> complex
NAATs	Nucleic acid amplification tests
NGS	Next-generation sequencing
nsSNPs	non-synonymous SNPs
ONT	Oxford Nanopore Technology
PacBio	Pacific Biosciences
PAS	<i>p</i> -aminosalicylic acid
PCR	Polymerase chain reaction
PE	Proline-Glutamate
PGAP	Prokaryotic Genome Annotation Pipeline
PGRS	Polymorphic GC-rich repetitive sequence
PPE	Proline-Proline-Glutamate
PTM	Pretomanid
PZA	Pyrazinamide

RDs	Regions of difference
RIF	Rifampicin
RNA	Ribonucleic acid
RR-TB	Rifampicin-resistant TB
SBS	Sequencing by synthesis
SMRT	Single-molecule real time
SMS	Single-molecule sequencing
SNPs	Single nucleotide polymorphisms
sSNPs	synonymous SNPs
TB	Tuberculosis
TbD1	<i>Mtb</i> -specific deletion 1
TFTRs	TetR family of transcriptional regulators
Th1	T helper 1
Th2	T helper 2
TST	Tuberculin skin test
VCF	Variant call file
WGS	Whole-genome sequencing
WHO	World Health Organisation
XDR-TB	Extensively drug-resistant TB

CHAPTER 1

Introduction

1.1. Global burden of tuberculosis disease

Human tuberculosis (TB), caused by *Mycobacterium tuberculosis* (*Mtb*) bacteria, has been present throughout the history of humankind, and caused the most mortality of any pathogen. During the 20th century, due to the introduction of the Bacilli Calmette-Guérin (BCG) vaccine, antibiotic treatments and better public health policies, TB morbidity and mortality trends decreased. However, these rates increased again at the end of the century, in part as a result of the AIDS epidemic and the emergence of anti-TB drug resistance [1], including rifampicin- (RIF) (RR-TB), multidrug- (MDR-TB) and extensively drug- (XDR-TB) resistant *Mtb*.

Nowadays, TB remains a global health problem being one of the deadliest infectious diseases worldwide [2]. One third of the world's population is considered to be infected; however, only 10% of these infected individuals will eventually develop the active form of the disease [3]. A total incidence of 9.9 million people was estimated for 2020, with most TB cases being found in South East Asia (43%), Africa (25%) and the Western Pacific (18%) World Health Organisation (WHO) regions (**Figure 1**). The number of deaths in 2020 showed an increase from 2019, with 1.3 million deaths among HIV-negative and a further 214,000 deaths among HIV-positive people [2]. Moreover, the emergence of resistant strains to the current anti-TB drugs threatens efforts to control the disease, accounting for 132,222 and 25,681 MDR/RR-TB and pre-XDR/XDR-TB cases respectively [2].

1.1.1. Tuberculosis and COVID-19

Although incidence and mortality rates have been declining in recent years, the COVID-19 pandemic has dramatically affected access to TB diagnosis and treatment, and therefore

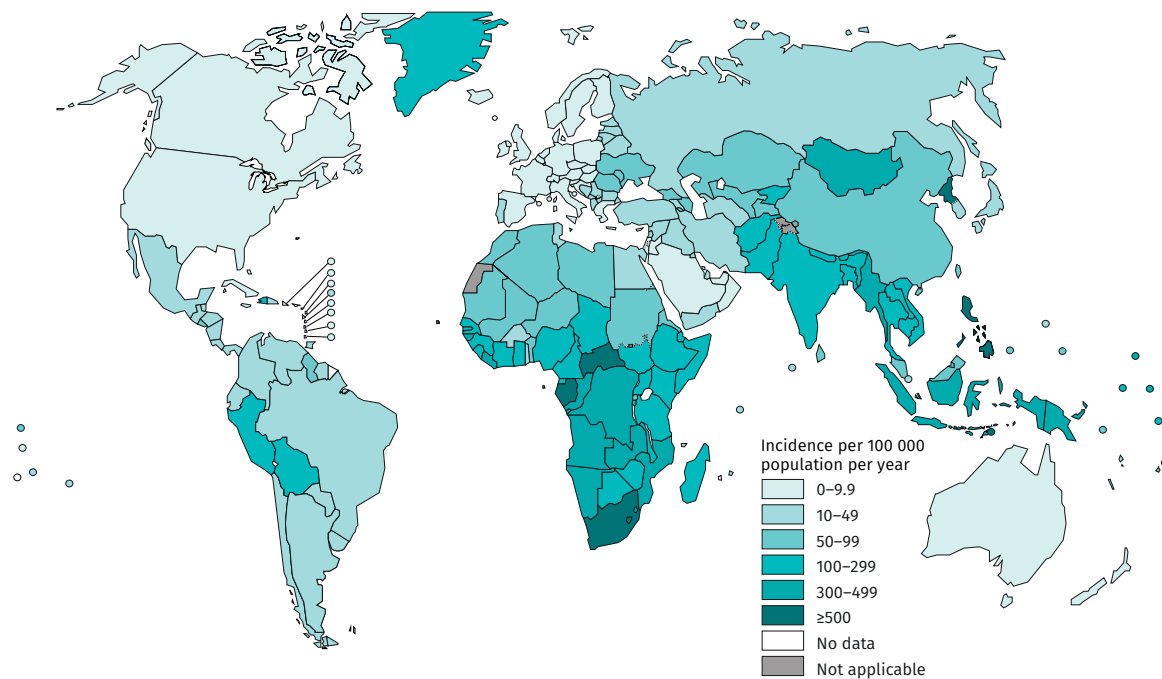


Figure 1. Estimated TB incidence rates per 100,000 population per year (taken from the WHO Global Tuberculosis Report 2021) [2].

substantially slowed down the progress achieved until 2020 in reducing the burden of the disease [2]. The number of TB cases notified in 2020 compared to 2019 has been reduced by 18% in average and up to 24% in high TB burden countries, showing a significant impact in case detection as a consequence of the COVID-19 pandemic [2, 4]. The acute reduction in detected cases points towards a reduced access to diagnosis rather than a result of decreased transmission. However, interventions such as lockdowns and mask-wearing may have had an impact in transmission whose extent is still unknown [4]. To mitigate some of the effects of the COVID-19 pandemic, the combined screening of COVID-19 and TB in high-burden settings has been suggested as a strategy to improve case-detection and reduce the potential risk of active TB associated with COVID-19 [4]. Modelling analyses have estimated an increase by 5-15% in TB mortality over the next 5 years [5]. Many of the TB-endemic countries have been the

most affected by COVID-19, with substantial economic impact that will probably translate into a long-term increasing trend in TB cases. Limited treatment support, reallocation of resources and restriction of movement have disrupted TB health services, especially in the most vulnerable settings [5]. Altogether, the COVID-19 pandemic has reversed gains in the fight against TB, which unfortunately, will force some TB-endemic countries to revise the 2025 milestones of the WHO End TB Strategy [2].

1.2. Disease aetiology, risk factors and host susceptibility

TB is an airborne infection transmitted by inhalation of aerosols containing viable bacilli from infected humans with active pulmonary disease. When the bacilli reach the host alveoli, they face the innate immune response mediated firstly by alveolar macrophages that phagocytose the bacteria (**Figure 2**). This first contact of the bacteria with the host is the beginning of a complex and yet not entirely understood interaction with the immune system. Although alveolar macrophages can eliminate the bacteria through the production of nitric oxide and reactive oxygen species, they play a dual role, also enabling the establishment of the bacilli [6]. *Mtb* bacteria becomes then resistant to clearance through different strategies of immune evasion, from inhibition of phago-lysosome fusion to dormancy [7,8]. Replication of the bacteria within the macrophage leads to cytolysis and infection of neighbouring cells [6]. In this early stage, lymphatic and haematogenous dissemination to other organs may occur [9]. The delay experienced in the initiation of the adaptive immune response (CD4 T Cells) enables the exponential growth and contributes to the survival of the bacilli [6]. Recruitment and confluence of lymphocytes, neutrophils and other immune cells at the primary site of infection forms the granuloma, and the consequent granulomatous inflammation that occurs in the periphery of

the lung constitutes what is known as “Ghon complex” [10]. Although granuloma formation has been associated with host protection, there is also growing evidence of its role in mycobacterial expansion [11]. These events represent the primary TB infection, usually asymptomatic, but sometimes the cause of non-specific symptoms, common with lower respiratory tract infections [10]. The Th1 cell-mediated immune response is believed to be principally responsible for containment of the initial infection, with the potential of *Mtb* elimination. However, immune evasion mechanisms by *Mtb* can trigger a gradual shift towards Th2 responses [12].

After primary infection, tubercle bacilli can remain in a dormant state for a long time, which is known as latent TB, the most common form of TB infection. The interior of the granuloma becomes necrotic and hypoxic, which triggers different metabolic adaptive pathways in the bacilli to enter a quiescent state. In this phase, a low proportion of the bacterial population, named “scouts”, become active and replicative, being constantly killed by the host immune response [9]. Interestingly, persistent bacteria has been found not only in the lung lesions, but also in different host locations, such as fat tissue [13]. When temporary or permanent immunological impairment occurs, tubercle bacilli replicate in an uncontrolled manner, so that the latent form of the disease shifts to active TB [9]. This post-primary TB infection can manifest as pulmonary (most common) or extrapulmonary, which includes tuberculosis meningitis or disseminated TB. Symptoms during this phase include fever, anorexia, reduced appetite, weight loss, night sweats, anaemia, persistent cough, sputum production and haemoptysis [14].

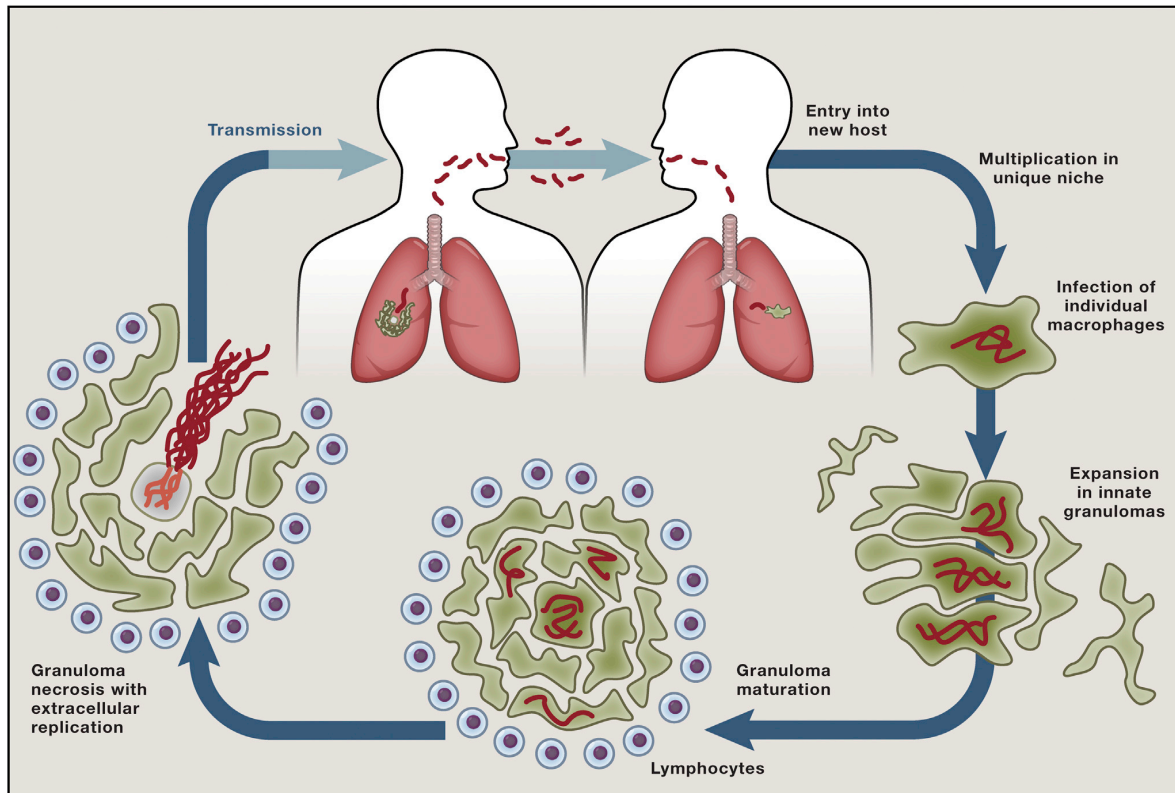


Figure 2. Transmission and granuloma formation during TB infection, taken from Cambier *et al.*, 2014 [15].

The main risk factors for the progression and development of the active form of the disease include HIV co-infection and drug-mediated immunosuppression. However, age, smoking, diabetes, malnutrition and other co-morbidities have also been reported to increase susceptibility to active TB [9, 16]. Additionally, socioeconomic levels, such as poverty and poor access to diagnosis and treatment are important aspects that can affect the population vulnerability [17]. Finally, several studies have demonstrated the impact that genetic factors have on resistance or susceptibility to TB. Twin and family studies have indicated evidence of heritable components of TB susceptibility [18]. Furthermore, polymorphisms in HLA genes related to ethnic and geographical differences have also been associated with increased susceptibility [19, 20],

as well as other variants identified in, for instance, genes involved in immune response and inflammation signalling pathways [21]. Along with host determinants, it is important to highlight pathogen factors, such as *Mtb* strain differences in virulence or drug resistance, that can influence the outcome of the infection, as a consequence of the dynamic host-pathogen interaction.

1.3. Diagnosis

A prompt and accurate diagnosis is crucial for the control of the TB disease. Different techniques are used for the diagnosis of the active and latent forms of TB. Although chest radiographies can be informative to identify pulmonary TB lesions, abnormalities in the lungs are often indicative of other pathologies [22], and therefore, bacteriological confirmation is required for the diagnosis of active TB. The mycobacterial culture is still the gold standard in many countries. This method can be performed in solid or liquid media, but due to the low replication rate of *Mtb*, it is highly time consuming (4-6 weeks in solid and 10-21 days in liquid media) [23]. Moreover, it requires trained personal and specific infrastructure. The sputum smear microscopy technique is also widely used. It is an inexpensive and simple method, although diagnostic quality is highly operator dependent. Its sensitivity is relatively low (~70%), increased by the application of LED fluorescent microscopy [24], and conditional on bacillary concentration, which limits its reliability in, for instance, HIV-positive patients [25]. For the early and correct diagnosis of active TB, the WHO currently recommends the use of endorsed molecular techniques as initial diagnostic tests [2]. The Xpert MTB/RIF, used worldwide, has shown good sensitivity and specificity. This test requires minimal processing and can be performed on sputum samples with the simultaneous detection of *Mtb* and resistance

to rifampicin [26]. Other nucleic acid amplification tests (NAATs) like line probe assays (LPAs), useful to detect resistant genotypes, have also been in use for a decade now [27]. Moreover, the recent development of NAATs has led to several other assays, like TB-LAMP [28] or True-nat MTB, which are also among the recommended tests in latest WHO guidelines [29]. The use of next-generation sequencing (NGS) for the detection of drug resistance significantly reduces the time of traditional phenotypic culture or culture-based testing [30,31]. Thus, several countries have already implemented NGS technology for surveillance of drug resistance [32]. Among the different approaches, target amplicon sequencing shows promising results and cost-effectiveness; however, only used for research purposes so far [32, 33]. Although *Mtb* culture is usually necessary prior to sequencing, the development of techniques performed directly from sputum have already been successful [34]. Finally, immunological tests are the methods of choice for the diagnosis of latent TB. There are two methods available: (i) the tuberculin skin test (TST) or Mantoux, and (ii) the interferon- γ release assays (IGRAs). One advantage of the latter over the TST is its improved specificity so it does not cause false positives after BCG vaccination [35]. However, due to its cost, the TST is still the preferred option in low-income regions [36]. With the present COVID-19 pandemic, fears of underdiagnosis of TB have grown. As they share common symptoms, suggestions on combined diagnostics for both infections have been proposed [32]. Additionally, research on host transcriptomic biomarkers of TB infection and progression poses an interesting field towards a pathogen-free diagnosis [37].

1.4. Treatment and vaccines

In view of its airborne transmission and the emergence of drug resistance, effective treatment is important for the management and control of tuberculosis. The anti-TB armamentar-

ium consists of first and second-line drugs (**Table 1**). First-line drugs include isoniazid (INH) and ethambutol (EMB) that target the synthesis of mycolic acids; pyrazinamide (PZA) that inhibits the synthesis of coenzyme A; and rifampicin (RIF) that inhibits RNA synthesis [38]. On the other hand, second-line drugs comprise different groups of drugs with various mechanisms of action, such as fluoroquinolones, injectable aminoglycosides, capreomycin class polypeptides, cycloserine and *p*-aminosalicylic acid (PAS) [38]. Recently, novel potent drugs like bedaquiline (BDQ), delamanid (DLM), linezolid (LZD) and pretomanid (PTM), have also been included for the treatment of drug resistant cases in different combination regimens [38, 39]. Second-line drugs are classified in three different classes (**Table 1**) based on their relative benefits and harms, and their use is reserved for the treatment of drug-resistant TB.

Current anti-TB treatment involves long regimens of a combination of bactericidal and sterilising drugs. This approach is based on the principle of a two-step treatment, with an initial bactericidal phase where replicative bacilli are killed, leading to clinical recovery, followed by a sterilising phase where semi-dormant bacilli are eliminated [40]. For susceptible cases, this regimen is composed of the combination of 4 first-line drugs for 6 months, consisting of 2 months with INH, PZA, EMB and RIF, followed by 4 months with RIF and INH [41]. The treatment of latent TB cases is recommended for high-risk patients (*e.g.*, HIV co-infection or household contacts of a bacteriologically confirmed TB case), where a 6-months monotherapy of INH is usually prescribed [42]. The emergence of drug resistance and consequent treatment failure requires the use of second-line drugs, which have a higher toxicity and side effects, and thus promote lower compliance. Moreover, in HIV-positive patients or those with other comorbidities, the management of the disease can be more complicated due to pharmacological interactions [40]. This is of concern as ultimately it can lead to poor treatment outcomes and

Table 1. Drugs used for TB treatment

First-line drugs	
	Isoniazid (INH); Ethambutol (EMB); Pyrazinamide (PZA); Rifampicin (RIF)
Second-line drugs	
Class A	Levofloxacin (LFX); Moxifloxacin (MFX); Bedaquiline (BDQ); Linezolid (LZD)
Class B	Ethambutol (EMB); Delamanid (DLM); Pyrazinamide (PZA); Imipenem-cilastatin (IPM-CLN); Meropenem (MPM); Amikacin (AMK); Streptomycin (STR); Ethionamide (ETO); Prothionamide (PTO); p-aminosalicylic (PAS)
Class C	Kanamycin; Capreomycin; Gatifloxacin; High-dose INH; Thioacetazone; Clavulanic acid

a higher risk of development of further drug resistance. The recommendation for multidrug-resistant TB (MDR-TB) cases is dependent on the resistance profile to the different anti-TB agents and the eligibility of each patient for the specific treatment, often requiring longer regimens of 18 months or more [39]. Although the introduction of the new drugs, such as BDQ or DLM, has brought promising results, efforts towards the discovery of novel compounds for the treatment of MDR-TB are still necessary.

Prevention of TB is based on the interruption of the transmission through early diagnosis and treatment of active TB. Although its effectiveness has been reported to be very variable [43], the BCG vaccine is the only one currently available and still widely used, as several studies support its protection against the most severe forms of childhood TB [16]. Ongoing efforts in the vaccine development pipeline are focused on different types, from attenuated or inactivated to subunit vaccine candidates, with more than a dozen of them undergoing clinical trials [2, 43]. An effective vaccine would be crucial in achieving the WHO goal of TB eradication by 2050 [44]. However, attempts to develop a more effective vaccine have been unsuccessful

so far.

1.5. Drug resistance

The emergence of drug resistance to the first and also second-line drugs is a public health concern that threatens the current therapeutic arsenal. TB drug resistance is classified into five categories as follows: (i) INH-resistant TB; (ii) RR-TB or rifampicin-resistant TB; (iii) MDR-TB or multidrug-resistant TB (resistant to RIF and INH); (iv) pre-XDR-TB or pre-extensively drug-resistant TB (additional resistance to any fluoroquinolone); and (v) XDR-TB or extensively drug-resistant TB (additional resistance to BDQ or LNZ) [2]. The detection of drug resistance requires culturing and further phenotypic drug susceptibility testing (DST), which can delay significantly the start of an adequate treatment. Nevertheless, in *Mtb*, drug resistance is mainly conferred by single nucleotide polymorphisms (SNPs) and small insertions and deletions (indels) located in genes coding for drug targets or enzymes responsible of activating prodrugs [45]. Moreover, acquisition and accumulation of resistance conferring mutations sometimes entails fitness loss, which triggers putative compensatory mechanisms [46, 47]. Through the comparative analysis of DST and genomics, mutations causing resistance have been characterised [48], and this has enabled the rapid detection of resistant genotypes with NAATs, such as Xpert MTB/RIF or others. However, these techniques are limited so that phenotypic tests involving culture are still being used for many drugs. In recent years, to overcome the limited number of loci tested by molecular techniques, whole-genome sequencing (WGS) has been proposed as a rapid alternative method of detection [48–50], even with the possibility to be performed from sputum samples [34]. With whole-genome sequence data, bioinformatic tools such as TBProfiler can predict drug resistance, based on known genetic markers [51].

Thanks to phenotypic-genotypic studies, mutations associated with resistance to several drugs have been well characterised, despite the lack of understanding of some mechanisms of action. For instance, mutations in *katG* and *inhA* are known to confer resistance to INH; mutations in *rpoB* to RIF; mutations in *embB* to EMB; mutations in *pncA* to PZA; and mutations in *gyrA* and *gyrB* to fluoroquinolones [38, 52]. Moreover, even though the roll-out of BDQ and DLM is relatively recent, the appearance of mutations conferring resistance to these new drugs in clinical isolates has already been reported [53, 54]. These include mutations in *atpE* for BDQ and mutations in *ddn* and the enzymes involved in the F₄₂₀ coenzyme system for DLM [38, 52]. It is also important to highlight the potential cross-resistance that can be given by shared mechanisms of action (*e.g.*, DLM and PTM [55]), or by the activity of efflux pumps on specific drugs. The latter situation can be exemplified by the cross-resistance of BDQ and clofazimine (CFZ), where mutations in the transcriptional repressor *mmpR5* of the efflux pump encoded by *mmpL5-mmpS5* leads to increased minimum inhibitory concentrations (MIC) to both drugs [56].

1.6. *Mycobacterium tuberculosis*

1.6.1. *Mycobacterium tuberculosis* complex and strain diversity

Mycobacterium tuberculosis, the aetiological agent of human tuberculosis, is a slow-growing acid-fast bacteria with a peculiar lipid-rich cell envelope structure. Despite being classified as Gram-positive bacteria, its cell wall contains an outer membrane similar to Gram-negative bacteria, composed by an asymmetric lipid bilayer with the characteristic mycolic acids, and a layer of peptidoglycan in the periplasmic space [57]. This particular mycobacterial cell wall provides protection against hydrophilic compounds, conferring natural resistance to specific

drugs. Moreover, some components of the cell wall, such as the well-known ESAT-6 secretion system (ESX1-5), play an important role in virulence and host-pathogen interaction [9, 58–60]. *Mtb* belongs to the *M. tuberculosis* complex (MTBC) along with other human-adapted species, such as *M. africanum*, animal-adapted lineages and the denominated “smooth tubercle bacilli” [61]. The phylogenetic classification of the main human-adapted *Mtb* strains consists of 7 lineages, 5 within the *M. tuberculosis sensu stricto* (L1-L4 and L7) and 2 *M. africanum* (L5-6), with a different geographical distribution: Indo-Oceanic (L1), East Asian including Beijing (L2), East African-Indian (L3), Euro-American (L4), Ethiopian (L7), West African 1 (L5) and West African 2 (L6) [61, 62] (**Figure 3A**). These lineages have also been divided into two clades based on the presence or absence of the TbD1 deletion [63], being L2-4 considered the “modern” lineages, whilst L1 and L5-6 the “ancient” ones. In this classification, L7 holds an intermediate position. Despite the temporal connotation, all MTBC lineages have simultaneously evolved from the common ancestor, and therefore it does not necessarily indicate an evolutionary time dimension [64]. In recent years, two other lineages designated L8 and L9 have been discovered in East Africa [65, 66], the latter as a divergent group within *M. africanum*. Interestingly, virulence and pathogenicity across the different lineages have shown to be variable. For instance, specific characteristics of Beijing strains (L2) result in a notably higher virulence and spreading capacity [67–71]. Overall, the study of the genetic diversity in the different strains has shed light on transmission dynamics, virulence, pathogenicity and acquisition of drug resistance.

1.6.2. Genomic diversity

The sequencing of *Mtb* in 1998 revealed a 4.4 Mb genome with a high GC content (~65%) that comprises more than 4,000 genes [72]. Due to the low mutation rate observed [73] and the lack of horizontal gene transfer, the *Mtb* genome has traditionally been considered to have

limited variability and, compared to other bacteria, to be stable and largely clonal [61, 74, 75]. In general, the maximum SNP distance between any two human-adapted strains is approximately 1,200 SNPs [76]. A strong linkage between sites is one of the consequences of its clonality, resulting in genetic hitchhiking and background selection, phenomena where variants are selected or deleted according to its linkage to, for example, selection of drug resistance mutations or deletion of deleterious variants respectively [77, 78]. Although MTBC populations have shown an overall purifying selection (average pairwise $dN/dS = 0.57$), the ratio of non-synonymous SNPs (nsSNPs) to synonymous SNPs (sSNPs) (dN/dS) is higher than in other bacteria [79, 80], with evidence of some genes being under positive selection [81]. It is interesting that a large proportion of nsSNPs present in coding regions in *Mtb* have been found to be highly conserved among other mycobacteria species or to be fixed within a lineage, thus suggesting functional consequences [79, 80]. Drug resistance conferring mutations are often found under positive selection [82, 83] and convergent evolution, evolving independently in a phylogenetic tree multiple times [84]. These mutations frequently involve deleterious effects and fitness costs that are compensated by the appearance of other mutations, for example, compensatory mutations in *rpoA* and *rpoC* in RIF resistant strains [85]. Even though the mutation rate is low, the acquisition of drug resistance mutations in bacterial sub-populations can happen within weeks of treatment, and some studies have shown how different genotypes can coexist within the host, with the associated risk of misdiagnosis on detection of drug resistance [86, 87]. Moreover, evidence of co-divergence between MTBC lineages and human mitochondrial populations suggests the existence of co-phylogenies of MTBC and the human host [66].

Genetic differences between and within lineages have been widely described, being driven

by SNPs, small insertions and deletions (indels), large genomic deletions, large duplications, and mobile and repetitive elements. Among the MTBC, L1 holds the highest genetic diversity followed by L4 and L6 (*Figure 3B*) [76]. Some of these variants have been traditionally used for the genotyping of MTBC strains. For example, *IS6110* is an MTBC-specific insertion element whose position and copy number has been used for strain characterisation [88]. The presence/absence of specific spacer regions between conserved repeated sequences, called direct repeats (DR), or the length of tandem repeats have also been utilised (Spoligotyping and MIRU-VNTR typing, respectively) [89]. Finally, large deletions relative to the H37Rv reference genome denominated as regions of difference (RDs) have been identified in the different lineages defining the currently used phylogeny, often associated with deleterious effects or attenuation [90,91]. The current availability of whole-genome sequencing technologies provides with a more comprehensive, precise and informative genotyping method. The whole-genome characterisation of different lineages and sub-lineages motivated the development of SNP barcodes that, in combination with *in silico* profiling tools such as TBProfiler, can be used for the phylogenetic and resistance classification of MTBC strains [51, 92, 93].

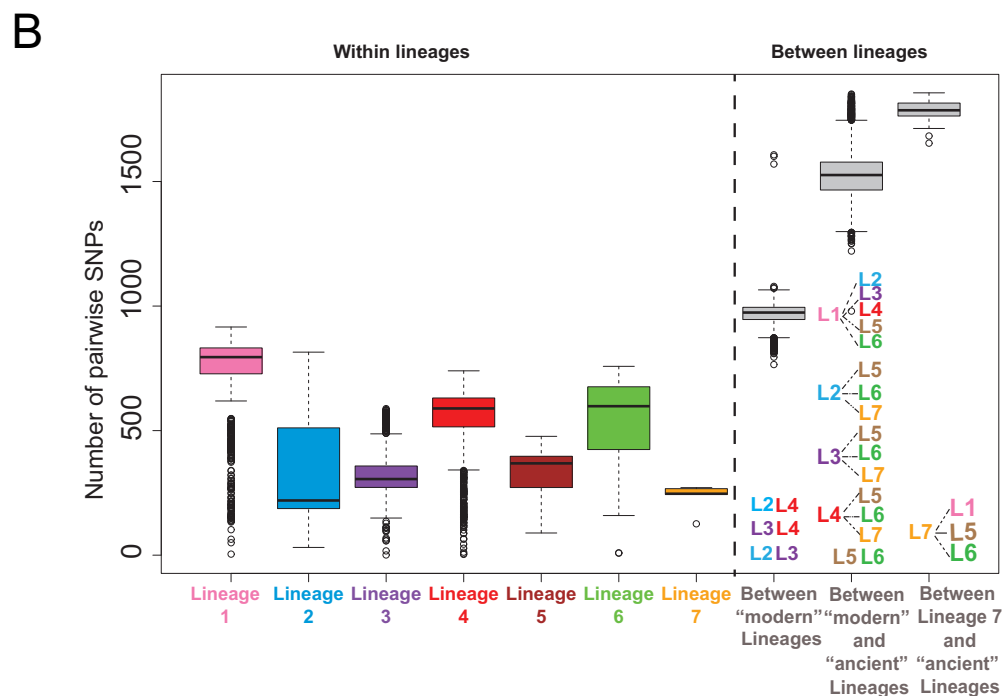


Figure 3. (A) Maximum-likelihood phylogenetic tree with 249 representative MTBC genomes from L1-9, taken from Coscolla *et al.*, 2021. [66] (B) Pairwise SNP distance within lineage (on the left) and between lineages (on the right) calculated for each pair of strains (total n=217 MTBC genomes; L1=44, L2=37, L3=36, L4=64, L5=16, L6=17, L7=4), taken from Coscolla *et al.*, 2014 [76].

1.6.3. Transcriptomics

Transcription is the next step in the central dogma of molecular biology, and thereby genetic diversity is likely to have a role in gene expression with potential phenotypic impact and implications in pathogenicity and clinical outcomes. The regulatory mechanisms of gene expression under different environmental cues have been broadly studied in *Mtb*. One of the most intriguing questions of the *Mtb* biology is the adaptation of the bacteria to the dormant state. Thereby, the investigation of the transcriptomic profiles under the conditions found within the granuloma, such as hypoxia or nutrient starvation, has revealed insights into the adaptation of the different metabolic pathways to these conditions [94]. Drug exposure has implications in gene expression too, with drug resistant isolates showing different transcriptomic profiles to susceptible ones [95, 96]. In spite of these observations of differential gene expression under environmental cues, Gao *et al.* showed how ten clinical isolates grown in liquid culture differed in their expression profiles, demonstrating the strain-to-strain variation at a transcription level [97]. Furthermore, lineage-specific transcriptomes, with a significant number of genes differentially expressed between ancient and modern strains, have been reported *in vitro* and during survival in macrophages [80, 98]. The DosR regulon, which comprises 48 genes involved in metabolism, anaerobic respiration and stress responses, closely related with dormancy and latent infection, represents a characteristic aspect of Beijing isolates, being constitutively over-expressed [99–101]. This transcriptional alteration has been suggested to be the result of a sSNP in the *dosR-dosS* operon [102], although a 350 kb gene duplication including the *dosR* operon probably has implications in the increased expression too [103, 104]. Therefore, genomic variants can have a direct impact on gene expression. The overexpression of the MmpS5/MmpL5 efflux pump as a result of mutations in its transcriptional regulator *mmpR5* is another example of mutations with expression consequences, and

in this case, associated drug resistance [105]. Nevertheless, little is known about the effect of genomic variants on gene expression levels at a genome-wide scale in *Mtb*, with a single study investigating the effect of mutations on transcriptional regulators and promoter regions [80]. Such studies can be performed through association analysis known as expression quantitative trait loci (eQTLs).

1.6.4. Epigenetics: DNA methylation

Epigenetic mechanisms involve changes in the chromosome, without altering the genetic sequence, which regulate different cell processes. One of these mechanisms is DNA methylation. In prokaryotic cells, epigenetic mechanisms control different biological processes such as timing of DNA replication and repair or chromosome partitioning, by regulating specific DNA-protein interactions, essentially via DNA methylation [106, 107]. Different types of methyltransferases (MTases) are involved in DNA methylation, sometimes as part of restriction-modification systems that constitute a defence mechanism against, for example, exogenous viral DNA [108]. In contrast to eukaryotic cells, these bacterial MTases target specific motifs, which are often found methylated in high proportions [106, 109]. The recent development of long-read sequencing platforms, such as PacBio single-molecule real time (SMRT) or Oxford Nanopore Technologies (ONT) has significantly facilitated the study of bacterial methylomes. Two types of DNA modification are predominantly established in bacteria: N⁶-methyl-adenine (6mA) and C⁴-methyl-cytosine (4mC). However, 6mA is better characterised as an epigenetic regulator in bacteria [106], and the only one found within the modified motifs identified in the MTBC: CTCCAG and GATN₄RTAC and their partner motifs, methylated on both strands, and the hemimethylated CACGCAG [109, 110]. Three MTases are responsible for the methylation of those motifs: MamA, HsdM and MamB, respectively. MamA and MamB are predicted to be type

II MTases, whilst HsdM is a type I MTase, with two specificity subunits (HsdS.1 and HsdS.2) [109–111]. Nevertheless, all of them are considered to be orphan enzymes, like Dam MTase in *E. coli*, as they do not have any cognate restriction enzyme associated. The recent study of different lineages across the MTBC revealed lineage-specific methylation profiles, where not all the three motifs were modified in some strains [110]. The concomitant identification of potential loss of function (LoF) mutations in the respective MTases has been proposed as an explanation of the absence of methylation patterns [109, 110].

Several studies have shown how methylation plays a role in gene expression regulation in bacteria through alteration of the DNA structure or steric hindrance so that binding to regulatory proteins becomes affected [106, 108]. Methylation-induced phase variation and phase-variable MTases that can cause genome-wide gene-expression changes and consequent phenotypic differences, as well as direct regulation of specific genes, have been described [108, 112, 113]. One notable consequence of epigenetic regulation recently reported is the emergence of drug resistance [114]. Thus, environmental cues can alter gene expression through competition between transcription factors and MTases [113]. Unlike eukaryotic cells, where DNA methylation is often associated with repression of gene expression, down-regulation of certain genes as a consequence of the absence of methylation has also been observed in bacteria [111, 112]. In *Mtb*, Shell *et al.* showed how disruption of the *mamA* gene decreased expression of several genes and affected survival during hypoxia [111]. Methylation sites were also found to overlap with sigma factor binding sites, all together suggesting its role in transcription. Moreover, changes in the transcriptome and methylome of INH or RIF resistant *Mtb* has given insights on the epigenetics mechanisms of induced antibiotic resistance [115].

1.7. The *pe* and *ppe* genes

With the sequencing of the whole genome of *Mtb* in 1998, the unique *pe* and *ppe* gene families were discovered [72]. The *pe* (100 loci) and *ppe* (69 loci) genes are found scattered throughout the genome and constitute approximately the 10% of the coding potential. They were characterised by the presence of their conserved N-terminal domains with the distinctive PE (proline-glutamate) and PPE (proline-proline-glutamate) motifs [72], which are found in the first ~110 and ~180 residues respectively. In comparison, the C-terminal domains vary significantly in size and sequence among members of these two families, often sharing particular motifs that classified them in further subfamilies [116]. Moreover, some of these genes are distinguished by their content on repetitive regions, like the polymorphic GC-rich repetitive sequences (PGRS) or the major polymorphic tandem repeat (MPTR) [117] (**Figure 4A**). Evolution studies of the *pe/ppe* genes have shown their close association with the ESX secretion system [116]. Their expansion has been proposed to occur through duplication of the ESAT-6 gene clusters, where insertion, deletions and homologous recombination are thought to have played a role too [116, 118, 119]. Thus, five sub-families can be distinguished in each family, with the *pe_pgrs* and the *ppe_mptr* being the most polymorphic and most recently originated [116] (**Figure 4B and 4C**). Interestingly, some of these genes, especially members of the subfamilies V (*pe_pgrs/ppe_mptr*), comprise some of the most variable regions of the *Mtb* genome, with hot spots for polymorphisms and recombination having been found among them [120–124]. For this reason, the accurate alignment and analysis of these genes is difficult and they have been systematically excluded from whole-genome studies [124–126].

Although the function of the PE/PPE proteins is still widely unknown, their cellular localisation together with their higher abundance in pathogenic mycobacteria compared to sapro-

phytic or avirulent strains, has suggested an important role during infection [116, 127, 128]. The study of individual genes has revealed their highly immunogenic nature, demonstrating their role in host-pathogen interactions and potential use as targets for vaccine and diagnostic development [129]. Moreover, due to their hypervariability, they have been proposed as mechanisms of antigenic variation and immune evasion [122, 127, 130], although the location of predicted T-cell epitopes in the highly conserved PE domains counters this hypothesis [131]. Additionally, PPE38 seems to play an essential role in the secretion of PE_PGRS and PPE_MPTR proteins, whose disruption in Beijing isolates and consequent lack of secretion is proposed to result in hypervirulence [132]. The known functions of PE/PPE proteins are various, from preventing phagosome maturation enhancing survival like PE_PGRS30 [133], to triggering autophagy like PE_PGRS29 [134], driving anti-inflammatory Th2 immune responses like PPE34 [135] or inducing pro-inflammatory cytokines like PE_PGRS33 [136]. Several pairs of *pe* and *ppe* genes are organised in operons, being transcribed together and are suggested to interact with each other forming heterodimers [137, 138]. For instance, the crystal structure of the PE25/PPE41 pair has been solved, demonstrating how protein folding is dependent on the protein interaction [139]. Nevertheless, structural data of PE/PPE proteins is scarce, which hinders the elucidation of the functional consequences of their variability [117].

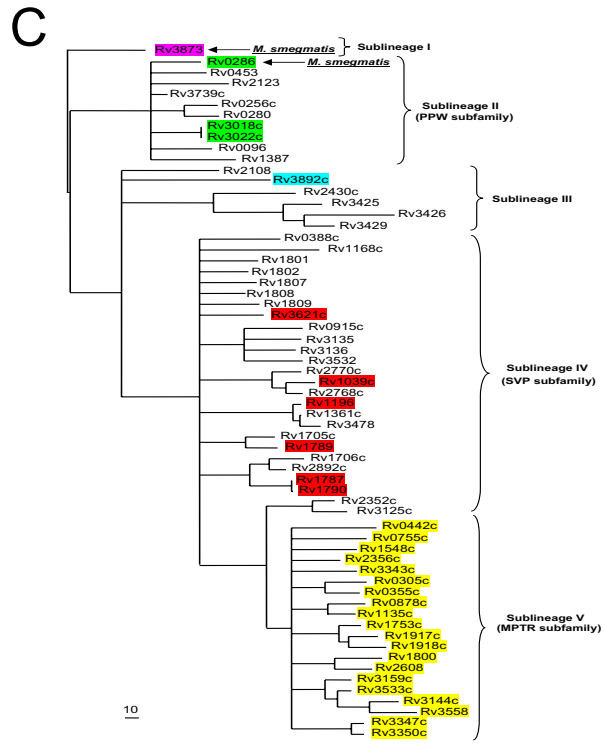
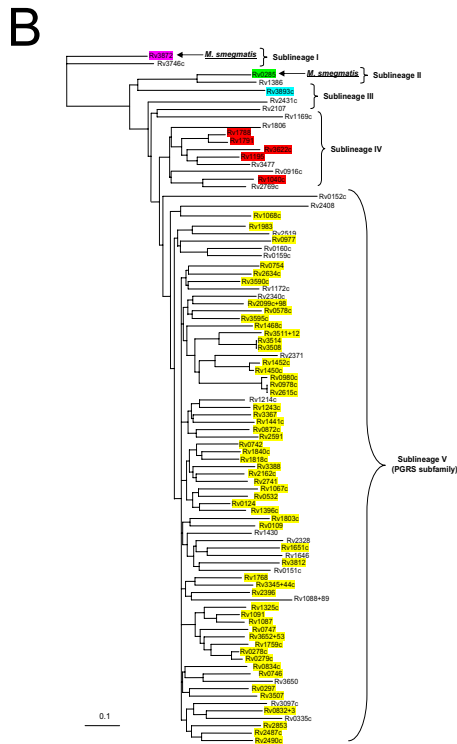
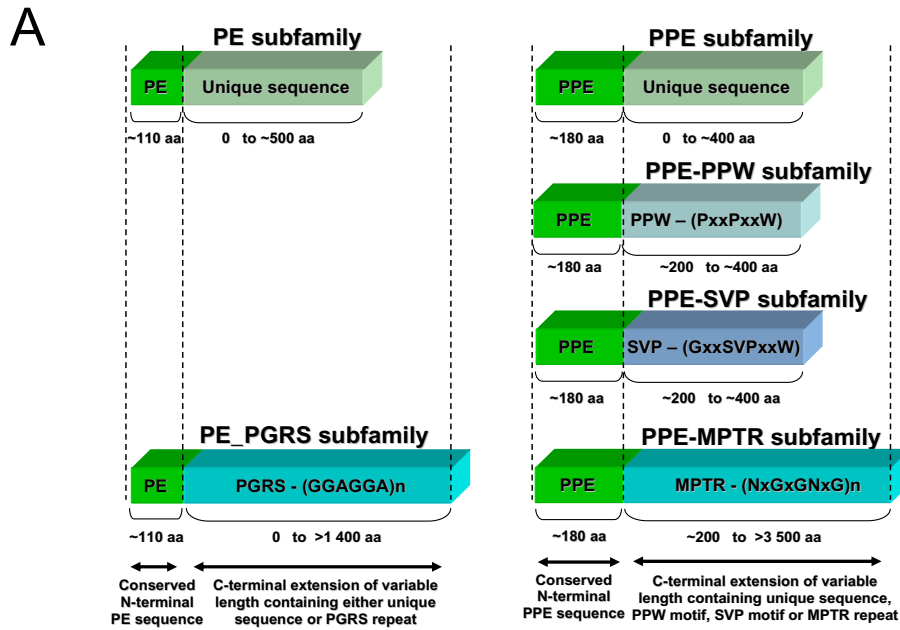


Figure 4. (A) Structure of the PE and PPE proteins. (B) PE proteins phylogenetic tree. (C) PPE proteins phylogenetic tree, taken from Gey Van Pittius *et al.*, 2006 [116].

1.8. Whole-genome sequencing and 'omics

1.8.1. Next-generation sequencing: short- and long-read sequencing technologies

The founding method of sequencing was Sanger technology, developed in 1977 [140], which, after improvements and automation processes, still remains the method of choice for some applications, like verification of plasmid constructs [141]. However, Sanger sequencing is inefficient for high throughput applications, and thus the relatively recent introduction of new whole-genome sequencing (WGS) technologies has provided the means to perform research at a faster and larger scale. On the basis of Sanger sequencing, the so called “next-generation sequencing” (NGS) technologies have recently expanded. These include mainly two types: (i) sequencing by synthesis (SBS), and (ii) single-molecule sequencing (SMS). NGS relies on the same principles of Sanger sequencing: fragmentation of the DNA/RNA of interest, generation of the sequencing library with the attachment of platform-specific adapters, and sequencing by template amplification [142]. But overall, NGS has an enhanced data-generation capacity, with the possibility of parallel sequencing of multiple genomes (“multiplexing”) at a reduced cost and higher throughput when compared to Sanger sequencing [142].

There are different approaches within the SBS methods, such as 454 pyrosequencing, Ion Torrent or Illumina sequencing, the latter being the most popular to date. In general, they separate the DNA molecules in millions of wells where they undergo PCR or isothermal amplification prior to sequencing in order to achieve the “massively parallel” principle of the technology. They generate short reads (~150-500 bp) with very high sequence coverage (millions of reads) [141, 142]. Nevertheless, one pitfall of SBS methods derives from the amplification process, where artefacts due to the inherent error rate of polymerases, can lead to false positive variants. Moreover, the performance of these technologies on repetitive regions and

high or low GC content fragments drops, which also limits the read lengths that can be obtained [126, 142]. For this reason, short-read data is more suitable for alignment to reference genomes than *de novo* assembly [142].

The development of SMS technologies has tried to overcome the drawbacks of SBS short-read. These methods have been mainly commercialised by Pacific Biosciences (PacBio) [143] and Oxford Nanopore Technologies (ONT) [144]. SMS methods attempt to sequence long DNA molecules and thus obviate the amplification step required in SBS, through a single molecule approach. Although they can produce longer reads (>15 kbp), they usually have a higher error rate than short-read technologies [142]. But overall, long-read data can lead to high quality *de novo* assembly and helps to characterise repetitive regions, such as the *pe/ppc* genes in MTBC [145]. The PacBio SMRT (Single Molecule Real Time) sequencing immobilises an engineered DNA polymerase together with the template DNA molecule inside small chambers, where incorporation of fluorescent labelled nucleotides is detected, thereby enabling real-time base-calling [143]. One advantage of this platform is that the inter-pulse duration (IPD) or speed at which each nucleotide is incorporated, can be captured and methylation of adenine and cytosine bases can be identified [146]. On the other hand, Nanopore technology method is based on characteristic electronic signals produced by the nucleotides as they travel through a pore. Methylated bases can also be distinguished, but in contrast to the modification detection by the polymerase kinetics from PacBio, Oxford Nanopore platforms rely on converting electric signal to base calls [141]. As a benefit of Oxford Nanopore, it is important to highlight the portable nature of the MinION device, powered only by a laptop, which reduces the infrastructure needed and facilitates its application on-site [147].

1.8.2. Application in 'omics

The use of NGS has brought several applications that arise from the access to whole-genome sequence data in combination with bioinformatic pipelines. These 'omics approaches, such as genomics and transcriptomics, together with multi-omic strategies have facilitated research in different fields. In pathogen genomics, genotyping methods were significantly improved with the introduction of NGS, which has enabled a better characterisation of different organisms and their genomic variation [148]. For instance, in *Mtb*, the establishment of a genetic barcode built with a subset of SNPs has been successfully achieved for lineage identification and implemented for profiling purposes [51,92,93]. Moreover, whole-genome sequence data have assisted with more accurate phylogenetic reconstructions and a better understanding of transmission dynamics [149]. An important application is also the *in silico* prediction of drug resistance, which can be accomplished with the use of NGS along with mutation libraries, previously identified through genotype-phenotype studies. TBProfiler constitutes an example of a bioinformatic tool for drug-resistance prediction based on whole-genome data [51]. Additionally, transcriptomics studies have also been benefited by NGS technologies, with the development of RNA-seq. With improved accuracy and resolution than the previous microarray methods, RNA-seq studies have been used for the better understanding of the biology of organisms, in this case, *Mtb*. One application of multi-omics is the study of eQTLs, which are genomic markers, in general SNPs, associated with the up- or down-regulation of a specific loci, classified as *cis* or *trans* depending on the physical distance from the gene they regulate [150]. Finally, as previously mentioned, DNA methylation analysis has become more accessible as a result of the SMS technologies, which, integrated with expression data, can inform on epigenetic regulation mechanisms.

1.8.3. Analysis of NGS data

The development of NGS technologies has been in parallel with the expansion of bioinformatic pipelines for its analysis. Sequencing outputs are generated as raw reads stored in different formats depending on the sequencing platform (*e.g.*, fastq files from Illumina, or fast5 files from MinION). Based on the desired downstream analysis, these reads can be either aligned to a reference genome or assembled to generate a complete genome without a reference (*de novo* assembly). There are multiple programs that perform *de novo* assembly, *e.g.* HGAP [151] or Flye [152], which overlap the reads creating longer “contigs” in an attempt to complete the genome. Similarly, different programs exist for mapping, BWA [153] being one of the most commonly used for short-read data, and minimap2 [154] for long-reads. The approach consists of algorithms that find the best possible alignment position of a read against the reference, generating a BAM file where this information is stored. Variants can be extracted from these BAM files, as well as obtained from assembled genomes through their alignment to a reference and are usually stored in a variant call file (VCF). Different steps of variant filtering can be performed prior to downstream analysis to discard the low-quality variants. Moreover, they can be used for the reconstruction of maximum-likelihood phylogenetic trees or population genetics analysis. RNA-seq data analysis involves the same methodology, where reads are aligned to an annotated reference, so that the number of reads mapped to each gene can be quantified. Programs like HTSeq [155] and DESeq2 [156] can be utilised for counting, normalising and carrying out differential expression analysis. For the DNA methylation analysis, PacBio provides a Motif and Methylation software pipeline, whilst different tools have been developed for the extraction and analysis of similar data from MinION reads. Overall, there is a wide range of bioinformatic programs and pipelines that enable and facilitate the analysis of high throughput NGS data.

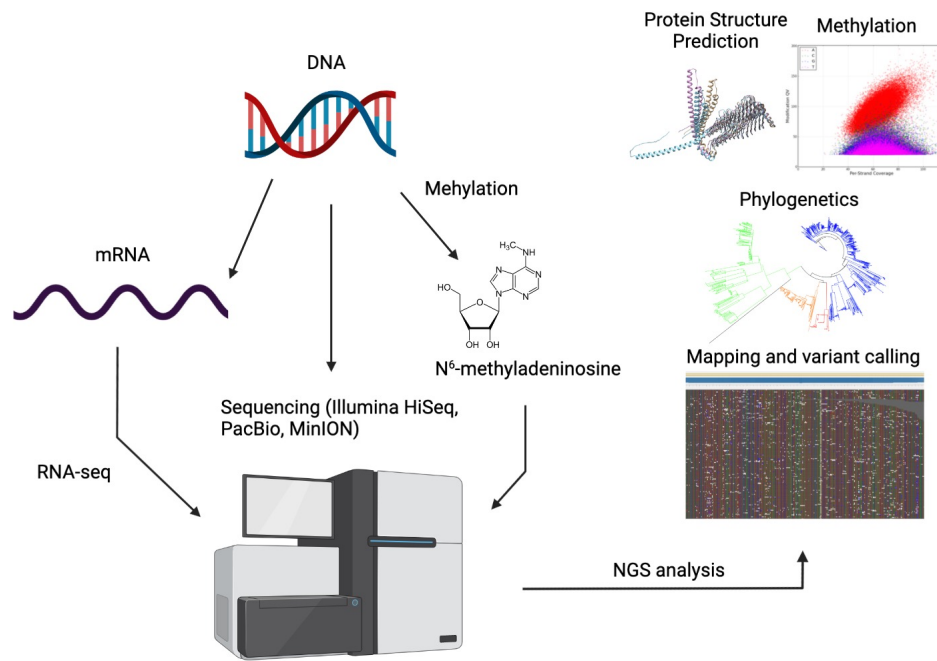


Figure 5. Schematic description of NGS data generation and analysis for DNA sequencing, RNA-seq and DNA methylation.

References

- [1] Daniel, T. M. The history of tuberculosis. *Respiratory Medicine* **100**, 1862–1870 (2006).
- [2] World Health Organization, W. Global Tuberculosis Report 2021. Tech. Rep. (2021).
- [3] Vynnycky, E. & Fine, P. E. Lifetime risks, incubation period, and serial interval of tuberculosis. *American Journal of Epidemiology* **152**, 247–263 (2000).
- [4] Dheda, K. *et al.* The intersection pandemics of tuberculosis and covid-19: population-level and patient-level impact, clinical presentation, and corrective interventions. *The Lancet Respiratory Medicine* **10**, 603–622 (2022).
- [5] McQuaid, C. F., Vassall, A., Cohen, T., Fiekert, K. & White, R. G. The impact of covid-19 on tb: A review of the data. *International journal of Tuberculosis and Lung Disease* **25**, 436–446 (2021).
- [6] Mayer-Barber, K. D. & Barber, D. L. Innate and Adaptive Cellular Immune Responses to *Mycobacterium tuberculosis* Infection. *Cold Spring Harbor Perspectives in Medicine* **5**, a018424 (2015).
- [7] Guirado, E., Schlesinger, L. S. & Kaplan, G. Macrophages in tuberculosis: friend or foe. *Seminars in Immunopathology* **35**, 563–583 (2013).
- [8] Gengenbacher, M. & Kaufmann, S. H. *Mycobacterium tuberculosis*: success through dormancy. *FEMS Microbiology Reviews* **36**, 514–532 (2012).

- [9] Delogu, G., Sali, M. & Fadda, G. The biology of *Mycobacterium tuberculosis* infection. *Mediterranean Journal of Hematology and Infectious Diseases* **5**, e2013070 (2013).
- [10] Loddenkemper, R., Lipman, M. & Zumla, A. Clinical Aspects of Adult Tuberculosis. *Cold Spring Harbor Perspectives in Medicine* **6**, a017848 (2016).
- [11] Pagán, A. J. & Ramakrishnan, L. Immunity and Immunopathology in the Tuberculous Granuloma. *Cold Spring Harbor Perspectives in Medicine* **5**, a018499 (2015).
- [12] Walzl, G. *et al.* Clinical Immunology and Multiplex Biomarkers of Human Tuberculosis. *Cold Spring Harbor Perspectives in Medicine* **5**, a018515–a018515 (2015).
- [13] Neyrolles, O. *et al.* Is Adipose Tissue a Place for *Mycobacterium tuberculosis* Persistence? *PLoS ONE* **1**, e43 (2006).
- [14] Lawn, S. D. & Zumla, A. I. Tuberculosis. *The Lancet* **378**, 57–72 (2011).
- [15] Cambier, C. J., Falkow, S. & Ramakrishnan, L. Host evasion and exploitation schemes of *Mycobacterium tuberculosis*. *Cell* **159**, 1497–1509 (2014).
- [16] Heemskerk, D., Caws, M., Marais, B. & Farrar, J. *Tuberculosis in Adults and Children*, vol. 2 of *Springer Briefs in Public Health* (Springer International Publishing, Cham, 2015).
- [17] Sulis, G., Roggi, A., Matteelli, A. & Raviglione, M. C. TUBERCULOSIS: EPIDEMIOLOGY AND CONTROL. *Mediterranean Journal of Hematology and Infectious Diseases* **6**, e2014070 (2014).
- [18] Bellamy, R. Susceptibility to mycobacterial infections: the importance of host genetics. *Genes & Immunity* **4**, 4–11 (2003).

- [19] Stead, W. W., Senner, J. W., Reddick, W. T. & Lofgren, J. P. Racial Differences in Susceptibility to Infection by *Mycobacterium tuberculosis*. *New England Journal of Medicine* **322**, 422–427 (1990).
- [20] Cantwell, M. F., McKenna, M. T., McCray, E. & Onorato, I. M. Tuberculosis and Race/Ethnicity in the United States: impact of socioeconomic status. *American Journal of Respiratory and Critical Care Medicine* **157**, 1016–1020 (1998).
- [21] Curtis, J. *et al.* Susceptibility to tuberculosis is associated with variants in the ASAP1 gene encoding a regulator of dendritic cell migration. *Nature Genetics* **47**, 523–527 (2015).
- [22] World Health Organization, W. Chest Radiography in Tuberculosis. Tech. Rep. (2016).
- [23] Lawn, S. D. Advances in Diagnostic Assays for Tuberculosis. *Cold Spring Harbor Perspectives in Medicine* **5**, a017806 (2015).
- [24] Minion, J., Pai, M., Ramsay, A., Menzies, D. & Greenaway, C. Comparison of LED and Conventional Fluorescence Microscopy for Detection of Acid Fast Bacilli in a Low-Incidence Setting. *PLoS ONE* **6**, e22495 (2011).
- [25] Gupta, R. K. *et al.* Impact of human immunodeficiency virus and CD4 count on tuberculosis diagnosis: analysis of city-wide data from Cape Town, South Africa. *The International Journal of Tuberculosis and Lung Disease* **17**, 1014–1022 (2013).
- [26] Lawn, S. D. *et al.* Advances in tuberculosis diagnostics: the Xpert MTB/RIF assay and future prospects for a point-of-care test. *The Lancet Infectious Diseases* **13**, 349–361 (2013).
- [27] World Health Organization, W. Molecular Line Probe Assays for Rapid Screening of Patients At Risk of Multidrug-Resistant Tuberculosis (MDR-TB). Tech. Rep. June (2008).

- [28] Shete, P. B., Farr, K., Strnad, L., Gray, C. M. & Cattamanchi, A. Diagnostic accuracy of TB-LAMP for pulmonary tuberculosis: a systematic review and meta-analysis. *BMC Infectious Diseases* **19**, 268 (2019).
- [29] World Health Organization (WHO). Consolidated Guidelines on Tuberculosis. Module 3 : Diagnosis. Rapid diagnostics for tuberculosis detection. Tech. Rep. (2021).
- [30] Lee, R. S. & Pai, M. Real-Time Sequencing of *Mycobacterium tuberculosis*: Are We There Yet? *Journal of Clinical Microbiology* **55**, 1249–1254 (2017).
- [31] Lam, C. *et al.* Value of routine whole genome sequencing for *Mycobacterium tuberculosis* drug resistance detection. *International Journal of Infectious Diseases* **113**, S48–S54 (2021).
- [32] MacLean, E. *et al.* Advances in Molecular Diagnosis of Tuberculosis. *Journal of Clinical Microbiology* **58**, 1–13 (2020).
- [33] Jouet, A. *et al.* Deep amplicon sequencing for culture-free prediction of susceptibility or resistance to 13 anti-tuberculous drugs. *European Respiratory Journal* **57**, 2002338 (2021).
- [34] Doyle, R. M. *et al.* Direct whole-genome sequencing of sputum accurately identifies drug-resistant *Mycobacterium tuberculosis* faster than MGIT culture sequencing. *Journal of Clinical Microbiology* **56**, 1–11 (2018).
- [35] Pai, M. *et al.* Gamma Interferon Release Assays for Detection of *Mycobacterium tuberculosis* Infection. *Clinical Microbiology Reviews* **27**, 3–20 (2014).
- [36] Zumla, A., Raviglione, M., Hafner, R. & Fordham von Reyn, C. Tuberculosis. *New England Journal of Medicine* **368**, 745–755 (2013).

- [37] Heyckendorf, J. *et al.* Pathogen-free diagnosis of tuberculosis. *The Lancet Infectious Diseases* **21**, 1066 (2021).
- [38] Peloquin, C. A. & Davies, G. R. The Treatment of Tuberculosis. *Clinical Pharmacology & Therapeutics* **110**, 1455–1466 (2021).
- [39] World Health Organization, W. WHO consolidated guidelines on tuberculosis. Module 4: treatment. Drug-resistant tuberculosis treatment. Tech. Rep. (2020).
- [40] Sotgiu, G., Centis, R., D’ambrosio, L. & Migliori, G. B. Tuberculosis Treatment and Drug Regimens. *Cold Spring Harbor Perspectives in Medicine* **5**, a017822–a017822 (2015).
- [41] World Health Organization, W. Guidelines for treatment of drug-susceptible tuberculosis and patient care. Tech. Rep. 12 (2017).
- [42] World Health Organization, W. Latent tuberculosis infection: updated and consolidated guidelines for programmatic management. Tech. Rep. (2018).
- [43] Sable, S. B., Posey, J. E. & Scriba, T. J. Tuberculosis Vaccine Development: Progress in Clinical Evaluation. *Clinical Microbiology Reviews* **33** (2019).
- [44] Li, J. *et al.* Tuberculosis vaccine development: from classic to clinical candidates. *European Journal of Clinical Microbiology & Infectious Diseases* **39**, 1405–1425 (2020).
- [45] Nebenzahl-Guimaraes, H., Jacobson, K. R., Farhat, M. R. & Murray, M. B. Systematic review of allelic exchange experiments aimed at identifying mutations that confer drug resistance in *Mycobacterium tuberculosis*. *Journal of Antimicrobial Chemotherapy* **69**, 331–342 (2014).
- [46] Cohen, T., Sommers, B. & Murray, M. The effect of drug resistance on the fitness of *Mycobacterium tuberculosis*. *The Lancet Infectious Diseases* **3**, 13–21 (2003).

- [47] de Vos, M. *et al.* Putative Compensatory Mutations in the *rpoC* Gene of Rifampin-Resistant *Mycobacterium tuberculosis* Are Associated with Ongoing Transmission. *Antimicrobial Agents and Chemotherapy* **57**, 827–832 (2013).
- [48] Coll, F. *et al.* Rapid determination of anti-tuberculosis drug resistance from whole-genome sequences. *Genome Medicine* **7**, 51 (2015).
- [49] Witney, A. A. *et al.* Clinical Application of Whole-Genome Sequencing To Inform Treatment for Multidrug-Resistant Tuberculosis Cases. *Journal of Clinical Microbiology* **53**, 1473–1483 (2015).
- [50] Phelan, J. *et al.* The variability and reproducibility of whole genome sequencing technology for detecting resistance to anti-tuberculous drugs. *Genome Medicine* **8**, 132 (2016).
- [51] Phelan, J. E. *et al.* Integrating informatics tools and portable sequencing technology for rapid detection of resistance to anti-tuberculous drugs. *Genome Medicine* **11**, 41 (2019).
- [52] Hameed, H. M. A. *et al.* Molecular Targets Related Drug Resistance Mechanisms in MDR-, XDR-, and TDR-*Mycobacterium tuberculosis* Strains. *Frontiers in Cellular and Infection Microbiology* **8** (2018).
- [53] Bloemberg, G. V., Gagneux, S. & Böttger, E. C. Acquired resistance to bedaquiline and delamanid in therapy for tuberculosis. *New England Journal of Medicine* **373**, 1986–1988 (2015).
- [54] Hoffmann, H. *et al.* Delamanid and Bedaquiline Resistance in *Mycobacterium tuberculosis* Ancestral Beijing Genotype Causing Extensively Drug-Resistant Tuberculosis in a Tibetan Refugee. *American Journal of Respiratory and Critical Care Medicine* **193**, 337–340 (2016).

- [55] Lee, B. M. *et al.* Predicting nitroimidazole antibiotic resistance mutations in *Mycobacterium tuberculosis* with protein engineering. *PLOS Pathogens* **16**, e1008287 (2020).
- [56] Hartkoorn, R. C., Uplekar, S. & Cole, S. T. Cross-Resistance between Clofazimine and Bedaquiline through Upregulation of MmpL5 in *Mycobacterium tuberculosis*. *Antimicrobial Agents and Chemotherapy* **58**, 2979–2981 (2014).
- [57] Jackson, M. The Mycobacterial Cell Envelope - Lipids. *Cold Spring Harbor Perspectives in Medicine* **4**, a021105–a021105 (2014).
- [58] Refai, A., Gritli, S., Barbouche, M.-R. & Essafi, M. Mycobacterium tuberculosis Virulent Factor ESAT-6 Drives Macrophage Differentiation Toward the Pro-inflammatory M1 Phenotype and Subsequently Switches It to the Anti-inflammatory M2 Phenotype. *Frontiers in Cellular and Infection Microbiology* **8**, 1–14 (2018).
- [59] Majlessi, L. *et al.* Influence of ESAT-6 Secretion System 1 (RD1) of *Mycobacterium tuberculosis* on the Interaction between Mycobacteria and the Host Immune System. *The Journal of Immunology* **174**, 3570–3579 (2005).
- [60] Bottai, D. *et al.* Disruption of the ESX-5 system of *Mycobacterium tuberculosis* causes loss of PPE protein secretion, reduction of cell wall integrity and strong attenuation. *Molecular Microbiology* **83**, 1195–1209 (2012).
- [61] Brites, D. & Gagneux, S. The Nature and Evolution of Genomic Diversity in the *Mycobacterium tuberculosis* Complex. In *Strain Variation in the Mycobacterium tuberculosis Complex: Its role in biology, epidemiology and control*, 1–26 (2017).
- [62] Firdessa, R. *et al.* Mycobacterial Lineages Causing Pulmonary and Extrapulmonary Tuberculosis, Ethiopia. *Emerging Infectious Diseases* **19**, 460–463 (2013).

- [63] Brosch, R. *et al.* A new evolutionary scenario for the *Mycobacterium tuberculosis* complex. *Proceedings of the National Academy of Sciences* **99**, 3684–3689 (2002).
- [64] Smith, N. H., Hewinson, R. G., Kremer, K., Brosch, R. & Gordon, S. V. Myths and misconceptions: the origin and evolution of *Mycobacterium tuberculosis*. *Nature Reviews Microbiology* **7**, 537–544 (2009).
- [65] Ngabonziza, J. C. S. *et al.* A sister lineage of the *Mycobacterium tuberculosis* complex discovered in the African Great Lakes region. *Nature Communications* **11**, 2917 (2020).
- [66] Coscolla, M. *et al.* Phylogenomics of *Mycobacterium africanum* reveals a new lineage and a complex evolutionary history. *Microbial Genomics* **7**, 1176–1182 (2021).
- [67] Glynn, J. R., Whiteley, J., Bifani, P. J., Kremer, K. & van Soolingen, D. Worldwide Occurrence of Beijing/W Strains of *Mycobacterium tuberculosis*: A Systematic Review. *Emerging Infectious Diseases* **8**, 843–849 (2002).
- [68] Bifani, P. J., Mathema, B., Kurepina, N. E. & Kreiswirth, B. N. Global dissemination of the *Mycobacterium tuberculosis* W-Beijing family strains. *Trends in Microbiology* **10**, 45–52 (2002).
- [69] Parwati, I., van Crevel, R. & van Soolingen, D. Possible underlying mechanisms for successful emergence of the *Mycobacterium tuberculosis* Beijing genotype strains. *The Lancet Infectious Diseases* **10**, 103–111 (2010).
- [70] Hanekom, M. *et al.* *Mycobacterium tuberculosis* Beijing genotype: A template for success. *Tuberculosis* **91**, 510–523 (2011).
- [71] Merker, M. *et al.* Evolutionary history and global spread of the *Mycobacterium tuberculosis* Beijing lineage. *Nature Genetics* **47**, 242–249 (2015).

- [72] Cole, S. T. *et al.* Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* **393**, 537–544 (1998).
- [73] Grant, A., Arnold, C., Thorne, N., Gharbia, S. & Underwood, A. Mathematical Modelling of *Mycobacterium tuberculosis* VNTR Loci Estimates a Very Slow Mutation Rate for the Repeats. *Journal of Molecular Evolution* **66**, 565–574 (2008).
- [74] Supply, P. *et al.* Genomic analysis of smooth tubercle bacilli provides insights into ancestry and pathoadaptation of *Mycobacterium tuberculosis*. *Nature Genetics* **45**, 172–179 (2013).
- [75] Boritsch, E. C. *et al.* Key experimental evidence of chromosomal DNA transfer among selected tuberculosis-causing mycobacteria. *Proceedings of the National Academy of Sciences* **113**, 9876–9881 (2016).
- [76] Coscolla, M. & Gagneux, S. Consequences of genomic diversity in *Mycobacterium tuberculosis*. *Seminars in Immunology* **26**, 431–444 (2014).
- [77] Eldholm, V. *et al.* Evolution of extensively drug-resistant *Mycobacterium tuberculosis* from a susceptible ancestor in a single patient. *Genome Biology* **15**, 490 (2014).
- [78] Pepperell, C. S. *et al.* The Role of Selection in Shaping Diversity of Natural *M. tuberculosis* Populations. *PLoS Pathogens* **9**, e1003543 (2013).
- [79] Hershberg, R. *et al.* High Functional Diversity in *Mycobacterium tuberculosis* Driven by Genetic Drift and Human Demography. *PLoS Biology* **6**, e311 (2008).
- [80] Rose, G. *et al.* Mapping of Genotype–Phenotype Diversity among Clinical Isolates of *Mycobacterium tuberculosis* by Sequence-Based Transcriptional Profiling. *Genome Biology and Evolution* **5**, 1849–1862 (2013).

- [81] Zhang, Y., Zhang, H., Zhou, T., Zhong, Y. & Jin, Q. Genes under positive selection in *Mycobacterium tuberculosis*. *Computational Biology and Chemistry* **35**, 319–322 (2011).
- [82] Farhat, M. R. *et al.* Genomic analysis identifies targets of convergent positive selection in drug-resistant *Mycobacterium tuberculosis*. *Nature Genetics* **45**, 1183–1189 (2013).
- [83] Zhang, H. *et al.* Genome sequencing of 161 *Mycobacterium tuberculosis* isolates from China identifies genes and intergenic regions associated with drug resistance. *Nature Genetics* **45**, 1255–1260 (2013).
- [84] Godfroid, M. *et al.* Insertion and deletion evolution reflects antibiotics selection pressure in a *Mycobacterium tuberculosis* outbreak. *PLOS Pathogens* **16**, e1008357 (2020).
- [85] Comas, I. *et al.* Whole-genome sequencing of rifampicin-resistant *Mycobacterium tuberculosis* strains identifies compensatory mutations in RNA polymerase genes. *Nature Genetics* **44**, 106–110 (2012).
- [86] Lieberman, T. D. *et al.* Genomic diversity in autopsy samples reveals within-host dissemination of HIV-associated *Mycobacterium tuberculosis*. *Nature Medicine* **22**, 1470–1474 (2016).
- [87] Trauner, A. *et al.* The within-host population dynamics of *Mycobacterium tuberculosis* vary with treatment efficacy. *Genome Biology* **18**, 71 (2017).
- [88] van Embden, J. D. *et al.* Strain identification of *Mycobacterium tuberculosis* by DNA fingerprinting: recommendations for a standardized methodology. *Journal of Clinical Microbiology* **31**, 406–409 (1993).

- [89] Merker, M., Kohl, T. A., Niemann, S. & Supply, P. The Evolution of Strain Typing in the *Mycobacterium tuberculosis* Complex. In *Strain Variation in the Mycobacterium tuberculosis Complex: Its role in biology, epidemiology and control*, 43–78 (2017).
- [90] Tsolaki, A. G. *et al.* Functional and evolutionary genomics of *Mycobacterium tuberculosis*: Insights from genomic deletions in 100 strains. *Proceedings of the National Academy of Sciences* **101**, 4865–4870 (2004).
- [91] Mostowy, S. *et al.* Genomic Analysis Distinguishes *Mycobacterium africanum*. *Journal of Clinical Microbiology* **42**, 3594–3599 (2004).
- [92] Coll, F. *et al.* A robust SNP barcode for typing *Mycobacterium tuberculosis* complex strains. *Nature Communications* **5**, 4812 (2014).
- [93] Napier, G. *et al.* Robust barcoding and identification of *Mycobacterium tuberculosis* lineages for epidemiological and clinical studies. *Genome Medicine* **12**, 114 (2020).
- [94] Kundu, M. & Basu, J. Applications of Transcriptomics and Proteomics for Understanding Dormancy and Resuscitation in *Mycobacterium tuberculosis*. *Frontiers in Microbiology* **12**, 1–16 (2021).
- [95] Briffotiaux, J., Liu, S. & Gicquel, B. Genome-Wide Transcriptional Responses of *Mycobacterium* to Antibiotics. *Frontiers in Microbiology* **10**, 1–14 (2019).
- [96] Yu, G. *et al.* Gene expression analysis of two extensively drug-resistant tuberculosis isolates show that two-component response systems enhance drug resistance. *Tuberculosis* **95**, 303–314 (2015).
- [97] Gao, Q. *et al.* Gene expression diversity among *Mycobacterium tuberculosis* clinical isolates. *Microbiology* **151**, 5–14 (2005).

- [98] Homolka, S., Niemann, S., Russell, D. G. & Rohde, K. H. Functional Genetic Diversity among *Mycobacterium tuberculosis* Complex Clinical Isolates: Delineation of Conserved Core and Lineage-Specific Transcriptomes during Intracellular Survival. *PLoS Pathogens* **6**, e1000988 (2010).
- [99] Reed, M. B., Gagneux, S., DeRiemer, K., Small, P. M. & Barry, C. E. The W-Beijing Lineage of *Mycobacterium tuberculosis* Overproduces Triglycerides and Has the DosR Dormancy Regulon Constitutively Upregulated. *Journal of Bacteriology* **189**, 2583–2589 (2007).
- [100] Fallow, A., Domenech, P. & Reed, M. B. Strains of the East Asian (W/Beijing) Lineage of *Mycobacterium tuberculosis* Are DosS/DosT-DosR Two-Component Regulatory System Natural Mutants. *Journal of Bacteriology* **192**, 2228–2238 (2010).
- [101] Sherman, D. R. *et al.* Regulation of the *Mycobacterium tuberculosis* hypoxic response gene encoding α -crystallin. *Proceedings of the National Academy of Sciences* **98**, 7534–7539 (2001).
- [102] Domenech, P. *et al.* Unique Regulation of the DosR Regulon in the Beijing Lineage of *Mycobacterium tuberculosis*. *Journal of Bacteriology* **199**, 1–19 (2017).
- [103] Domenech, P., Kolly, G. S., Leon-Solis, L., Fallow, A. & Reed, M. B. Massive Gene Duplication Event among Clinical Isolates of the *Mycobacterium tuberculosis* W/Beijing Family. *Journal of Bacteriology* **192**, 4562–4570 (2010).
- [104] Weiner, B. *et al.* Independent Large Scale Duplications in Multiple *M. tuberculosis* Lineages Overlapping the Same Genomic Region. *PLoS ONE* **7**, e26038 (2012).
- [105] Andries, K. *et al.* Acquired Resistance of *Mycobacterium tuberculosis* to Bedaquiline. *PLoS ONE* **9**, e102135 (2014).

- [106] Casadesús, J. & Low, D. Epigenetic Gene Regulation in the Bacterial World. *Microbiology and Molecular Biology Reviews* **70**, 830–856 (2006).
- [107] Sánchez-Romero, M. A. & Casadesús, J. The bacterial epigenome. *Nature Reviews Microbiology* **18**, 7–20 (2020).
- [108] Seong, H. J., Han, S.-W. & Sul, W. J. Prokaryotic DNA methylation and its functional roles. *Journal of Microbiology* **59**, 242–248 (2021).
- [109] Zhu, L. *et al.* Precision methylome characterization of *Mycobacterium tuberculosis* complex (MTBC) using PacBio single-molecule real-time (SMRT) technology. *Nucleic Acids Research* **44**, 730–743 (2016).
- [110] Phelan, J. *et al.* Methylation in *Mycobacterium tuberculosis* is lineage specific with associated mutations present globally. *Scientific Reports* **8**, 160 (2018).
- [111] Shell, S. S. *et al.* DNA Methylation Impacts Gene Expression and Ensures Hypoxic Survival of *Mycobacterium tuberculosis*. *PLoS Pathogens* **9**, e1003419 (2013).
- [112] Balbontin-del Portillo, F., Hinton, J. C. D. & Casadesus, J. DNA Adenine Methylation Regulates Virulence Gene Expression in *Salmonella enterica* Serovar Typhimurium. *Journal of Bacteriology* **188**, 8160–8168 (2006).
- [113] Beaulaurier, J., Schadt, E. E. & Fang, G. Deciphering bacterial epigenomes using modern sequencing technologies. *Nature Reviews Genetics* **20**, 157–172 (2019).
- [114] Yuan, W. *et al.* Multiple antibiotic resistance and DNA methylation in Enterobacteriaceae isolates from different environments. *Journal of Hazardous Materials* **402**, 123822 (2021).

- [115] Chen, L. *et al.* Genome-wide DNA methylation and transcriptome and proteome changes in *Mycobacterium tuberculosis* with para-aminosalicylic acid resistance. *Chemical Biology & Drug Design* **95**, 104–112 (2020).
- [116] Gey van Pittius, N. C. *et al.* Evolution and expansion of the *Mycobacterium tuberculosis* PE and PPE multigene families and their association with the duplication of the ESAT-6 (*esx*) gene cluster regions. *BMC evolutionary biology* **6**, 95 (2006).
- [117] Fishbein, S., van Wyk, N., Warren, R. M. & Sampson, S. L. Phylogeny to function: PE/PPE protein evolution and impact on *Mycobacterium tuberculosis* pathogenicity. *Molecular Microbiology* **96**, 901–916 (2015).
- [118] Abdallah, A. M. *et al.* Type VII secretion — mycobacteria show the way. *Nature Reviews Microbiology* **5**, 883–891 (2007).
- [119] Medha, Sharma, S. & Sharma, M. Proline-Glutamate/Proline-Proline-Glutamate (PE/PPE) proteins of *Mycobacterium tuberculosis*: The multifaceted immunomodulators. *Acta Tropica* **222**, 106035 (2021).
- [120] Talarico, S. *et al.* Variation of the *Mycobacterium tuberculosis* PE_PGRS33 Gene among Clinical Isolates. *Journal of Clinical Microbiology* **43**, 4954–4960 (2005).
- [121] Karboul, A. *et al.* Frequent Homologous Recombination Events in *Mycobacterium tuberculosis* PE/PPE Multigene Families: Potential Role in Antigenic Variability. *Journal of Bacteriology* **190**, 7838–7846 (2008).
- [122] Talarico, S. *et al.* *Mycobacterium tuberculosis* PE_PGRS16 and PE_PGRS26 genetic polymorphism among clinical isolates. *Tuberculosis* **88**, 283–294 (2008).

- [123] McEvoy, C. R. E. *et al.* Comparative Analysis of *Mycobacterium tuberculosis* *pe* and *ppe* Genes Reveals High Sequence Variation and an Apparent Absence of Selective Constraints. *PLoS ONE* **7**, e30593 (2012).
- [124] Phelan, J. E. *et al.* Recombination in *pe/pppe* genes contributes to genetic variation in *Mycobacterium tuberculosis* lineages. *BMC Genomics* **17**, 151 (2016).
- [125] Meehan, C. J. *et al.* Whole genome sequencing of *Mycobacterium tuberculosis*: current standards and open issues. *Nature Reviews Microbiology* **17**, 533–545 (2019).
- [126] Modlin, S. J. *et al.* Exact mapping of Illumina blind spots in the *Mycobacterium tuberculosis* genome reveals platform-wide and workflow-specific biases. *Microbial Genomics* **7** (2021).
- [127] Akhter, Y., Ehebauer, M. T., Mukhopadhyay, S. & Hasnain, S. E. The PE/PPE multigene family codes for virulence factors and is a possible source of mycobacterial antigenic variation: Perhaps more? *Biochimie* **94**, 110–116 (2012).
- [128] McGuire, A. *et al.* Comparative analysis of *Mycobacterium* and related actinomycetes yields insight into the evolution of *Mycobacterium tuberculosis* pathogenesis. *BMC Genomics* **13**, 120 (2012).
- [129] Qian, J., Chen, R., Wang, H. & Zhang, X. Role of the PE/PPE Family in Host–Pathogen Interactions and Prospects for Anti-Tuberculosis Vaccine and Diagnostic Tool Design. *Frontiers in Cellular and Infection Microbiology* **10**, 1–8 (2020).
- [130] Tundup, S. *et al.* The Co-Operonic PE25/PPE41 Protein Complex of *Mycobacterium tuberculosis* Elicits Increased Humoral and Cell Mediated Immune Response. *PLoS ONE* **3**, e3586 (2008).

- [131] Copin, R. *et al.* Sequence Diversity in the *pe_pgrs* Genes of *Mycobacterium tuberculosis* Is Independent of Human T Cell Recognition. *mBio* **5**, 1–11 (2014).
- [132] Ates, L. S. *et al.* Mutations in *ppe38* block PE_PGRS secretion and increase virulence of *Mycobacterium tuberculosis*. *Nature Microbiology* **3**, 181–188 (2018).
- [133] Iantomasi, R. *et al.* PE_PGRS30 is required for the full virulence of *Mycobacterium tuberculosis*. *Cellular Microbiology* **14**, 356–367 (2012).
- [134] Chai, Q. *et al.* A *Mycobacterium tuberculosis* surface protein recruits ubiquitin to trigger host xenophagy. *Nature Communications* **10**, 1973 (2019).
- [135] Bansal, K. *et al.* Src Homology 3-interacting Domain of *Rv1917c* of *Mycobacterium tuberculosis* Induces Selective Maturation of Human Dendritic Cells by Regulating PI3K-MAPK-NF- κ B Signaling and Drives Th2 Immune Responses. *Journal of Biological Chemistry* **285**, 36511–36522 (2010).
- [136] Basu, S. *et al.* Execution of Macrophage Apoptosis by PE_PGRS33 of *Mycobacterium tuberculosis* Is Mediated by Toll-like Receptor 2-dependent Release of Tumor Necrosis Factor- α . *Journal of Biological Chemistry* **282**, 1039–1050 (2007).
- [137] Tundup, S., Akhter, Y., Thiagarajan, D. & Hasnain, S. E. Clusters of PE and PPE genes of *Mycobacterium tuberculosis* are organized in operons: Evidence that PE *Rv2431c* is co-transcribed with PPE *Rv2430c* and their gene products interact with each other. *FEBS Letters* **580**, 1285–1293 (2006).
- [138] Tiwari, B., Soory, A. & Raghunand, T. R. An immunomodulatory role for the *Mycobacterium tuberculosis* region of difference 1 locus proteins PE35 (*Rv3872*) and PPE68 (*Rv3873*). *FEBS Journal* **281**, 1556–1570 (2014).

- [139] Strong, M. *et al.* Toward the structural genomics of complexes: Crystal structure of a PE/PPE protein complex from *Mycobacterium tuberculosis*. *Proceedings of the National Academy of Sciences* **103**, 8060–8065 (2006).
- [140] Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences* **74**, 5463–5467 (1977).
- [141] Slatko, B. E., Gardner, A. F. & Ausubel, F. M. Overview of Next-Generation Sequencing Technologies. *Current Protocols in Molecular Biology* **122**, e59 (2018).
- [142] McCombie, W. R., McPherson, J. D. & Mardis, E. R. Next-Generation Sequencing Technologies. *Cold Spring Harbor Perspectives in Medicine* **9**, a036798 (2019).
- [143] Nakano, K. *et al.* Advantages of genome sequencing by long-read sequencer using SMRT technology in medical area. *Human Cell* **30**, 149–161 (2017).
- [144] Jain, M. *et al.* MinION Analysis and Reference Consortium: Phase 2 data release and analysis of R9.0 chemistry. *F1000Research* **6**, 760 (2017).
- [145] Elghraoui, A., Modlin, S. J. & Valafar, F. SMRT genome assembly corrects reference errors, resolving the genetic basis of virulence in *Mycobacterium tuberculosis*. *BMC Genomics* **18**, 302 (2017).
- [146] Flusberg, B. A. *et al.* Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nature Methods* **7**, 461–465 (2010).
- [147] Runtuwene, L. R., Tuda, J. S. B., Mongan, A. E. & Suzuki, Y. On-Site MinION Sequencing. In Suzuki, Y. (ed.) *Advances in Experimental Medicine and Biology*, chap. 1129, 143–150 (Springer New York LLC, 2019).

- [148] Niemann, S. & Supply, P. Diversity and Evolution of *Mycobacterium tuberculosis*: Moving to Whole-Genome-Based Approaches. *Cold Spring Harbor Perspectives in Medicine* **4**, a021188–a021188 (2014).
- [149] Guerra-Assunção, J. *et al.* Large-scale whole genome sequencing of *M. tuberculosis* provides insights into transmission in a high prevalence area. *eLife* **4**, 1–17 (2015).
- [150] Nica, A. C. & Dermitzakis, E. T. Expression quantitative trait loci: present and future. *Philosophical Transactions of the Royal Society B: Biological Sciences* **368**, 20120362 (2013).
- [151] Jayakumar, V. & Sakakibara, Y. Comprehensive evaluation of non-hybrid genome assembly tools for third-generation PacBio long-read sequence data. *Briefings in Bioinformatics* **20**, 866–876 (2019).
- [152] Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long, error-prone reads using repeat graphs. *Nature Biotechnology* **37**, 540–546 (2019).
- [153] Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
- [154] Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
- [155] Anders, S., Pyl, P. T. & Huber, W. HTSeq - a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).
- [156] Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15**, 550 (2014).

CHAPTER 2

Objectives and Structure of the Thesis

2.1. Objectives

Through analysis of WGS data from different 'omics approaches, this thesis focuses on the investigation of the diversity observed in clinical isolates of *Mtb* to improve our understanding of the pathogen biology and inform in aspects such as pathogenicity or drug resistance. The questions addressed in this thesis include:

- (i) the role of genetic and DNA methylation diversity on regulation of gene expression **(Chapter 3)**;
- (ii) the frequency and distribution of drug resistance associated mutations to bedaquiline, delamanid and pretomanid in a large set of clinical isolates **(Chapter 4)**;
- (iii) characterisation and investigation of the diversity in the *pe* and *ppe* gene families across different lineages by using long-read sequencing data **(Chapter 5)**;
- (iv) and the application of cost-effective sequencing technologies for epidemiological and drug resistance detection investigations **(Chapter 6)**.

For the completion of this work, well characterised *Mtb* clinical isolates from the Karonga Prevention Study were cultured and DNA/RNA extracted in the Biosafety Level 3 containment facilities at LSHTM. The sequencing was outsourced through The Applied Genomics Centre and involved Illumina HiSeq4000, PacBio and Oxford Nanopore Technologies (ONT) platforms. The generated and collected data, together with publicly available sequences, was analysed using a range of bioinformatic tools and resources.

2.2. Structure of the Thesis

The thesis is divided in four chapters corresponding to individual manuscripts (2 published, 2 submitted). The research papers and manuscripts included in this thesis are the following:

Chapter	Title	Status, journal and year of publication
2	An integrated whole genome analysis of <i>Mycobacterium tuberculosis</i> reveals insights into relationship between its genome, transcriptome and methylome	Published; Scientific Reports 2019
3	Genetic diversity of candidate loci linked to <i>Mycobacterium tuberculosis</i> resistance to bedaquiline, delamanid and pretomanid	Published; Scientific Reports 2021
4	Functional genetic variation in <i>pe/ppe</i> genes contributes to diversity in <i>Mycobacterium tuberculosis</i> lineages and potential interactions with the human host	Submitted; Genome Biology
5	Portable sequencing of <i>Mycobacterium tuberculosis</i> for clinical and epidemiological applications	Submitted; Briefings in Bioinformatics

The role of DNA methylation in transcription has been described in bacteria, including few studies carried out in *Mtb*. Changes in gene expression affect the bacterial phenotype, and therefore are likely to have clinical implications. On this premise, the understanding of the different existing methylation patterns among lineages and their consequences is important.

Chapter 3 presents a joint study of the genome, transcriptome and methylation profiles of three of the major lineages of *Mtb* to interrogate the role of genetic variants and modification patterns in the regulation of gene expression at a genome-wide scale. For this purpose, PacBio long-read sequencing and Illumina RNA-seq data are analysed to obtain variants, methylated motifs and gene expression levels. Through statistical associations established by expression quantitative trait loci studies (eQTLs), this analysis aims to provide candidate variants and methylated sites potentially involved in changes in expression.

The availability of WGS data from different strains, collection times and geographical locations enables the performance of large-scale analysis. In **Chapter 4**, a collection of WGS from > 30k *Mtb* isolates is used for the study of 9 drug resistance associated candidate loci to the new anti-TB drugs bedaquiline, delamanid and pretomanid. With reports of resistant strains to these drugs soon after their roll-out, there are increasing concerns on the rapid acquisition of resistant mutations or the existence of intrinsic conferring-resistance variants leading to treatment failure. The lack of drug susceptibility testing (DST) for the new drugs limits the sample sizes for association studies and discovery of new mechanisms of resistance. **Chapter 4** describes a comprehensive analysis of the frequency and distribution of variants in candidate loci by applying phylogenetic methods and using the phenotypic information available in the literature.

Some long-read sequencing technologies have high error rates, however, their application can provide better resolution of complex regions with high GC content and repetitive sequences, as well as the base modifications mentioned earlier. These regions include the *pe* and *ppe* gene families. **Chapter 5** describes an analysis looking at the organisation and diversity of the 169 *pe/pppe* genes using >70 high quality PacBio genomes representing different lineages. For improved resolution, hybrid assembly approaches that combine PacBio and Illumina data are used, and population genomics methods are applied to inform on the conservation across the two gene families, and ultimately, improve the knowledge of these immunogenic proteins, often targeted as vaccine candidates.

Although WGS platforms have been implemented for the standard diagnosis of resistant-TB in countries like the UK, their high cost limits its accessibility in, for example, high burden TB settings. Nevertheless, these economic and infrastructure restraints can be overcome by cost-

effective and portable platforms, such as the ONT MinION sequencer. Based on recent reports that have suggested the use of MinION for detection of drug resistance mutations, **Chapter 6** assesses its application for epidemiological analysis and *in silico* drug resistance prediction. MinION performance is compared to the gold standard Illumina platform. Moreover, the ability to identify variants in *pe/ppa* genes (analysed in **Chapter 5**), loci typically excluded from analysis, is assessed, with the impact of additional characterised variants on phylogenetically resolution evaluated. **Chapter 7** contains the thesis discussion and conclusions.

CHAPTER 3

An integrated whole-genome analysis of
Mycobacterium tuberculosis reveals
insights into relationship between its
genome, transcriptome and methylome

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	lsh1704009	Title	
First Name(s)	Paula Josefina		
Surname/Family Name	Gómez González		
Thesis Title	Analysis of Mycobacterium tuberculosis 'omics data to inform on loci linked to drug resistance, pathogenicity and virulence		
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?	2019		
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

<p>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)</p>	<p>I received the raw sequence data from collaborators. I designed and ran the analysis pipeline, consisting in mapping, variant calling, modification analysis through the SMRT portal and read count for RNA-seq. I performed the statistical analysis and plotting with custom scripts. I wrote the first draft of the manuscript and circulated to co-authors. After several iterations of including comments, I submitted the manuscript to Scientific Reports and dealt with any subsequent revisions.</p>
---	--

SECTION E

Student Signature	
Date	January 28, 2022

Supervisor Signature	
Date	January 28, 2022

SCIENTIFIC REPORTS

OPEN

An integrated whole genome analysis of *Mycobacterium tuberculosis* reveals insights into relationship between its genome, transcriptome and methylome

Paula J. Gomez-Gonzalez¹, Nuria Andreu¹, Jody E. Phelan¹, Paola Florez de Sessions², Judith R. Glynn³, Amelia C. Crampin^{3,4}, Susana Campino¹, Philip D. Butcher⁵, Martin L. Hibberd^{1,2} & Taane G. Clark^{1,3}

Human tuberculosis disease (TB), caused by *Mycobacterium tuberculosis* (*Mtb*), is a complex disease, with a spectrum of outcomes. Genomic, transcriptomic and methylation studies have revealed differences between *Mtb* lineages, likely to impact on transmission, virulence and drug resistance. However, so far no studies have integrated sequence-based genomic, transcriptomic and methylation characterisation across a common set of samples, which is critical to understand how DNA sequence and methylation affect RNA expression and, ultimately, *Mtb* pathogenesis. Here we perform such an integrated analysis across 22 *M. tuberculosis* clinical isolates, representing ancient (lineage 1) and modern (lineages 2 and 4) strains. The results confirm the presence of lineage-specific differential gene expression, linked to specific SNP-based expression quantitative trait loci: with 10 eQTLs involving SNPs in promoter regions or transcriptional start sites; and 12 involving potential functional impairment of transcriptional regulators. Methylation status was also found to have a role in transcription, with evidence of differential expression in 50 genes across lineage 4 samples. Lack of methylation was associated with three novel variants in *mamA*, likely to cause loss of function of this enzyme. Overall, our work shows the relationship of DNA sequence and methylation to RNA expression, and differences between ancient and modern lineages. Further studies are needed to verify the functional consequences of the identified mechanisms of gene expression regulation.

Human tuberculosis disease (TB), caused by *Mycobacterium tuberculosis* (*Mtb*), is a major global public health issue¹. A deeper understanding of the biology of *Mtb* should reveal new insights that may help to improve diagnostics, treatments, vaccines and other much needed control measures. *Mtb* belongs to the *M. tuberculosis* complex (MTC), which consists of seven main lineages classified into modern (lineages 2–4), ancient (lineages 1, 5 and 6), and intermediate (lineage 7) strains². The lineages vary in their geographic distribution and spread, with lineage 2 being particularly mobile with evidence of recent spread from Asia to Europe and Africa. Lineage 4 is common in Europe and southern Africa, coinciding with regions of high TB incidence and high levels of HIV co-infection. The lineages may vary in their propensity to transmit and to cause disease, and in the site and severity of disease^{3–5}. A set of SNPs in the *Mtb* genome (size 4.4 Mb) has been identified that can be used to barcode sub-lineages⁶, leading to informatic tools that position sequenced samples within a global phylogeny⁷.

¹Faculty of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine, London, United Kingdom. ²Genome Institute of Singapore, Biopolis, Singapore. ³Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London, United Kingdom. ⁴Malawi Epidemiology and Intervention Research Unit, Lilongwe, Malawi. ⁵Institute for Infection & Immunity, St George's University of London, London, United Kingdom. Martin L. Hibberd and Taane G. Clark jointly supervised this work. Correspondence and requests for materials should be addressed to T.G.C. (email: taane.clark@lshtm.ac.uk)

Genetic diversity, accessible through whole genome sequencing, plays an important role also in transcription. Gene expression differences have been observed, with 15% of the genes found to be differentially expressed among different *Mtb* clinical isolates⁸, and lineage-specific transcriptome differences have been observed *in vitro* and during survival in macrophages^{9,10}. The mechanisms controlling expression of candidate genes, such as the upregulation of the *dosR* operon specific to Beijing strains, have been broadly investigated^{11–13}. However, little is known about the effect of genomic variation on transcription at a whole genome scale. These effects can be explored through an association analysis of polymorphisms, such as single nucleotide polymorphisms (SNPs), and gene expression levels to determine expression quantitative trait loci (eQTL). eQTLs are genetic variants that explain variation in gene expression levels, and can be classified as *cis* or *trans* depending on the physical distance from the gene they regulate¹⁴. In *Mtb*, one previous study focusing on lineage 1 and 2 strains, highlighted two types of mechanisms where polymorphisms may change gene expression: through impairment of transcriptional regulators or by affecting the promoter regions¹⁰.

In addition to genomic variants, epigenetic mechanisms such as DNA methylation have an effect on gene expression. Several lines of evidence have revealed N6-methyladenine (m6A) and 5-methylcytosine (m5C) methylation mechanisms within *Mtb* genomes, and these can be characterised using single-molecule real time (SMRT) sequencing from Pacific Biosciences technology^{15,16}. Motifs within three DNA methyltransferases (MTases), *mamA*, *mamB*, and *hdsM* are responsible for m6A modification^{15–17}. In *Mtb* it has been shown that the loss of *mamA* MTase can decrease gene expression and affect survival during hypoxia¹⁷. Methylation sites have been found to overlap with sigma factor binding sites, suggesting that if methylation affects sigma factor binding, methylation status may play a role in transcription¹⁷. Lineage-specific methylation patterns have been reported for *Mtb* strains¹⁶, which indicates the potential for novel functional differences between them. In eukaryotic cells, DNA methylation is often associated with repression of gene expression; however, in prokaryotes, methylation has been associated with both induction and repression of gene expression^{17,18}.

To date, no studies have integrated sequence-based genomic, transcriptomic and methylation characterisation across a common set of samples. This integration is critical to understand how DNA sequence and methylation affect RNA expression and, ultimately, *Mtb* pathogenesis. Here we seek to investigate the relationship between the genome, transcriptome and methylome in a panel of 22 *Mtb* isolates, belonging to the Karonga Prevention Study, a longitudinal epidemiological project focused on mycobacterial disease¹⁹. We present a differential gene expression study correlated with lineage, as well as an eQTL study linked with SNPs and methylated bases at a whole genome scale. Differential transcription between lineages was found, and genetic variants revealed as potential candidate eQTLs. Methylation status was also found to have a potential role in transcription, with evidence of differential gene expression between samples with non-methylated and methylated genes.

Results

Genomic analysis. *Mtb* was isolated from 22 sputum samples from 22 different TB patients collected between 2003 and 2009 in Karonga, a northern district of Malawi. The majority of individuals were HIV positive (16/22). Genomic DNA was extracted and sequenced using PacBio single-molecule real time (SMRT) and Illumina sequencing technologies. One ancient (L1, n = 8) and two modern lineages (L2 and L4, n = 14) were represented (Supplementary Table S1). For each isolate, the raw sequence data was aligned to the H37Rv reference genome, leading to >100-fold average coverage. Across all samples 9,384 unique SNPs were characterised, with ~40% of them identified in single isolates. Only 1,446 of the 9,384 SNPs were located in intergenic regions. The average number of SNPs per isolate varied by lineage (L1: 2,613; L2: 1,675; L4: 1,101); the sub-lineage 4.9 (H37Rv-like) was the least polymorphic (~600 variants). Using the 9,384 SNPs, a maximum-likelihood phylogenetic tree was constructed (Fig. 1) and the isolates clustered by lineage as expected.

Transcriptomic analysis and lineage-specific expression. *Mtb* RNA was extracted from the 22 clinical isolates following liquid culture at mid-log phase growth and sequenced using Illumina HiSeq technology. Short reads were aligned to the H37Rv reference genome and counts per gene were obtained. A total of 3,987 genes were transcribed in at least two clinical isolates with a minimum of 10 counts. The average number of transcripts in the sample set is 3,864. A differential expression test was performed by clade, between the ancient (L1; n = 8) and the modern (L2 and L4; n = 14) strains in our sample set (Supplementary Fig. S1A). At a significance level of $p < 1.24 \times 10^{-5}$ (corresponding to a Bonferroni adjusted $p < 0.05$), 105 genes were revealed as differentially expressed (Fig. 2, Supplementary Table S2). Five of them (*Rv1524-wbbL2*, *Rv2652c-Rv2653c-Rv2658c*) correspond to known deletions in ancient isolates. *PE_PGRS57* was also absent in ancient genomes of our samples, which has also been observed to be deleted in other ancient (L5; *M. africanum*) strains in other studies^{20,21}. As expected, *Rv1524-wbbL2*, *Rv2652c-Rv2653c-Rv2658c* and *PE_PGRS57* transcripts were down-regulated in ancient strains. Forty-eight of the 105 (45.7%) genes found to be differentially expressed by clade have been reported in previous transcriptomic analyses performed between ancient and modern strains or L1 and L2^{9,10}, leading to 57 newly described genes here. The main functional ontological categories for the 105 identified genes were conserved hypotheticals and intermediary metabolism and respiration. Enrichment in nitrogen metabolism ($p = 2.75 \times 10^{-5}$) and PE-PGRS ($p = 7.2 \times 10^{-3}$) associated genes was found. Within clade-specific patterns, genes associated with transcriptional regulation were also identified. For ancient strains, *Rv0273c*, *Rv0275c*, and *Rv2160A* were the most under-expressed, whilst *pknH*, *Rv2282c*, *virS*, and *Rv3167c*, were over-expressed. In addition, several of the 105 differentially expressed genes were associated with virulence. Three of them belonged to the *vapBC* toxin-antitoxin system (*vapB10*, *vapC10*, *vapB22*), which were up- or down-regulated in ancient strains. Also, the *mce4A* gene, involved in cholesterol uptake during macrophage survival and associated with long term persistence²², and *yrbE4B*, forming part of the *mce4* operon, were found over-expressed in ancient isolates. Finally, genes associated with drug resistance, such as the efflux pump *Rv2994* and the isoniazid related *iniA* and *iniB* genes, were revealed as differentially expressed between the ancient and modern lineages studied.

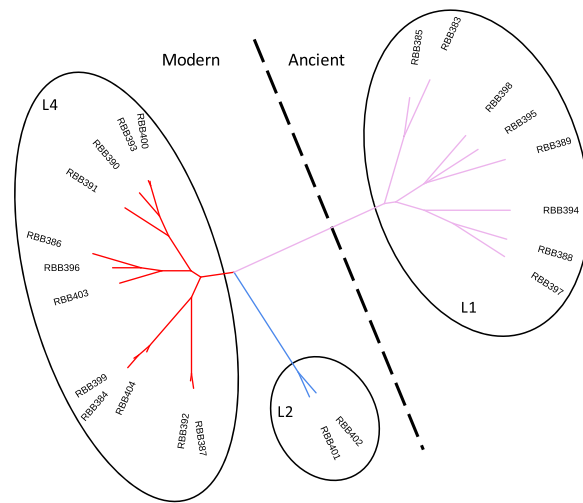


Figure 1. Phylogenetic tree of the 22 Karonga strains. Maximum-likelihood phylogenetic tree of the 22 isolates analysed, covering lineages 1 (L1), 2 (L2) and 4 (L4).

Rv2994 has found to be over-expressed in multi-drug resistant isolates²³, and the *iniA* and *iniB* genes are related with higher persistence under isoniazid conditions^{24,25}.

Identification of Expression Quantitative Trait Loci (eQTL). An eQTL analysis was performed at a whole genome scale across the 22 isolates, and we attempted to associate SNP alleles with differential transcription signal. Association testing was performed between 9,384 SNPs and 3,987 transcripts using a linear regression modelling approach (Supplementary Fig. S1B). We identified potential eQTLs from the 38,949 significant associations between 5,608 SNP positions and 118 differential transcribed genes ($p < 1.32 \times 10^{-9}$; adjusted $p < 0.05$). The 5,608 SNPs considered as eQTLs were located in 2,279 genes and intergenic regions. Forty-two of the 118 (35.6%) genes were differentially expressed due to large deletions and were subsequently excluded from further analysis (Supplementary Table S3), leaving 76 genes as potentially affected by SNP eQTLs (Supplementary Table S4). More than half of these 76 genes had a lineage or sub-lineage-specific expression profile. Moreover, a large number of the eQTLs associations were due to both lineage-specific SNPs and expressed genes. Thereby, a group of 790 common SNPs across all ancient isolates was associated with the expression of 24 genes; a group of 169 SNPs present in all L1 and L2 isolates was associated with the expression of 9 genes, and 584 SNPs present in Beijing (L2) isolates were associated with the expression of 3 genes (Supplementary Table S4). To assign the most likely causative genetic variation of the eQTLs, we investigated SNPs with a potential *cis* regulatory function and those within transcriptional regulatory proteins.

***Cis*-regulatory eQTLs.** A *cis*-eQTL analysis was performed at SNPs, within each gene or < 200 bp upstream from their start codon, tested for differential expression (Supplementary Fig. S1C). This analysis identified 99 potential *cis*-eQTLs associated with the differential expression of 83 genes ($p < 4.04 \times 10^{-6}$, adjusted $p < 0.05$), involving 92 SNPs (Supplementary Table S5). The majority (65/92) of these candidate *cis*-eQTL SNPs were located within the gene, 15 were located in the upstream intergenic region and 8 within the upstream gene. Among those in the upstream intergenic region, 8 were in predicted promoter regions. Eleven upstream SNPs (11/15) were common (allele frequency > 5%) in a global set of strains ($n = 6,218$)²⁶. Also, 6 SNPs within the upstream gene (6/8) were common (Table 1). Among them, the antitoxin *vapB22*, is known to be over-expressed in ancient isolates when compared to modern strains, and was found to harbour a SNP in its promoter (T3137237C) in all ancient isolates, thereby providing a possible explanation for the change in expression. Further, all the SNPs identified as potential *cis*-eQTLs were aligned to a map of transcriptional start sites (TSS)²⁷. We found that three were located within the TSS of three genes shown to be differentially expressed in L1 compared to modern strains, with *PE_PGRS38* (A2424864G) and *fadD31* (T2177073C) under-expressed, and *virS* (A3447480C) over-expressed in ancient isolates. Overall, five SNPs present in ancient strains identified in this study as potential *cis*-eQTLs have already been reported as potentially associated with variation in gene transcription¹⁰, giving us confidence in our approach.

Transcriptional regulatory proteins. We next considered candidate SNP eQTLs with non-synonymous mutations in transcriptional regulatory proteins (Supplementary Fig. S1D). These mutations could affect the DNA binding function of the protein. In total, 46 SNPs in 38 different transcriptional regulatory proteins (Table 2) were associated in the eQTL analysis with the differential transcription of 56 genes, accounting for a total of 376 potential eQTL associations. Ten of these 46 SNPs have been previously reported as having a potential effect in transcriptional regulation¹⁰. Functional effects were investigated through the SIFT algorithm, and 16 of the 38 (42.1%) transcriptional regulators were predicted to have SNP mutations affecting functional impairment. For

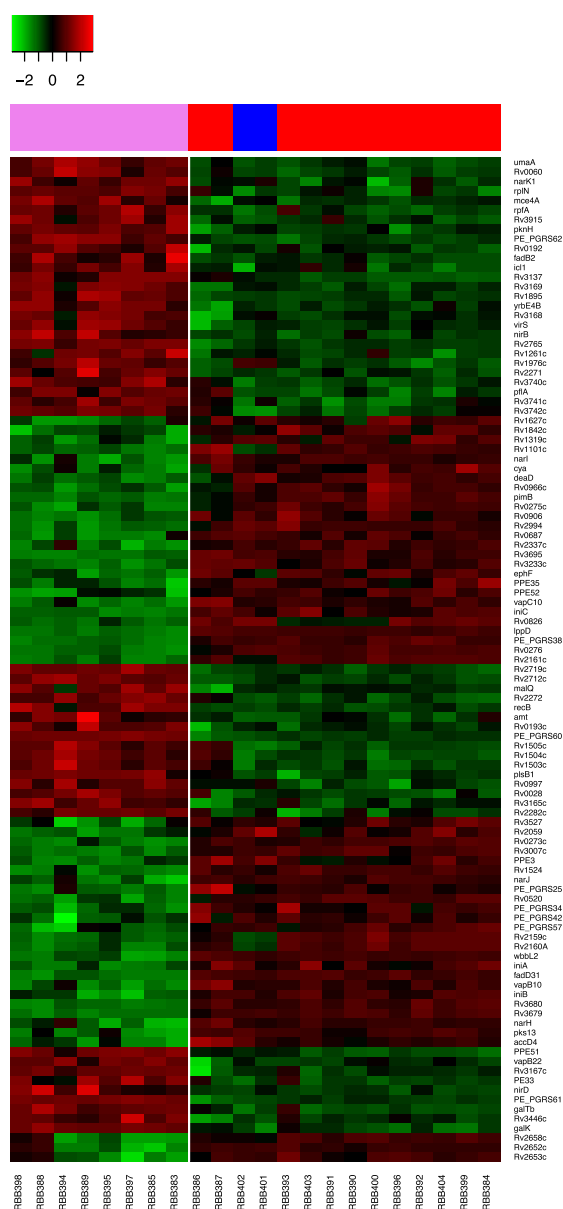


Figure 2. Gene expression differences between modern (lineage 2 and 4) and ancient (lineage 1) strains. A heatmap showing the 105 genes differentially expressed between ancient and modern strains, constructed with the gene expression distances between rows. Rows and columns are ordered based on row or column means. Over-expressed genes are coloured in red whilst under-expressed ones in green. Ancient strains (n = 8) represented on the left of the white vertical line and modern strains (n = 14) on the right. Lineage 1 represented in violet, Lineage 2 in blue and Lineage 4 in red.

the majority of the regulatory genes (20/38; 52.6%), the SIFT software did not predict a functional consequence of the mutations, due to the lack of homology with sequences in its database.

Mutations in the *sirR* and *Rv0195* genes resulted in stop codons and led to truncated proteins. The stop codon in *sirR*, a manganese-dependent transcriptional repressor²⁸, was observed in all L1 samples. While, mutations in *Rv0195*, a LuxR family regulatory gene, were observed in one L1 sample. Some of the 38 transcriptional regulators belonged to other known regulatory families such as TetR. The TetR family of transcriptional regulators (TFTRs) are one-component prokaryotic signal transduction systems controlling different biochemical functions. Although they were thought to be expression repressors, work in other bacteria has shown that they can act also as activators²⁹. The TFTR *Rv2160A* carried a SNP (C155R) and an insertion (304insGGAA) causing a change in the

	Transcript differentially expressed	Annotation	SNP	Position SNP			Regulation	Strain Lineage	Allele frequency**	
				Gene	Distance (bp) from start codon	Promoter (P)/TSS			Ancient	Modern
SNPs in upstream region	<i>Rv0193c</i>	1	G226676A	IGR	-105	—	Up	1	0.973	0
	<i>Rv0326</i>	—	T392261C	<i>Rv0325</i>	-12	—	Up	1,2	0.978	0.324
	<i>Rv0377</i>	6	T454295C	<i>Rv0376c</i>	-126	—	Up	1,2,4.1,4.3,4, 4.8,4.9	1	0.994
	<i>gpdA1</i>	4	T655986G	IGR	-37	P	Up	1,2	0.976	0.324
	<i>mce2D</i>	6	A690450C	<i>mce2C</i>	-51	—	Up	1,2	0.976	0.324
	<i>Rv0669c</i>	3	T769663G	IGR	-66	P	Down	4.3,3	0	0.050
	<i>Rv0958</i>	3	C1069871T	IGR	-12	P	Up	1.1,3	0.220	0
	<i>Rv1096</i>	3	T1224367C	IGR	-18	P	Down	1,2,4.1,4.3,4,8	1	0.976
	<i>Rv1503c</i>	1	A1694547C	IGR	-3	—	Up	1	0.973	0
	<i>fadD31</i>	4	T2177073C	IGR	-14	TSS/P	Down	1	0.973	0
	<i>Rv2036</i>	3	C2282058T	<i>Rv2035</i>	-41	—	Up	1.2,2*	0.157	0
	<i>Rv2159c</i>	1	A2421816G	<i>Rv2160A</i>	-151	—	Down	1,2	0.977	0.323
	<i>PE_PGRS38</i>	7	A2424864G	IGR	-18	TSS	Down	1	0.973	0
	<i>Rv2712c</i>	1	C3025431T	IGR	-103	P	Up	1	0.971	0
	<i>vapB22</i>	5	T3137237C	IGR	-13	P	Up	1	0.973	0
	<i>yrbE4B</i>	5	G3920109T	<i>yrbE4A</i>	-47	—	Up	1	0.971	0
	<i>Rv3695</i>	2	T4137190C	IGR	-16	—	Down	1	0.973	0

Table 1. Putative functional SNPs associated with expression (*cis*-eQTLs with allele frequencies >5%; adjusted $p < 0.05$). Table showing the candidate transcripts differentially expressed due to SNPs in upstream intergenic regions (IGRs) or within the upstream gene. Annotation of the transcript differentially expressed: 1 – Conserved hypotheticals, 2 – Cell wall and cell processes, 3 – Intermediary metabolism and respiration, 4 – Lipid metabolism, 5 – Virulence, detoxification, adaptation, 6 – Regulatory proteins, 7 – PE/PPE, 8 – information pathways. Distance of the SNP location from the start codon of the transcript is shown as negative when it is upstream and positive when it is located within the gene. TSS = Transcriptional Start Site. *Only one or two samples from the lineage out of the 3 analysed. **Allele frequency refers to the fraction of strains harbouring the SNP in a larger data set ($n = 6,218$)³⁰; “—” when not available.

reading frame in isolates from L1 and L2. *Rv2160A* is likely to form part of the operon *Rv2159c/Rv2160A/Rv2161c*. In our analysis, *Rv2159c* and *Rv2161c* were revealed as highly down-regulated in ancient strains compared to modern ones, and marginally down-regulated in L2 compared to L4 isolates. These observations suggest the operon may act as an activator, and that the mutations may lead to a loss of its function.

In *Streptomyces* it has been shown that TFTRs can regulate divergently oriented neighbouring genes³⁰, and previous studies in *Mtb*^{10,31} have found differential expression of genes adjacent to TFTRs. We looked for similar effects in *Mtb* TFTRs carrying potential eQTLs. *Rv0275c* is a potential regulator of its divergent oriented neighbouring gene *Rv0276*. The ancient strains carried a mutation (S24L) in *Rv0275c*, which was associated with the under-expression of *Rv0276*. Similarly, *Rv3167c* is a potential regulator of its divergent oriented neighbour gene *Rv3168*. Although, the ancient strains carried a mutation (P17Q) in *Rv3167c*, and *Rv3168* appeared slightly over-expressed, this effect did not reach the stringent significance cut off imposed in the eQTL analysis.

In order to study the consequential effects of mutations in the transcriptional regulators of the genes found as being differentially expressed, network gene regulation was analysed through the Environment and Gene Regulatory Influence Network (EGRIN) model from the MTB Network Portal³² and the regulatory network map from the TB database³³. We compared the predicted induced and repressed genes by the transcriptional regulators harbouring non-synonymous SNPs with the differentially expressed genes in our samples. This analysis revealed the association of genes differentially expressed with five of our candidate transcriptional regulators (Supplementary Table S6). *Rv0275c*, which is predicted to auto-induce its expression, was found to be down-regulated in ancient strains (with S24L mutation), although this effect did not reach the statistical significance cut-off. In addition to the under-expression of *Rv0276*, discussed above, three other genes (*Rv0520*, *Rv2162c* and *Rv0826*) were found to be under-expressed in ancient strains and are predicted to be regulated by *Rv0275c*. Genes regulated by *ramB*, were up- or down-regulated in ancient strains carrying *ramB* P91Q and Q121R mutations. Other genes were regulated by the transcriptional regulators *Rv1776c*, *Rv3167c* and *Rv3249c*, which harboured potential impairment mutations, leading to under- or over-expression in those isolates carrying the mutations. For the remaining regulators within known control networks, no statistically significant associations of variable gene expression with mutations were found.

Sigma and anti-sigma factors are critical to the gene expression regulatory network³⁴, and here we hypothesised that polymorphisms in these factors might affect the transcription of those genes regulated by them. We found three anti-sigma factors (*rseA*, *rskA* and *rsfA*) harbouring non-synonymous SNPs that were considered as potential eQTLs (adjusted $p < 0.05$) associated with six genes differentially expressed between the isolates carrying and not carrying the mutations (Supplementary Table S7).

Gene	Mutation	Family	Lineage of strains carrying mutation	Allele frequency	
				Ancient	Modern
<i>whiB5</i>	S21G	whiB	1.2.2**	0.021	0
<i>Rv0023</i>	G217D		4.9**	0	0.001
<i>Rv0042c</i>	L186R*	MarR	4.9**	0	0
<i>Rv0144</i>	P36L*	tetR	4.9**	0	0
<i>Rv0195</i>	C41STOP	LuxR	1.2.2**	0.021	0
<i>Rv0275c</i>	S24L	tetR	1	0.973	0
<i>iniR</i>	E23K		1.2.2**	0.019	0
<i>Rv0377</i>	P302R*	LysR	1	0.973	0
<i>Rv0386</i>	L475R*	LuxR/UhpA	4.1.1.3	0	0.003
<i>ramB</i>	P91Q		1	0.973	0
	T118A		4.9**	0	0.001
	Q121R		1	0.973	0
<i>Rv0576</i>	R233H*	ArsR	1,2	0.978	0.334
<i>Rv0691c</i>	A140T		2	0.003	0.114
<i>Rv0818</i>	P227L*		4.1.1.3	0	0.003
	E246K*		4.1.2	0	0.009
<i>narL</i>	G169R*		2	0.003	0.147
<i>Rv0890c</i>	E234G*	LuxR	2	0.003	0.111
	E303K*		4.1.2	0	0.009
<i>Rv0891c</i>	V37G*		1,2,4.1,4.3,4.8	1	0.974
<i>kdpE</i>	G60S*	KDPD/KDPE	2	0.003	0.111
<i>Rv1219c</i>	R11T		1.2.2**	0.148	0
<i>embr</i>	A70S		4.1.2	0	0.009
	C110Y		1	0.973	0
<i>Rv1453</i>	D208N		1.1.3	0.230	0
	D218N		1.2.2**	0.021	0
	P405Q		1,2,4.1,4.3,4.8	1	0.974
<i>Rv1674c</i>	E189G*		4.3	0.014	0.281
<i>cmr</i>	V59A	CRP/FNR	1	0.974	0
	A125S		1.1.3*	0.072	0
<i>Rv1776c</i>	R154S		1.2.2**	0.019	0
<i>blaI</i>	L57R		1	0.970	0
<i>mce3R</i>	D148Y*	tetR	1.1.3**	—	—
<i>Rv2017</i>	A262E		1,2,4.1,4.3,4.8	0.998	0.973
<i>Rv2160A</i>	C155R	tetR	1,2	0.977	0.323
<i>zur</i>	H64R*		1	0.973	0
<i>Rv2488c</i>	D184Y*	LuxR	1.2.2**	0.018	0
<i>Rv2621c</i>	A110V		2	0.003	0.148
<i>sirR</i>	Q131STOP		1	0.973	0
<i>Rv3060c</i>	G420D	GntR	4.1.2	0	0.009
<i>virS</i>	L316R*	AraC/XylS	1	0.973	0
<i>Rv3167c</i>	P17Q	tetR	1	0.973	0
<i>Rv3249c</i>	T154A	tetR	4.1.1.3	0.003	0.049
<i>whiB4</i>	S2L	whiB	1.1.3	0.223	0
<i>Rv3736</i>	G144R*	AraC/XylS	1	0.971	0
<i>whiB6</i>	G71D	whiB	1.2.2**	0.014	0

Table 2. Non-synonymous variants in transcriptional regulatory genes with eQTL associations, with potential functional impairment. Table showing non-synonymous mutations in transcriptional regulatory genes found as potential eQTLs. *Sorting Intolerant from tolerant (SIFT) predicted scores (p value) < 0.05 and considered to have functional impact; whilst for the others the SIFT software was unable to predict functional effects of mutations; **Only one or two samples available from the lineage. Allele frequency refers to the fraction of strains harbouring the SNP in a larger data set ($n = 6,218$)²⁶.

Methylation analysis. Motif and methylation finding was performed through the Modification and Motif Analysis pipeline provided by the SMRT portal (<https://github.com/PacificBiosciences/SMRT-Analysis>). By analysing the kinetic variation through the inter-pulse duration ratio (IPD) at each nucleotide in the genome, a large

Gene	Position	strand	Motif	Distance from start codon (bp)	Promoter/TSS	Regulation in non-methylated samples
<i>Rv0565c</i>	657533	–	CTGGAG	–63	–	Down
<i>ompA</i>	1002711	+	CTCCAG	–101	–	Down
<i>Rv1371</i>	1543277	+	CTCCAG	–82	–	Up
<i>scpB</i>	1938088	+	CTCCAG	–58	P, TSS	Up
<i>moaC3</i>	3710411	–	CTCCAG	–163	–	Up
<i>Rv3324A</i>	3710411	–	CTCCAG	–32	–	Up
<i>Rv3325</i>	3710408	+	CTGGAG	–25	–	Down
<i>PE_PGRS60</i>	4093563	+	CTGGAG	–69	–	Down

Table 3. *cis*-eQTLs located in upstream intergenic regions linked with methylation in Lineage 4 strains. Table showing genes differentially expressed potentially due to the lack of methylation in the upstream region. The name of the gene, the position of the eQTL (methylation site), strand, motif, distance of the methylated base from start codon of the transcript (negative shown as upstream), prediction of promoter or TSS (P = promoter region, TSS = Transcriptional Start Site), and type of regulation of the gene in non-methylated samples is shown.

number of modifications were identified. Only high quality 6-methyl-adenine (m6A) levels were found within motifs, where m6A is a well characterised epigenetic regulator in other prokaryotes^{35,36}. The three motifs previously reported in *Mtb*^{15–17} were identified: CTCCAG and GATN₄RTAC and their partner motifs (CTGGAG and GTAYN₄ATC, respectively), and the hemi-methylated CACGCAG. The distribution and numbers of the different motifs were similar across the samples regardless of lineage and sub-lineage, with an average number of 1,934 for CTCCAG, 357 for GATN₄RTAC and 813 for CACGCAG. However, the fraction of methylated motifs varied across isolates and (sub-)lineage patterns (Supplementary Table S8), consistent with a previous report¹⁶. In particular, within L4, two sub-lineages patterns were found with methylation in GATN₄RTAC and CACGCAG motifs. Moreover, the CTCCAG motif was not methylated in either of the two L2 isolates. Among L1, methylation in CTCCAG and CACGCAG motifs was absent in some samples. When methylated, the percentage of motifs modified across all the samples varied from 50% to ~100%.

To explain the lack of methylation observed in some isolates, the presence of SNPs in the MTases genes was investigated. Three SNP mutations were identified: (i) E270A in *mamA* in L2, (ii) P306L in *hdsM* in sub-lineages 4.3, 4.8 and 4.9, and (iii) S253L in *mamB* in sub-lineage 1.1.3; which have been reported previously to be associated with the loss of function of the enzymes^{15,16} (Supplementary Table S9). Two novel mutations (Q340K and G152S) and a deletion (1232delG) were also identified in *mamA*, potentially associated with the lack of methylation of CTCCAG in two isolates belonging to L1 and L4. For the remaining samples with an absence of methylation in any of the three motifs, there were no SNPs uniquely found in these samples that could be correlated with the loss of function of the enzyme.

Differential gene expression linked with methylation. In order to understand how the methylation status of the genes affects their expression, a differential transcription analysis was performed on the L1 and L4 strains (n = 20) (Supplementary Fig. S1E). The analysis involved stratifying by lineage to overcome the lineage-specific transcriptional profiles seen above. L2 was discarded due to the low number of clinical isolates represented. Firstly, 5,326 different intragenic methylation sites were used. A linear regression analysis was applied to obtain the correlation between methylation status and gene expression level at a whole-genome scale. Across L4, 44 genes were found to be differentially expressed (Benjamini-Hochberg (BH) adjusted $p < 0.05$), whose over- or under-expression was potentially associated with their methylation status. Twenty-eight (of the 44; 63.6%) genes, mostly down-regulated, were deficient in methylation only in the CTCCAG motif in one sample, which was associated with the presence of the mutation G152S in *mamA* (Supplementary Fig. S2). These genes were enriched for metabolic pathways ($p < 0.05$). The remaining 16 genes differentially expressed in L4 were non-methylated in >1 isolate and mostly in the CTCCAG motif (Supplementary Fig. S3). For L1, none of the genes that were found to be differentially expressed were significantly associated with methylation status. Methylation of the upstream intergenic regions may have a role in gene expression, and we performed a lineage-stratified *cis*-eQTL analysis with the 393 unique methylation sites located within 200bp upstream from the start codons of the genes. In L4, seven eQTLs (BH adjusted $p < 0.05$) for 8 genes differentially expressed were revealed (Table 3, Supplementary Fig. S4), including one located in the predicted promoter region and overlapping with the TSS. Among ancient strains, none of the genes that were found to be differentially expressed were significantly associated with methylation of upstream regions.

Overlap between eQTLs linked with SNPs and methylation. Finally, we assessed whether there is a link between the SNPs and methylated motifs associated with the differentially expressed genes identified. To this end, we evaluated the degree of overlap between the different associations (Fig. 3). We considered three types of association: (i) genes differentially expressed due to SNPs in promoter regions, TSS or within the gene, denoted as *cis*-eQTLs; (ii) genes differentially expressed due to potential impairing mutations in transcriptional regulators that are predicted to control their expression, denoted as *tr*-eQTLs; and (iii) genes differentially expressed as a consequence of methylation of either the promoter, TSS, upstream region or the gene, denoted as *mod*-eQTLs. We found that 5 genes with variable transcription were associated with both, *mod*-eQTLs and *cis*-eQTLs, and another

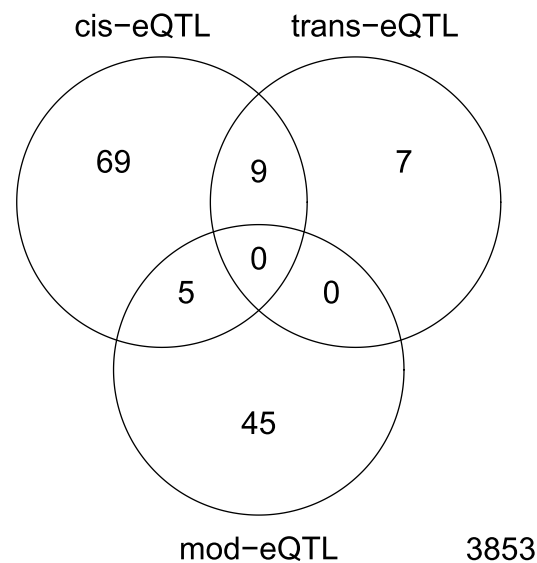


Figure 3. Venn diagram showing the overlap of genes differentially expressed (from the 3,987 investigated) associated with the different eQTL types (*cis*, *trans* and modified). The numbers represent the number of genes differentially expressed associated with the different types of eQTLs: *cis*-eQTLs, SNPs in promoter regions, transcriptional start sites (TSS), upstream (up to -200 bp) or within the gene; *tr*-eQTLs, potentially impairing non-synonymous SNPs located in transcriptional regulators; and *mod*-eQTLs, methylated bases located either within the gene or upstream including promoter regions and TSS.

9 were associated with *cis*-eQTLs and *tr*-eQTLs. There was no overlap between genes differentially expressed due to *tr*-eQTLs and *mod*-eQTLs, and the majority of the genes were uniquely assigned to one of the mechanisms responsible for their differential expression.

Discussion

Genetic mutations and variations in gene expression have an important impact on MTC virulence and pathogenicity^{4,5}. Previous studies have shown how genomic variants or methylation can affect the level of gene expression^{9,10,17}, but have not shown how one analysis may influence another. In this study, for the first time, we performed an integrated analysis of the genome, methylome and transcriptome, across 3 major *Mtb* lineages. We have revealed clade-specific differences in the core transcriptomes between ancient and modern strains, as previously observed⁹, but in addition our analysis has revealed genes linked to virulence and pathogenicity (e.g. *vapBC* family), drug resistance and efflux pumps (e.g. *Rv2994*²³ or *iniA* and *iniB*^{24,25}). An eQTL analytical approach revealed 5,608 SNPs associated with differential gene expression (a total of 38,949 candidate eQTLs) and reinforced the lineage-specific genetic diversity and its effects on transcriptomes. To achieve improved resolution, *cis*-eQTLs based on regions upstream or within the genes differentially expressed were considered. This approach revealed ten SNPs within the promoter regions or TSS of genes differentially expressed, as well as others within coding regions of the genes, doubling the number of previously reported associations¹⁰. Among these variants, lineage-specific SNPs were associated with the genes differentially expressed, thereby revealing a potential explanation for the differential core transcription.

The high proportion of non-synonymous mutations present in coding regions in *Mtb* has been suggested to have a functional impact⁴, with consequences for transcription when found within transcriptional regulators¹⁰. In our study, functional impairment was predicted for sixteen of the transcriptional regulators found among the 38,949 potential eQTLs, including in *sirR* and *Rv0195* that contained premature stop codons. The number of regulators found is likely to be an under-estimate, as databases accessible to SIFT are incomplete, leading to no prediction for the vast majority of loci. Most of the potential impairing mutations were found to be lineage-specific. In particular, we identified a mutation and an insertion in L1 and L2 strains in *Rv2160A*, which act as a transcriptional activator of the adjacent genes *Rv2159c* and *Rv2161c*, with which it likely forms an operon²⁹. Similarly, the protein encoded by *Rv3167c* was predicted to function as a repressor of its contiguous gene *Rv3168*, over-expressed in ancient samples with the P17Q mutation. Whilst *Rv0275c* was shown as a candidate activator of the adjacent gene *Rv0276*, and under-expressed in the L1 strains with the S24L mutation, consistent with previously reported associations^{10,31}. The analysis of the regulatory networks of the transcriptional regulators was performed in order to look for *trans*-eQTLs, and found 11 of the genes differentially expressed from the primary eQTL analysis were regulated by one of the transcriptional regulators harbouring potential impairing mutations. Three mutations affecting the function of three anti-sigma factors (*rseA*, *rskA* and *rsfA*) were associated with the up-regulation of 6 genes. This result suggests that the functional impairment of sigma and anti-sigma factors can be the cause of variable gene expression.

Our study confirmed the same motifs and patterns of methylation as previously reported^{15,16} but in addition identified three novel variants (Q340K, 121delG and G152S) in *mamA*, which could explain the lack of methylation in the CTCCAG motif in the samples harbouring them. DNA methylation has been hypothesised to affect gene expression in bacteria³⁵, and the disruption of *mamA* in *Mtb* has been shown to result in altered gene expression¹⁷. In *E. coli* it has been suggested that an overrepresented motif in the genome is more likely to be involved in gene expression regulation mediated by methylation³⁷. Different hypotheses concerning the control of gene expression by *dam* MTase have been proposed, including regulation by motifs found in promoter³⁸ and coding regions³⁹. Further, it has been suggested that DNA methylation is a mechanism of switching regulatory states in phase variation systems³⁷. Across the three lineages studied here, CTCCAG was the most abundant motif and was predominantly found in coding regions. An investigation of the relationship between the methylation status and gene expression levels revealed that the CTCCAG motif has the highest impact. In L4, the differential expression of 38 genes was potentially associated with CTCCAG methylation status, compared to 4 and 2 genes associated with CACGCAG and GATN_nRTAC methylation, respectively. A subset of these genes (28/44), mostly down-regulated, were found to be uniquely non-methylated in the sample with the *mamA* G152S mutation. These included genes associated with metabolic pathways or regulatory proteins (e.g. *Rv0348*, *virS* or *Rv1359*), and from the *pe/ppe* families (e.g. *PE17*, *PPE17* or *PE_PGRS2*). We also found that non-methylated CTCCAG motifs in upstream regions and TSS have an effect on gene expression, which is consistent with previous work¹⁷. In L1 no genes significantly associated with methylation were found. Overall our results show that methylation in the promoter regions and coding regions is likely to be involved in gene expression, with the CTCCAG motif as the main candidate with a role in regulation.

The functional impairment of MTases may have implications in biological processes of the *Mtb* controlled by genes whose expression is affected by the methylation status. This could eventually influence the *Mtb*'s virulence, pathogenicity or drug resistance. For instance, variable methylation status was found to be related to the differential transcription of genes associated with metabolic pathways, among others, which suggests the potential role of methylation on regulation of biological processes related with growth or persistence. However, further work is needed to understand how methylation regulates gene expression under different environmental cues including those encountered by *Mtb* inside the host.

In *Mtb*, virulence and the ability to become drug resistant vary across lineages^{40,41}. Hence, the study of lineage-specific transcriptomic profiles and the mechanisms that regulate gene expression can give insights into mechanisms underlying these biological differences. Such insights will be useful to identify potential targets for the development of new anti-tuberculosis drugs or vaccines. The small sample size is a potential limitation of the study, but our integrated analysis has detected known variants and methylated motifs, and putative candidate eQTLs for follow-up experiments. Future studies should consider larger sample sizes, including more lineages (e.g. other ancient lineages, such as L5 and L6), in order to confirm the candidate associations found in this analysis. In addition, there is a need for complementary proteomic analyses, to perform a comprehensive integrated study of *Mtb* genetic and epigenetic mechanisms of gene expression control. Overall, our data has identified common functional variants that affect transcriptional control, which gives further support to differential pathophysiology in ancient and modern *Mtb* lineages.

Materials and Methods

Bacterial strains, DNA and RNA sequencing. All 22 *Mtb* isolates listed in Supplementary Table S1 were sourced from 22 TB patients from Karonga (Malawi) between 2003 and 2009, and cultured in the LSHTM. *Mtb* isolates were grown by liquid culture (in the absence of antimicrobial drugs) from frozen stocks of Lowenstein-Jensen or liquid cultures derived from patient's sputum specimens already isolated. *Mtb* strains were grown to mid-log phase (OD = 0.6–0.8) in Middlebrook 7H9 supplemented with 0.05% Tween 80 and 10% albumin-dextrose-catalase (ADC) at 37 °C in standing 25 cm² vented tissue culture flasks and subcultured in 75 cm² vented tissue culture flasks. DNA and RNA were extracted from the same cultures (passage 3–4 from original sputum sample) using the phenol-chloroform-isoamyl alcohol method and the trizol method with bead-beating as previously described^{42,43}. The samples were sequenced at the Genome Institute of Singapore. Single-molecule real time (SMRT) sequencing from Pacific Biosciences (PacBio) RSII long read technology was used with the parameter of 6 hours per SMRTcell (PacBio RS II SMRT Cells 8Pac). The library preparation involved the use of the template prep kit 1.0, and the binding chemistry involved the use of DNA/Polymerase binding kit P6. The sequencing kit used was the DNA Sequencing Reagent Kit 4.0.

For RNA sequencing, total RNA extracts were run on the Agilent 4200 TapeStation System (Agilent Technologies, Santa Clara, CA, USA) using the RNA TapeStation Assay to determine the RNA integrity values. TruSeq Stranded mRNA sample preparation was used according to the manufacturer's instructions for next generation library preparation. Briefly, library preparation started with purification of mRNA using poly-T oligo attached magnetic beads, fragmentation of mRNA, 1st and 2nd strand cDNA synthesis, A-tailing and ligation of adapters with multiplex indexes. Samples were enriched with 15 PCR cycles followed by Agencourt AMPure XP magnetic bead (Beckman Coulter, Brea, CA, USA) clean up as per the manufacturer's instructions. Quality of cDNA libraries was checked with Agilent D1000 TapeStation Assay (Agilent 4200 TapeStation System, Agilent Technologies, Santa Clara, CA, USA). Next generation sequencing was performed using Illumina HiSeq4000 flow cell, with 2 × 151 base pair-end runs. PhiX was used as a control.

Bioinformatic and association analysis. PacBio long reads were analysed using the pipelines provided by the SMRT Portal software. Briefly, raw sequence data were aligned to the H37Rv (GCA_00019595.2) reference genome and small variants (SNPs and indels) were called over the consensus sequences. Single nucleotide polymorphisms (SNPs) were used to build the maximum-likelihood phylogenetic tree using *RAXML* software⁴⁴. The Modification and Motif Analysis pipeline was used then for the methylation study and motif finding. Detection of

base modification was performed with a minimum QV score of 30 and coverage of 20-fold. Six-methyl-adenine (m6A) was determined within motifs with an inter-pulse duration ratio (IPD ratio) between 3 and 10. Statistical enrichment analysis was performed using *DAVID* software⁴⁵. Functional impairment prediction for proteins harbouring non-synonymous mutations was performed using the *Sorting Intolerant from tolerant* (SIFT) algorithm⁴⁶.

Pair-end short reads generated by Illumina HiSeq technology for RNA sequencing were assessed for quality and trimmed using *Trimmomatic* v0.36⁴⁷. High quality reads were mapped to the H37Rv reference genome (GCA_000195955.2) using the *Burrows-Wheeler Alignment* (BWA-mem) v0.7.15 tool⁴⁸. *HTSeq* 0.9.1⁴⁹ was used to quantify the number of reads per transcript. Lowly expressed genes were filtered out by a minimum count per million (CPM) value of 0.6, equivalent to 10 counts. For differential transcription analysis, counts were then normalised using the trimmed mean of M-values normalization (TMM) method⁵⁰. To compare expression levels between ancient and modern strains as well as for the eQTL studies linked with SNPs and methylation, significant differences were obtained through linear regression tests. Adjusted *p* values for multiple testing were calculated through the Bonferroni and Benjamini-Hochberg corrections for statistical significance. The prediction of promoter regions was performed using Neural Network Promoter Prediction (http://www.fruitfly.org/seq_tools/promoter.html). The EGRIN model from the MTB Network Portal³² and the regulatory network map from the TB Database³³ were used for the study of the association between transcriptional regulators and genes differentially expressed. The allele frequencies of variants identified in the eQTL analysis were calculated in an independent set of ancient and modern strains using a large published dataset (n = 6,218), described previously²⁶.

Data Availability

All pathogen raw sequencing data is available from the ENA short read archive (accession number PRJEB29197).

References

1. WHO. Global Tuberculosis Report 2017. WHO, WHO/HTM/TB/2017.23 (2017).
2. Brosch, R. *et al.* A new evolutionary scenario for the *Mycobacterium tuberculosis* complex. *Proc. Natl. Acad. Sci.* **99**, 3684–3689 (2002).
3. Koser, C. U., Feuerriegel, S., Summers, D. K., Archer, J. A. C. & Niemann, S. Importance of the Genetic Diversity within the *Mycobacterium tuberculosis* Complex for the Development of Novel Antibiotics and Diagnostic Tests of Drug Resistance. *Antimicrob. Agents Chemother.* **56**, 6080–6087 (2012).
4. Hershberg, R. *et al.* High functional diversity in *Mycobacterium tuberculosis* driven by genetic drift and human demography. *PLoS Biol.* **6**, 2658–2671 (2008).
5. Coscolla, M. & Gagneux, S. Consequences of genomic diversity in *Mycobacterium tuberculosis*. *Semin Immunol.* **26**, 431–444 (2014).
6. Coll, F. *et al.* PolyTB: A genomic variation map for *Mycobacterium tuberculosis*. *Tuberc.* **94**, 346–354 (2014).
7. Benavente, E. D. *et al.* PhyTB: Phylogenetic tree visualisation and sample positioning for *M. tuberculosis*. *BMC Bioinformatics* **16**, 155 (2015).
8. Gao, Q. *et al.* Gene expression diversity among *Mycobacterium tuberculosis* clinical isolates. *Microbiology* **151**, 5–14 (2005).
9. Homolka, S., Niemann, S., Russell, D. G. & Rohde, K. H. Functional genetic diversity among *Mycobacterium tuberculosis* complex clinical isolates: Delineation of conserved core and lineage-specific transcriptomes during intracellular survival. *PLoS Pathog.* **6**, 1–17 (2010).
10. Rose, G. *et al.* Mapping of genotype-phenotype diversity among clinical isolates of *Mycobacterium tuberculosis* by sequence-based transcriptional profiling. *Genome Biol. Evol.* **5**, 1849–1862 (2013).
11. Reed, M. B., Gagneux, S., DeRiemer, K., Small, P. M. & Barry, C. E. The W-Beijing lineage of *Mycobacterium tuberculosis* overproduces triglycerides and has the DosR dormancy regulon constitutively upregulated. *J. Bacteriol.* **189**, 2583–2589 (2007).
12. Fallow, A., Domenech, P. & Reed, M. B. Strains of the East Asian (W/Beijing) lineage of *Mycobacterium tuberculosis* are DosS/DosT-DosR two-component regulatory system natural mutants. *J. Bacteriol.* **192**, 2228–2238 (2010).
13. Domenech, P. *et al.* Unique regulation of the DosR regulon in the Beijing lineage of *Mycobacterium tuberculosis*. *J. Bacteriol.* **199**, 1–19 (2017).
14. Nica, A. C. & Dermizakis, E. T. Expression quantitative trait loci: present and future. *Philos. Trans. R. Soc. B Biol. Sci.* **368**, 20120362 (2013).
15. Zhu, L. *et al.* Precision methylome characterization of *Mycobacterium tuberculosis* complex (MTBC) using PacBio single-molecule real-time (SMRT) technology. *Nucleic Acids Res.* **44**, 730–743 (2016).
16. Phelan, J. *et al.* Methylation in *Mycobacterium tuberculosis* is lineage specific with associated mutations present globally. *Sci. Rep.* **8**, 1–7 (2018).
17. Shell, S. S. *et al.* DNA Methylation Impacts Gene Expression and Ensures Hypoxic Survival of *Mycobacterium tuberculosis*. *PLoS Pathog.* **9**, 24–28 (2013).
18. Balbontin, R. *et al.* DNA adenine methylation regulates virulence gene expression in *Salmonella enterica* serovar typhimurium. *J. Bacteriol.* **188**, 8160–8168 (2006).
19. Crampin, A. C., Glynn, J. R. & Fine, P. E. M. What has Karonga taught us? Tuberculosis studied over three decades. *Int. J. Tuberc. Lung Dis.* **13**, 153–164 (2009).
20. Roetzer, A. *et al.* Whole Genome Sequencing versus Traditional Genotyping for Investigation of a *Mycobacterium tuberculosis* Outbreak: A Longitudinal Molecular Epidemiological Study. *PLoS Med.* **10**, e1001387 (2013).
21. Winglee, K. *et al.* Whole Genome Sequencing of *Mycobacterium africanum* Strains from Mali Provides Insights into the Mechanisms of Geographic Restriction. *PLoS Negl. Trop. Dis.* **10**, 1–28 (2016).
22. Sinha, R. *et al.* Methyl-accepting chemotaxis like Rv3499c (Mce4A) protein in *Mycobacterium tuberculosis* H37Rv mediates cholesterol-dependent survival. *Tuberculosis* **109**, 52–60 (2018).
23. Li, G. *et al.* Efflux pump gene expression in multidrug-resistant *Mycobacterium tuberculosis* clinical isolates. *PLoS One* **10**, 1–12 (2015).
24. Colangeli, R. *et al.* The *Mycobacterium tuberculosis* *iniA* gene is essential for activity of an efflux pump that confers drug tolerance to both isoniazid and ethambutol. *Mol. Microbiol.* **55**, 1829–1840 (2005).
25. Li, Y., Zeng, J., Zhang, H. & He, Z. G. The characterization of conserved binding motifs and potential target genes for *M. tuberculosis* MtrAB reveals a link between the two-component system and the drug resistance of *M. smegmatis*. *BMC Microbiol.* **10** (2010).
26. Coll, F. *et al.* Genome-wide analysis of multi- and extensively drug-resistant *Mycobacterium tuberculosis*. *Nat. Genet.* **50** (2018).
27. Cortes, T. *et al.* Genome-wide Mapping of Transcriptional Start Sites Defines an Extensive Leaderless Transcriptome in *Mycobacterium tuberculosis*. *Cell Rep.* **5**, 1121–1131 (2013).
28. Pandey, R. *et al.* MntR(Rv2788) a transcriptional regulator that controls manganese homeostasis in *Mycobacterium tuberculosis*. *Mol. Microbiol.* **98**, 1168–1183 (2015).

29. Balhana, R. J. C., Singla, A., Sikder, M. H., Withers, M. & Kendall, S. L. Global analyses of TetR family transcriptional regulators in mycobacteria indicates conservation across species and diversity in regulated functions. *BMC Genomics* **16**, 1–12 (2015).
30. Ahn, S. K., Cuthbertson, L. & Nodwell, J. R. Genome Context as a Predictive Tool for Identifying Regulatory Targets of the TetR Family Transcriptional Regulators. *PLoS One* **7**, e50562 (2012).
31. Quigley, J. *et al.* The cell wall lipid PDIM contributes to phagosomal escape and host cell exit of *Mycobacterium tuberculosis*. *MBio* **8**, 1–12 (2017).
32. Turkarslan, S. *et al.* A comprehensive map of genome-wide gene regulation in *Mycobacterium tuberculosis*. *Sci. Data* **2**, 1–10 (2015).
33. Galagan, J. E. *et al.* The *Mycobacterium tuberculosis* regulatory network and hypoxia. *Nature* **499**, 178–183 (2013).
34. Chauhan, R. *et al.* Reconstruction and topological characterization of the sigma factor regulatory network of *Mycobacterium tuberculosis*. *Nat. Commun.* **7** (2016).
35. Casadesús, J. & Low, D. Epigenetic gene regulation in the bacterial world. *Microbiol. Mol. Biol. Rev.* **70**, 830–856 (2006).
36. Suzuki, M. M. & Bird, A. DNA methylation landscapes: Provocative insights from epigenomics. *Nat. Rev. Genet.* **9**, 465–476 (2008).
37. Adhikari, S. & Curtis, P. D. DNA methyltransferases and epigenetic regulation in bacteria. *FEMS Microbiol. Rev.* **40**, 575–591 (2016).
38. Oshima, T. *et al.* Genome-wide analysis of deoxyadenosine methyltransferase-mediated control of gene expression in *Escherichia coli*. *Mol. Microbiol.* **45**, 673–695 (2002).
39. Hénaut, A., Rouxel, T., Gleizes, A., Moszer, I. & Danchin, A. Uneven distribution of GATC motifs in the *Escherichia coli* chromosome, its plasmids and its phages. *J. Mol. Biol.* **257**, 574–585 (1996).
40. Merker, M. *et al.* Evolutionary history and global spread of the *Mycobacterium tuberculosis* Beijing lineage. *Nat. Genet.* **47**, 242–249 (2015).
41. Parwati, I., van Crevel, R. & van Soolingen, D. Possible underlying mechanisms for successful emergence of the *Mycobacterium tuberculosis* Beijing genotype strains. *Lancet Infect. Dis.* **10**, 103–111 (2010).
42. Benjak, A., Sala, C. & Hartkoorn, R. C. Whole-Genome Sequencing for Comparative Genomics and De Novo Genome Assembly. In 1–16, https://doi.org/10.1007/978-1-4939-2450-9_1 (2015).
43. Tischler, A. D., Leistikow, R. L., Kirksey, M. A., Voskuil, M. I. & McKinney, J. D. *Mycobacterium tuberculosis* requires phosphate-responsive gene regulation to resist host immunity. *Infect. Immun.* **81**, 317–328 (2013).
44. Stamatakis, A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
45. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **4**, 44–57 (2009).
46. Kumar, P., Henikoff, S. & Ng, P. C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.* **4**, 1073–1082 (2009).
47. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
48. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
49. Anders, S., Pyl, P. T. & Huber, W. HTSeq-A Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).
50. Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* **11**, R25 (2010).

Acknowledgements

We thank Teresa Cortes for useful comments. P.J.G.-G. is funded by an MRC-LID PhD studentship. J.P. is funded by a Newton Institutional Links Grant (British Council. 261868591). T.G.C. is funded by the Medical Research Council UK (Grant Nos MR/M01360X/1, MR/N010469/1, MR/R025576/1, and MR/R020973/1) and BBSRC (Grant No. BB/R013063/1). S.C. is funded by Medical Research Council UK grants (MR/M01360X/1, MR/R025576/1, and MR/R020973/1). We gratefully acknowledge the Scientific Computing Group for data management and compute infrastructure at Genome Institute of Singapore for their help. The MRC eMedLab computing resource was used for bioinformatics and statistical analysis. The authors declare no conflicts of interest.

Author Contributions

M.L.H. and T.G.C. conceived and directed the project. A.C.C. and J.R.G. coordinated sample collection. N.A. undertook sample processing and DNA/RNA extraction. N.A., P.F.d.S. and M.L.H. coordinated sequencing. P.J.G.-G. performed bioinformatic and statistical analyses under the supervision of M.L.H. and T.G.C. P.J.G.-G., J.E.P., S.C., P.D.B., M.L.H. and T.G.C. interpreted results. P.J.G.-G. wrote the first draft of the manuscript with inputs from T.G.C. and M.L.H. All authors commented and edited on various versions of the draft manuscript and approved the final manuscript. P.J.G.-G., M.L.H. and T.G.C. compiled the final manuscript.

Additional Information

Supplementary information accompanies this paper at <https://doi.org/10.1038/s41598-019-41692-2>.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2019

An integrated whole genome analysis of *Mycobacterium tuberculosis* reveals insights into relationship between its genome, transcriptome and methylome

Paula J. Gomez-Gonzalez^{1,*}, Nuria Andreu^{1,*}, Jody Phelan¹, Paola Florez de Sessions², Judith R. Glynn³, Amelia C. Crampin^{3,4}, Susana Campino¹, Philip D. Butcher⁵, Martin L. Hibberd^{1,2,**}, Taane G. Clark^{1,3,**}

¹ Faculty of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine, London, United Kingdom

² Genome Institute Singapore, Singapore

³ Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London, United Kingdom

⁴ Malawi Epidemiology and Intervention Research Unit, Lilongwe, Malawi.

⁵ Institute for Infection & Immunity, St George's University of London, UK

* joint authors

** joint authors

Corresponding author:

Professor Taane G. Clark

Pathogen Molecular Biology Department

Faculty of Infectious and Tropical Diseases

London School of Hygiene and Tropical Medicine, London, United Kingdom

Supplementary Table S1

Characteristics of the strains analysed

Isolate ID	Year of collection	Sub-lineage*	Number of SNPs	Number of transcripts**	HIV status	Age	Gender	INH	STR
RBB389	2008	<u>1.1.2</u> (EAI3; EAI5)	2751	3836	+	35	F	S	S
RBB395	2009	<u>1.1.2</u> (EAI3; EAI5)	2612	3950	+	37	F	S	S
RBB398	2009	<u>1.1.2</u> (EAI3; EAI5)	2583	3961	+	31	M	S	S
RBB383	2007	<u>1.1.3</u> (EAI6)	2320	3934	+	35	M	S	S
RBB385	2007	<u>1.1.3</u> (EAI6)	2619	3948	+	40	M	S	S
RBB388	2008	<u>1.2.2</u> (EAI1)	2632	3935	-	36	F	S	S
RBB394	2009	<u>1.2.2</u> (EAI1)	2612	3765	+	53	M	R	S
RBB397	2009	<u>1.2.2</u> (EAI1)	2643	3953	+	33	F	S***	S***
RBB401	2010	<u>2.2.1</u> (Beijing)	1651	3891	+	46	F	S	S
RBB402	2010	<u>2.2.1</u> (Beijing)	1676	3937	-	74	M	S	S
RBB384	2007	<u>4.1.1.3</u> (Haarlem X1, X3)	1473	3931	+	43	M	S	S
RBB399	2010	<u>4.1.1.3</u> (Haarlem X1, X3)	1423	3934	+	26	F	S	S
RBB404	2004	<u>4.1.1.3</u> (Haarlem X1, X3)	1217	3950	-	33	F	S	S
RBB387	2007	<u>4.1.2</u> (X-type)	1482	3933	+	45	M	S	S
RBB392	2008	<u>4.1.2</u> (X-type)	1426	3953	+	63	M	S	S
RBB386	2007	<u>4.3.3</u> (LAM)	1298	3345	+	50	F	S	S
RBB396	2009	<u>4.3.3</u> (LAM)	1102	3950	-	33	M	S	S
RBB403	2003	<u>4.3.4.2.1</u> (LAM)	1063	3929	-	67	M	S	S
RBB391	2008	<u>4.8</u> (T)	857	3950	+	34	M	S	S
RBB390	2008	<u>4.9</u> (T1-H37Rv)	528	3967	+	35	M	R	R
RBB393	2008	<u>4.9</u> (T1-H37Rv)	635	3083	-	49	F	S	S
RBB400	2010	<u>4.9</u> (T1-H37Rv)	634	3965	+	18	M	S***	S***

* Lineages are underlined; isoniazid (INH) and streptomycin (STR) drug susceptibility test (R: resistant; S: susceptible); ** number of genes transcribed with at least 10 counts; *** inferred by whole-genome sequencing.

Supplementary Table S2

105 genes found to be differentially expressed between ancient (lineage 1) and modern (lineages 2 and 4) isolates

Gene	Log2 Fold-change (ancient vs modern)	Adjusted <i>p</i> value
<i>Rv0028</i>	0.930	3.48x10 ⁻⁴
<i>Rv0060</i>	1.260	1.22x10 ⁻⁵
<i>ephF (Rv0134)</i>	-1.215	1.53x10 ⁻³
<i>Rv0192</i>	0.240	5.67x10 ⁻³
<i>Rv0193c</i>	1.663	2.94x10 ⁻³
<i>nirB (Rv0252)</i>	2.701	7.38x10 ⁻⁴
<i>nirD (Rv0253)</i>	2.661	5.12x10 ⁻³
<i>Rv0273c</i>	-1.642	8.89x10 ⁻⁸
<i>Rv0275c</i>	-2.572	9.48x10 ⁻⁵
<i>Rv0276</i>	-5.341	1.89x10 ⁻⁹
<i>PPE3 (Rv0280)</i>	-2.303	5.11x10 ⁻³
<i>iniB (Rv0341)</i>	-4.054	6.23x10 ⁻⁵
<i>iniA (Rv0342)</i>	-2.367	3.36x10 ⁻⁴
<i>iniC (Rv0343)</i>	-2.388	3.35x10 ⁻⁷
<i>icl1 (Rv0467)</i>	2.587	2.85x10 ⁻²
<i>fadB2 (Rv0468)</i>	1.426	6.48x10 ⁻³
<i>umaA (Rv0469)</i>	1.563	9.27x10 ⁻⁷
<i>Rv0520</i>	-2.961	3.65x10 ⁻⁶
<i>galTb (Rv0619)</i>	3.691	3.84x10 ⁻⁵
<i>galK (Rv0620)</i>	5.493	2.81x10 ⁻⁶
<i>recB (Rv0630c)</i>	1.550	4.31x10 ⁻⁵
<i>Rv0687</i>	-1.190	4.83x10 ⁻⁵
<i>rplN (Rv0714)</i>	0.509	1.17x10 ⁻²
<i>Rv0826</i>	-3.974	3.82x10 ⁻²
<i>rpfA (Rv0867c)</i>	1.647	3.96x10 ⁻⁴
<i>Rv0906</i>	-1.416	7.57x10 ⁻⁵
<i>Rv0966c</i>	-1.646	1.14x10 ⁻³
<i>Rv0997</i>	0.981	5.84x10 ⁻³
<i>Rv1101c</i>	-1.319	2.58x10 ⁻³
<i>narH (Rv1162)</i>	-2.399	6.90x10 ⁻³
<i>narJ (Rv1163)</i>	-2.577	2.19x10 ⁻³
<i>narI (Rv1164)</i>	-1.682	1.78x10 ⁻⁴
<i>deaD (Rv1253)</i>	-1.116	1.41x10 ⁻³
<i>Rv1261c</i>	0.313	4.52x10 ⁻²
<i>pknH (Rv1266c)</i>	1.069	5.04x10 ⁻⁵
<i>Rv1319c</i>	-0.796	1.58x10 ⁻²
<i>PE_PGRS25 (Rv1396c)</i>	-2.440	2.14x10 ⁻²

<i>vapC10 (Rv1397c)</i>	-3.070	6.82x10 ⁻⁵
<i>vapB10 (Rv1398c)</i>	-3.289	1.86x10 ⁻⁴
<i>Rv1503c</i>	1.836	1.98x10 ⁻³
<i>Rv1504c</i>	2.443	5.41x10 ⁻³
<i>Rv1505c</i>	2.284	8.78x10 ⁻³
<i>Rv1524*</i>	-2.051	2.57x10 ⁻⁷
<i>wbbL2* (Rv1525)</i>	-7.438	1.67x10 ⁻⁹
<i>plsB1 (Rv1551)</i>	2.205	1.391x10 ⁻⁴
<i>cya (Rv1625c)</i>	-0.956	3.67x10 ⁻²
<i>Rv1627c</i>	-0.585	4.60x10 ⁻²
<i>malQ (Rv1781c)</i>	1.548	2.24x10 ⁻²
<i>PE_PGRS34 (Rv1840c)</i>	-1.116	5.21x10 ⁻³
<i>Rv1842c</i>	-0.860	9.05x10 ⁻³
<i>Rv1895</i>	2.647	3.33x10 ⁻⁶
<i>lppD (Rv1899c)</i>	-4.085	1.19x10 ⁻¹⁵
<i>PPE35 (Rv1918c)</i>	-1.220	4.30x10 ⁻³
<i>fadD31 (Rv1925)</i>	-2.890	7.10x10 ⁻¹³
<i>Rv1976c</i>	0.753	1.08x10 ⁻²
<i>Rv2059</i>	-1.244	1.14x10 ⁻²
<i>Rv2159c</i>	-5.676	3.92x10 ⁻³
<i>Rv2160A</i>	-5.798	7.13x10 ⁻⁴
<i>Rv2161c</i>	-5.754	5.41x10 ⁻⁷
<i>PE_PGRS38 (Rv2162c)</i>	-4.112	9.36x10 ⁻¹²
<i>pimB (Rv2188c)</i>	-3.360	7.74x10 ⁻⁶
<i>Rv2271</i>	0.950	6.28x10 ⁻⁴
<i>Rv2272</i>	0.984	3.94x10 ⁻⁵
<i>Rv2282c</i>	0.801	1.96x10 ⁻²
<i>narK1 (Rv2329c)</i>	1.844	3.99x10 ⁻²
<i>Rv2337c</i>	-2.377	6.43x10 ⁻⁴
<i>PE_PGRS42 (Rv2487c)</i>	-1.238	3.06x10 ⁻²
<i>Rv2652c*</i>	-2.103	4.43x10 ⁻²
<i>Rv2653c*</i>	-2.666	4.22x10 ⁻²
<i>Rv2658c*</i>	-2.663	4.22x10 ⁻²
<i>Rv2712c</i>	1.896	1.98x10 ⁻⁹
<i>Rv2719c</i>	1.951	1.99x10 ⁻⁹
<i>Rv2765</i>	4.360	1.14x10 ⁻⁸
<i>vapB22 (Rv2830c)</i>	1.551	2.03x10 ⁻³
<i>amt (Rv2920c)</i>	1.091	3.90x10 ⁻²
<i>Rv2994</i>	-1.690	3.48x10 ⁻⁷
<i>Rv3007c</i>	-2.168	2.43x10 ⁻⁹
<i>virS (Rv3082c)</i>	3.052	3.34x10 ⁻³
<i>PPE51 (Rv3136)</i>	2.571	1.23x10 ⁻⁶
<i>Rv3137</i>	2.665	2.45x10 ⁻⁴

<i>pfIA</i> (Rv3138)	1.119	2.17x10 ⁻²
<i>PPE52</i> (Rv3144c)	-0.995	1.33x10 ⁻²
<i>Rv3165c</i>	0.592	6.34x10 ⁻³
<i>Rv3167c</i>	2.247	2.70x10 ⁻³
<i>Rv3168</i>	2.106	4.94x10 ⁻²
<i>Rv3169</i>	2.295	2.65x10 ⁻⁴
<i>Rv3233c</i>	-1.733	6.25x10 ⁻⁷
<i>Rv3446c</i>	2.177	4.47x10 ⁻³
<i>mce4A</i> (Rv3499c)	0.616	1.96x10 ⁻³
<i>yrbE4B</i> (Rv3500c)	1.697	5.17x10 ⁻⁴
<i>PE_PGRS57*</i> (Rv3514)	-2.748	1.52x10 ⁻⁶
<i>Rv3527</i>	-1.003	4.06x10 ⁻²
<i>PE33</i> (Rv3650)	1.136	3.74x10 ⁻²
<i>PE_PGRS60</i> (Rv3652)	3.675	3.69x10 ⁻¹¹
<i>PE_PGRS61</i> (Rv3653)	4.647	1.49x10 ⁻⁹
<i>Rv3679</i>	-4.162	3.05x10 ⁻¹²
<i>Rv3680</i>	-3.146	7.88x10 ⁻¹¹
<i>Rv3695</i>	-2.202	1.58x10 ⁻⁹
<i>Rv3740c</i>	1.246	2.69x10 ⁻⁴
<i>Rv3741c</i>	1.682	2.60x10 ⁻²
<i>Rv3742c</i>	2.550	3.70x10 ⁻³
<i>accD4</i> (Rv3799c)	-1.195	7.26x10 ⁻³
<i>pks13</i> (Rv3800c)	-1.390	6.95x10 ⁻⁴
<i>PE_PGRS62</i> (Rv3812)	1.249	1.82x10 ⁻⁷
<i>Rv3915</i>	0.244	5.50x10 ⁻³

* Genes deleted in ancient (lineage 1) isolates.

Adjusted *p* value obtained by Bonferroni correction.

Supplementary Table S3

42 genes found to be under-expressed (adjusted $p < 0.05$) and associated with large genomic deletions

Gene	Lineage/sub-lineage with the deletion
<i>Rv0072</i>	L2
<i>Rv0073</i>	L2
<i>msrA (Rv0137c)</i>	4.3.4.2.1
<i>Rv0195</i>	4.1.2
<i>aac (Rv0262c)</i>	1.2.2*
<i>Rv0265c</i>	1.2.2*
<i>oplA (Rv0266c)</i>	1.2.2*
<i>Rv1524</i>	1
<i>wbbL2 (Rv1525)</i>	1
<i>gabD2 (Rv1731)</i>	1.1.3
<i>PE18 (Rv1788)</i>	1.2.2*
<i>PE26 (Rv1789)</i>	1.2.2*
<i>Rv1993c</i>	4.3.4.2.1
<i>cmtR (Rv1994c)</i>	4.3.4.2.1
<i>plcC (Rv2349c)</i>	1.1.3
<i>plcB (Rv2350c)</i>	1.1.3
<i>plcA (Rv2351c)</i>	1.1.3
<i>PPE39 (Rv2353c)</i>	2
<i>Rv2645</i>	1*
<i>Rv2646</i>	1*
<i>Rv2647</i>	1*
<i>Rv2651c</i>	1*
<i>Rv2652c</i>	1*
<i>Rv2655c</i>	1*
<i>Rv2656c</i>	1*
<i>Rv2657c</i>	1*
<i>Rv2658c</i>	1*
<i>Rv2819c</i>	2
<i>PPE55 (Rv3347c)</i>	4.3.4.2.1
<i>Rv3349c</i>	4.3.4.2.1
<i>PPE56 (Rv3350c)</i>	4.3.4.2.1
<i>Rv3351c</i>	4.3.4.2.1
<i>lytB1 (Rv3382c)</i>	2*
<i>cmaA1 (Rv3392c)</i>	4.9*
<i>PPE58 (Rv3426)</i>	1,2,4.8,4.9
<i>Rv3468c</i>	4.8
<i>mhpE (Rv3469c)</i>	4.8
<i>ilvB2 (Rv3470c)</i>	4.8
<i>Rv3471c</i>	4.8
<i>Rv3472</i>	4.8
<i>bpoA (Rv3473c)</i>	4.8
<i>kgtP (Rv3476c)</i>	4.8

* Not all the clinical isolates from the lineage or sub-lineage.

Supplementary Table S4

76 genes found to be differentially expressed (adjusted $p < 0.05$) through eQTL analysis

Gene	Number of SNPs associated**	Lineage/sub-lineage	Regulation
<i>Rv0273c</i>	798	1	Down
<i>Rv0276</i>	790	1	Down
<i>iniC (Rv0343)</i>	790	1	Down
<i>umaA (Rv0469)</i>	790	1	Up
<i>Rv0520</i>	790	1	Down
<i>Rv0576</i>	4	1.1.2*, 1.2.2*	Up
<i>mce2R (Rv0586)</i>	1	1.1.2*, 1.1.3*	Up
<i>mce2D (Rv0592)</i>	169	1, 2	Up
<i>galK (Rv0620)</i>	790	1	up
<i>recB (Rv0630c)</i>	7	1*	Up
<i>mazF2 (Rv0659c)</i>	297	4.1.2	Down
<i>mazE2 (Rv0660c)</i>	297	4.1.2	Down
<i>Rv0687</i>	84	1	Down
<i>Rv0750</i>	368	4.1.1.3	Up
<i>Rv0958</i>	398	1.1.3	Up
<i>Rv0959</i>	398	1.1.3	Up
<i>Rv1096</i>	93	1, 2, 4.1, 4.3, 4.8	Up
<i>Rv1101c</i>	169	1, 2	Down
<i>bpoB (Rv1123c)</i>	368	4.1.1.3	Down
<i>narH (Rv1162)</i>	6	1*	Down
<i>narJ (Rv1163)</i>	7	1*	Down
<i>narI (Rv1164)</i>	7	1*	Down
<i>Rv1318c</i>	137	4.3	Up
<i>Rv1371</i>	93	1, 2, 4.1, 4.3, 4.8	Up
<i>vapC10 (Rv1397c)</i>	7	1*	Down
<i>vapB10 (Rv1398c)</i>	7	1*	Down
<i>Rv1429</i>	368	4.1.1.3	Up
<i>bisC (Rv1442)</i>	484	1.2.2*	Up
<i>Rv1489</i>	297	1.2.2*	Down
<i>Rv1489A</i>	297	1.2.2*	Down
<i>Rv1490</i>	297	1.2.2*	Down
<i>Rv1491c</i>	297	1.2.2*	Down
<i>Rv1764</i>	94	1, 2, 4.1, 4.3, 4.8	Down
<i>Rv1895</i>	798	1	Up
<i>lppD (Rv1899c)</i>	791	1	Down
<i>fadD31 (Rv1925)</i>	805	1	Down
<i>Rv1976c</i>	169	1, 2	Up
<i>vapC36 (Rv1982c)</i>	121	4.1	Up
<i>Rv2077c</i>	127	1, 2, 4.1, 4.3	Down
<i>Rv2159c</i>	170	1, 2	Down
<i>Rv2160A</i>	169	1, 2	Down
<i>Rv2161c</i>	963	1, 2	Down
<i>PE_PGRS38 (Rv2162c)</i>	790	1	Down
<i>Rv2271</i>	1	1	Up
<i>Rv2324</i>	4	1.1.2*, 1.2.2*	Up

<i>Rv2337c</i>	7	1*	Down
<i>vapB38 (Rv2493)</i>	121	4.1	Up
<i>vapC38 (Rv2494)</i>	121	4.1	Up
<i>arsC (Rv2643)</i>	226	1.1.3*	Up
<i>Rv2712c</i>	790	1	Up
<i>Rv2719c</i>	790	1	Up
<i>Rv2765</i>	797	1	Up
<i>Rv2915c</i>	226	1.1.3*	Up
<i>Rv2972c</i>	1	1, 2	Up
<i>recG (Rv2973c)</i>	1	1, 2	Up
<i>Rv2974c</i>	169	1, 2	Up
<i>Rv2994</i>	790	1	Down
<i>Rv3007c</i>	790	1	Down
<i>PPE51 (Rv3136)</i>	797	1*	Up
<i>Rv3169</i>	7	1*	Up
<i>Rv3233c</i>	790	1	Down
<i>Rv3322c</i>	1	4.1.2, 4.9	Down
<i>moaC3 (Rv3324c)</i>	93	4.9	Down
<i>spoU (Rv3366)</i>	584	2	Up
<i>fadD17 (Rv3506)</i>	177	4.9*	Up
<i>PE_PGRS60 (Rv3652)</i>	790	1	Up
<i>PE_PGRS61 (Rv3653)</i>	790	1	Up
<i>Rv3679</i>	791	1	Down
<i>Rv3680</i>	790	1	Down
<i>Rv3695</i>	790	1	Down
<i>Rv3706c</i>	198	1.1.3*	Up
<i>Rv3750c</i>	127	4.8, 4.9	Up
<i>tcrX (Rv3765c)</i>	198	1.1.3*	Up
<i>PE_PGRS62 (Rv3812)</i>	790	1	Up
<i>Rv3829c</i>	584	2	Up
<i>Rv3830c</i>	584	2	up

* Not all the clinical isolates from the lineage or sub-lineage.

** Number of common SNPs in isolates with a gene over- or under-expressed compared to the rest of isolates not carrying the SNPs. All the lineage or sub-lineage specific SNPs are therefore associated with genes differentially expressed by lineage or sub-lineage.

Supplementary Table S5

Functional SNPs located in the upstream intergenic region, upstream gene or within the gene associated with differential expression (*cis*-eQTLs, adjusted $p < 0.05$)

	Transcript differentially expressed	Annotation	SNP	Position SNP			Regulation	Strain Lineage	Allele frequency**	
				Gene	Distance (bp) from start codon	Promoter (P)/TSS			Ancient	Modern
	<i>Rv0068</i>	3	C75231T	IGR	-70	P	Up	4.1.2	0	0.009
	<i>Rv0193c</i>	1	G226676A	IGR	-105	-	Up	1	0.973	0
	<i>gpdA1</i>	4	T655986G	IGR	-37	P	Up	1,2	0.976	0.324
	<i>Rv0669c</i>	3	T769663G	IGR	-66	P	Down	4.3.3	0	0.050
	<i>Rv0750</i>	1	C841924T	IGR	-109	-	Up	4.1.1.3	0.003	0.038
SNPs in upstream intergenic region (IGR)	<i>Rv0958</i>	3	C1069871T	IGR	-12	P	Up	1.1.3	0.220	0
	<i>Rv1096</i>	3	T1224367C	IGR	-18	P	Down	1,2,4.1,4.3,4.8	1	0.976
	<i>Rv1503c</i>	1	A1694547C	IGR	-3	-	Up	1	0.973	0
	<i>fadD31</i>	4	T2177073C	IGR	-14	TSS/P	Down	1	0.973	0
	<i>PE_PGRS38</i>	7	A2424864G	IGR	-18	TSS	Down	1	0.973	0
	<i>Rv2712c</i>	1	C3025431T	IGR	-103	P	Up	1	0.971	0
	<i>vapB22</i>	5	T3137237C	IGR	-13	P	Up	1	0.973	0
	<i>Rv2923c</i>	1	G3238516A	IGR	-17	-	Up	4.1.2	0	0.009
	<i>Fpg</i>	8	G3239476A	IGR	-6	-	Up	4.1.2	0	0.008
	<i>Rv3695</i>	2	T4137190C	IGR	-16	-	Down	1	0.973	0
	<i>Rv0060</i>	1	C64028T	<i>Rv0060</i>	119	-	Up	1	0.973	0.002
	<i>ephF</i>	5	G162226A	<i>ephF</i>	455	-	Down	1	0.973	0

SNPs within gene or upstream gene	<i>Rv0193c</i>	1	C225668T	<i>Rv0193c</i>	903	-	Up	1	0.971	0
	<i>Rv0275c</i>	6	G331588A	<i>Rv0275c</i>	70	-	Down	1	0.973	0
	<i>Rv0276</i>	1	G331588A	<i>Rv0275c</i>	160	-	Down	1	0.973	0
	<i>PPE3</i>	7	C339508T	<i>PPE3</i>	144	-	Down	1	0.968	0
	<i>PPE5</i>	7	C370229T	<i>PPE5</i>	2535	-	Down	1.1.3	0.230	0
	<i>Rv0326</i>	-	T392261C	<i>Rv0325</i>	-12	-	Up	1,2	0.978	0.324
	<i>iniA</i>	2	T412280G	<i>iniA</i>	1442	-	Down	1	0.973	0
	<i>Rv0376c</i>	1	T454295C	<i>Rv0376c</i>	77	-	Up	1,2,4.1,4.3.4,4.8,4.9	1	0.994
	<i>Rv0377</i>	6	T454295C	<i>Rv0376c</i>	-126	-	Up	1,2,4.1,4.3.4,4.8,4.9	1	0.994
	<i>umaA</i>	4	C560664T	<i>umaA</i>	776	-	Up	1	0.973	0
			A560666G		778	-			0.973	0
	<i>mce2R</i>	6	C684611T	<i>mce2R</i>	201	-	Up	1.1.2*	0.024	0
	<i>mce2C</i>	5	A690450C	<i>mce2C</i>	1391	-	Up	1,2	0.976	0.324
	<i>mce2D</i>	6	A690450C	<i>mce2C</i>	-51	-	Up	1,2	0.976	0.324
	<i>recB</i>	8	G722852A	<i>recB</i>	2161	-	Up	1	0.973	0
	<i>Rv0669c</i>	3	A768395G	<i>Rv0669c</i>	1202	-	Down	4.3.3*	0	0
	<i>rpIN</i>	8	C811492G	<i>rpIN</i>	119	-	Up	1	0.970	0
	<i>Rv0750</i>	1	C842111G	<i>Rv0750</i>	78	-	Up	4.1.1.3	0.029	0.060
	<i>Rv0906</i>	1	C1009490T	<i>Rv0906</i>	546	-	Down	1	0.971	0
	<i>Rv0966c</i>	1	C1077754T	<i>Rv0966c</i>	81	-	Down	1	0.971	0
	<i>Rv1048c</i>	1	G1171183A	<i>Rv1048c</i>	970	-	Up	1.2.2*	0.021	0
	<i>bpoB</i>	5	G1246845A	<i>bpoB</i>	207	-	Down	4.1.1.3	0	0.003
	<i>deaD</i>	8	A1400396G	<i>deaD</i>	426	-	Down	1	0.971	0
	<i>Rv1318c</i>	3	G1480024T	<i>Rv1318c</i>	800	-	Up	4.3	0.013	0.277
	<i>Rv1319c</i>	3	T1481602G	<i>Rv1319c</i>	899	-	Down	1	0.970	0
<i>vapC10</i>	5	T1574206C	<i>vapC10</i>	307	-	Down	1	0.970	0	

<i>Rv1429</i>	1	C1605149T	<i>Rv1429</i>	271	-	Up	4.1.1.3	0.005	0.049
<i>bisC</i>	3	G1619841A	<i>bisC</i>	50	-	Up	1.2.2*	0.157	0
<i>Rv1505c</i>	1	G1695674A	<i>Rv1505c</i>	272	-	Up	1	0.973	0
<i>vapB11</i>	5	G1764812A	<i>vapB11</i>	57	-	Up	4.3.3*	0	0
<i>vapC11</i>	5	G1764812A	<i>vapB11</i>	-167	-	Up	4.3.3*	0	0
<i>Rv1773c</i>	6	G2007502A	<i>Rv1773c</i>	264	-	Up	4.1	0.003	0.176
<i>Rv1776c</i>	6	G2010096T	<i>Rv1776c</i>	459	-	Up	1.2.2*	0.019	0
<i>lldD2</i>	3	C2123181T	<i>Rv1873</i>	-30	-	Up	4.1.2	0.022	0.025
<i>lppD</i>	2	A2145878G	<i>lppD</i>	367	-	Down	1*	0.973	0
<i>fadD31</i>	4	G2177968T	<i>fadD31</i>	881	-	Down	1	0.973	0
<i>Rv1982c</i>	5	A2225456T	<i>Rv1982c</i>	376	-	Up	4.1	0.003	0.176
<i>Rv2036</i>	3	C2282058T	<i>Rv2035</i>	-41	-	Up	1.2.2*	0.157	0
<i>Rv2077c</i>	2	A2334007G	<i>Rv2077c</i>	287	-	Down	1,2,4.1,4.3	1	0.882
<i>Rv2159c</i>	1	A2421816G	<i>Rv2160A</i>	-151	-	Down	1,2	0.977	0.323
<i>Rv2160A</i>	6	A2421816G	<i>Rv2160A</i>	462	-	Down	1,2	0.977	0.323
<i>PE_PGRS38</i>	7	C2423785T	<i>PE_PGRS38</i>	1053	-	Down	1	0.962	0.001
<i>pimB</i>	4	G2450045A	<i>pimB</i>	1105	-	Down	1	0.971	0
		C2451081G		69	-			0.973	0
<i>Rv2263</i>	3	C2536599T	<i>Rv2263</i>	958	-	Down	2	0.003	0.126
<i>plcC</i>	3	G2627377T	<i>plcC</i>	1321	-	Down	1.1.3	0.232	0
<i>Rv2719c</i>	2	A3031285T	<i>Rv2719c</i>	252	-	Up	1	0.973	0
<i>Rv2765</i>	3	C3074830T	<i>Rv2765</i>	194	-	Up	1	0.973	0
<i>Rv2994</i>	2	G3351472A	<i>Rv2994</i>	203	-	Down	1	0.973	0
<i>Rv3027c</i>	1	G3386782A	<i>Rv3027c</i>	137	-	Up	4.1.2	0	0.009
<i>Rv3081</i>	1	C3446699G	<i>Rv3081</i>	659	-	Up	2	0.011	0.159
<i>virS</i>	5	A3447480C	<i>virS</i>	946	TSS	Up	1	0.973	0

<i>Rv3167c</i>	6	G3536008T	<i>Rv3167c</i>	49	-	Up	1	0.973	0
<i>Rv3180c</i>	1	C3549576A	<i>Rv3180c</i>	112	-	Up	1.1.2*	0.013	0
<i>lhr</i>	8	C3678298T	<i>lhr</i>	1523	-	Up	4.1.2*	0	0.004
<i>PPE55</i>	7	A3746409G	<i>PPE55</i>	6775	-	Up	1,2,4.1,4.3.3,4.8,4.9	-	-
		A3752207G		977	-			-	
<i>spoU</i>	8	GG3778011AT	<i>spoU</i>	274	-	Up	2	0.003	0.147
<i>PPE57</i>	7	T3842425A	<i>PPE57</i>	186	-	Down	4.1/4.3	-	-
		AG3842581GT		342	-			-	
<i>Rv3446c</i>	1	G3864041A	<i>Rv3446c</i>	490	-	Up	1	0.970	0
<i>kgtP</i>	2	A3892671G	<i>kgtP</i>	1049	-	Up	1,2,4.1,4.3,4.9	1	0.999
<i>yrbE4B</i>	5	G3920109T	<i>yrbE4A</i>	-47	-	Up	1	0.971	0
<i>fadD17</i>	4	C3925702T	<i>fadD17</i>	812	-	Up	4.9*	0.003	0
<i>PPE65</i>	7	A4060742G	<i>PPE65</i>	1147	-	Down	4.3.3*	0	0
<i>PE_PGRS60</i>	7	G4093719A	<i>PE_PGRS60</i>	87	-	Up	1	0.971	0
<i>Rv3679</i>	2	T4119246C	<i>Rv3679</i>	470	-	Down	1	0.968	0
<i>PE_PGRS62</i>	7	G4277032C	<i>PE_PGRS62</i>	461	-	Up	1	0.971	0

* Only one or two samples from the lineage out of the 3 analysed; ** Allele frequency refers to the proportion of strains harbouring the SNP in a larger data set (n = 6,218)⁵⁰.

Supplementary Table S6

Genes differentially expressed (adjusted $p < 0.05$) associated with transcriptional regulators carrying candidate impairing mutations

Transcriptional Regulator	Mutation	Genes Differentially Expressed	Regulation	Lineage	Allele frequency**	
					Ancient	Modern
<i>Rv0275c</i>	S24L	<i>Rv0276</i> , <i>Rv0520</i> , <i>Rv2162c</i> , <i>Rv0275c*</i> , <i>Rv0826</i>	Down	1	0.973	0
<i>ramB</i>	Q121R P91Q	<i>Rv1895</i> , <i>Rv3233c</i> , <i>Rv1164*</i> , <i>Rv1163*</i> , <i>Rv1162*</i>	Up/Down	1	0.973	0
<i>Rv1776c</i>	R154S	<i>Rv1048c</i> , <i>Rv1776c</i> , <i>Rv3136</i>	Up	1.2.2	0.019	0
<i>Rv3167c</i>	P17Q	<i>Rv1895</i>	Up	1	0.973	0
<i>Rv3249c</i>	T154A	<i>Rv1429</i> , <i>Rv1123c</i>	Up/Down	4.1.1.3	0.003	0.049

* Genes that are differentially expressed but didn't reach the cut off (adjusted $p < 0.05$).

** Allele frequency refers to the proportion of strains harbouring the mutation in a larger data set (n = 6,218)⁵⁰.

Supplementary Table S7

Mutations found in anti-sigma factors (as per H37Rv reference annotation) related with differential gene expression

Sigma Factor	Mutation	Lineage	Genes differentially expressed	Regulation	Allele frequency	
					Ancient	Modern
<i>rseA</i>	A23T	4.1.1.3	<i>Rv0750, Rv1429</i>	Up	0.003	0.049
<i>rskA</i>	E81D	2	<i>spoU, Rv3829c, Rv3830c</i>	Up	0.003	0.126
<i>rsfA</i>	L125R	1.2.2*	<i>bisC</i>	Up	0.148	0

* Not all the clinical isolates from the lineage or sub-lineage; Allele frequency refers to the proportion of strains harbouring the mutation in a larger data set (n = 6,218)⁵⁰.

Supplementary Table S8

Fractions of methylation for each identified motif

Sample	CTCCAG	CTGGAG	GATN ₄ RTAC	GTAYN ₄ ATC	CACGCAG	Lineage
RBB389	1795/1927 (0.93)	1704/1930 (0.88)	311/347 (0.9)	307/348 (0.88)	749/804 (0.93)	1.1.2
RBB398	1749/1934 (0.9)	1632/1937 (0.84)	304/356 (0.85)	289/355 (0.81)	739/811 (0.91)	1.1.2
RBB395	<u>6/1929</u> (0.003)	0/1930 (0)	296/351 (0.84)	289/348 (0.83)	759/806 (0.94)	1.1.2
RBB383	<u>211/1912</u> (0.11)	193/1911 (0.1)	<u>36/352</u> (0.1)	<u>31/351</u> (0.08)	<u>1/803</u> (0.001)	1.1.3
RBB385	1580/1941 (0.81)	1394/1940 (0.72)	272/361 (0.75)	269/362 (0.74)	<u>9/817</u> (0.01)	1.1.3
RBB388	1366/1924 (0.71)	1246/1925 (0.65)	249/347 (0.72)	237/348 (0.68)	640/804 (0.8)	1.2.2
RBB394	1208/1929 (0.63)	1134/1929 (0.59)	255/355 (0.72)	246/353 (0.69)	<u>0/808 (0)</u>	1.2.2
RBB397	1618/1927 (0.84)	1475/1928 (0.77)	294/347 (0.85)	280/347 (0.81)	720/805 (0.89)	1.2.2
RBB401	<u>0/1937 (0)</u>	<u>0/1935 (0)</u>	201/360 (0.56)	178/360 (0.5)	485/815 (0.6)	2.2.1
RBB402	<u>2/1938</u> (0.001)	<u>4/1938 (0.002)</u>	304/360 (0.85)	302/360 (0.84)	690/815 (0.84)	2.2.1
RBB384	1791/1930 (0.93)	1709/1934 (0.88)	307/355 (0.86)	305/357 (0.85)	766/810 (0.95)	4.1.1.3
RBB399	1772/1933 (0.92)	1685/1934 (0.87)	305/359 (0.85)	303/359 (0.84)	756/810 (0.93)	4.1.1.3
RBB404	1524/1930 (0.79)	1400/1930 (0.73)	273/355 (0.77)	252/358 (0.7)	701/810 (0.87)	4.1.1.3
RBB387	1767/1942 (0.91)	1622/1943 (0.83)	306/361 (0.85)	284/361 (0.79)	761/819 (0.93)	4.1.2
RBB392	1774/1942 (0.91)	1712/1943 (0.88)	305/361 (0.84)	302/360 (0.84)	756/819 (0.92)	4.1.2
RBB386	<u>104/1930</u> (0.05)	<u>109/1933</u> (0.06)	<u>0/361 (0)</u>	<u>0/362 (0)</u>	672/812 (0/83)	4.3.3
RBB396	1497/1944 (0.77)	1332/1943 (0.69)	<u>0/363 (0)</u>	<u>0/363 (0)</u>	732/819 (0.89)	4.3.3
RBB403	1586/1927 (0.82)	1408/1928 (0.73)	<u>0/362 (0)</u>	<u>0/362 (0)</u>	659/816 (0.89)	4.3.3
RBB391	1754/1935 (0.91)	1643/1936 (0.85)	<u>0/357 (0)</u>	<u>0/358 (0)</u>	<u>0/813 (0)</u>	4.8
RBB390	1827/1946 (0.94)	1760/1947 (0.9)	<u>0/363 (0)</u>	<u>0/363 (0)</u>	<u>0/819 (0)</u>	4.9
RBB393	1718/1943 (0.88)	1593/1945 (0.82)	<u>0/360 (0)</u>	<u>0/361 (0)</u>	<u>0/819 (0)</u>	4.9
RBB400	1812/1946 (0.93)	1751/1946 (0.9)	<u>0/362 (0)</u>	<u>0/362 (0)</u>	<u>0/820 (0)</u>	4.9

Methylated motifs/Total motifs (fraction of methylation). Cells coloured in red correspond to isolates with non-methylated motifs. Underlined in the motif shows the methylated nucleotide (m6A).

Supplementary Table S9

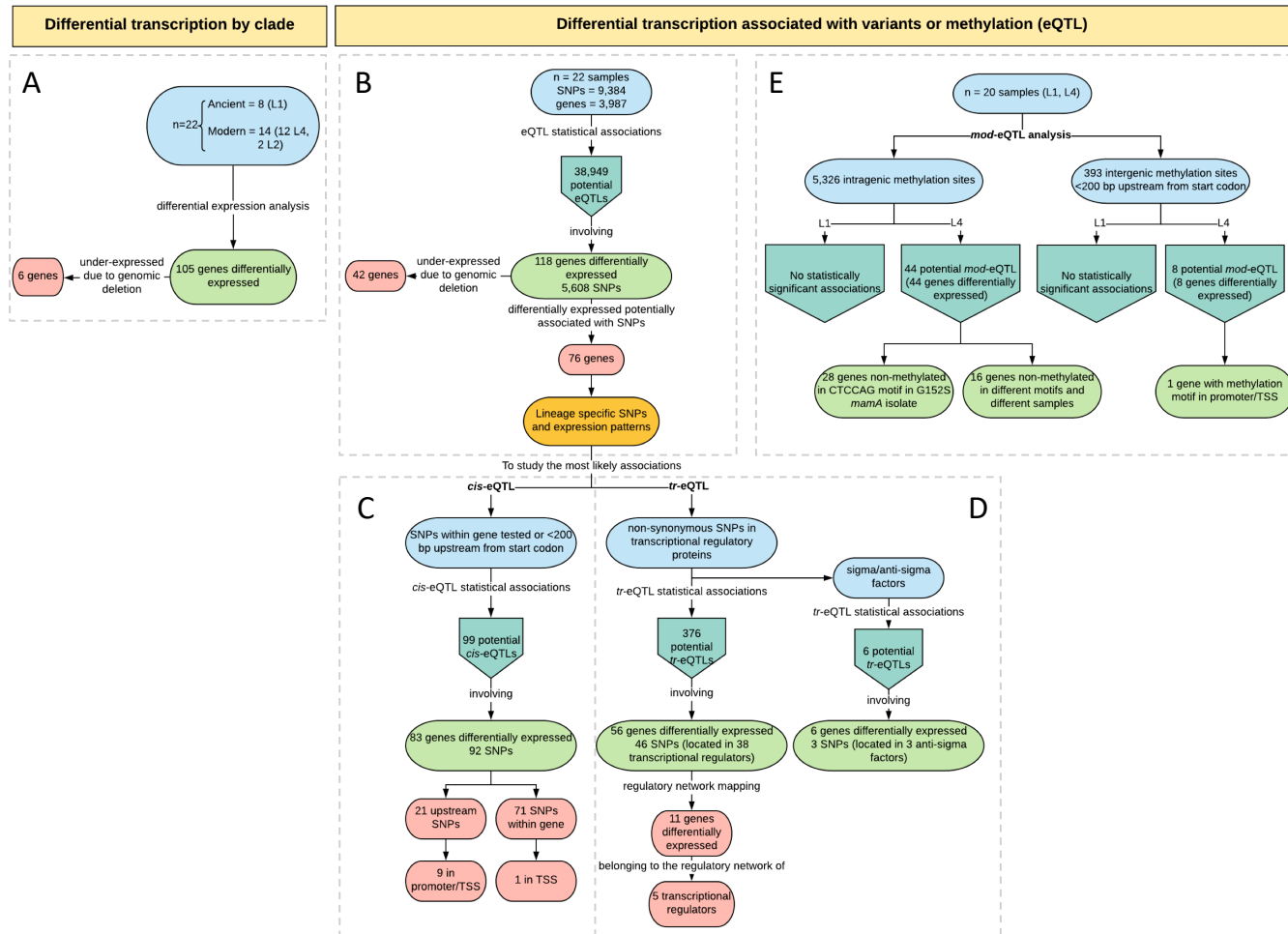
Mutations found in each *Mtb* MTase

Sample	<i>mamA</i>	<i>hsdM</i>	<i>mamB</i>	Lineage
RBB389	-	V93V	W47R, D154G, 1515delC	1.1.2
RBB398	-	V93V	W47R, D154G	1.1.2
RBB395	<u>Q340K, 121delG</u>	V93V	W47R, D154G, 1515delC	1.1.2
RBB383	-	V93V	W47R, D154G, S253L	1.1.3
RBB385	-	V93V	W47R, D154G, S253L	1.1.3
RBB388	-	V93V, T450T, K211Q	W47R, D154G	1.2.2
RBB394	-	V93V	W47R, D154G	1.2.2
RBB397	-	V93V	W47R, D154G	1.2.2
RBB401	E270A	-	W47R, D154G, S232S	2.2.1
RBB402	E270A	-	W47R, D154G, S232S	2.2.1
RBB384	-	-	W47R, D154G	4.1.1.3
RBB399	-	-	W47R, D154G, 1515insG	4.1.1.3
RBB404	-	-	W47R, D154G	4.1.1.3
RBB387	-	-	W47R, D154G	4.1.2
RBB392	-	-	W47R, D154G	4.1.2
RBB386	G152S, G72G	P306L	W47R, D154G	4.3.3
RBB396	G72G	P306L	W47R, D154G	4.3.3
RBB403	-	P306L	W47R, D154G	4.3.3
RBB391	-	P306L	W47R, D154G	4.8
RBB390	-	P306L	W47R	4.9
RBB393	-	P306L	W47R	4.9
RBB400	-	P306L	W47R	4.9

Mutations found in the three methyltransferases (MTases): *mamA*, *hsdM* and *mamB*. In bold, mutations involving amino-acidic changes potentially associated with the loss of function of the MTases, with novel candidates that might impact function of MTases underlined. Cells in red correspond to strains that did not present any of the motifs modified by those MTases methylated.

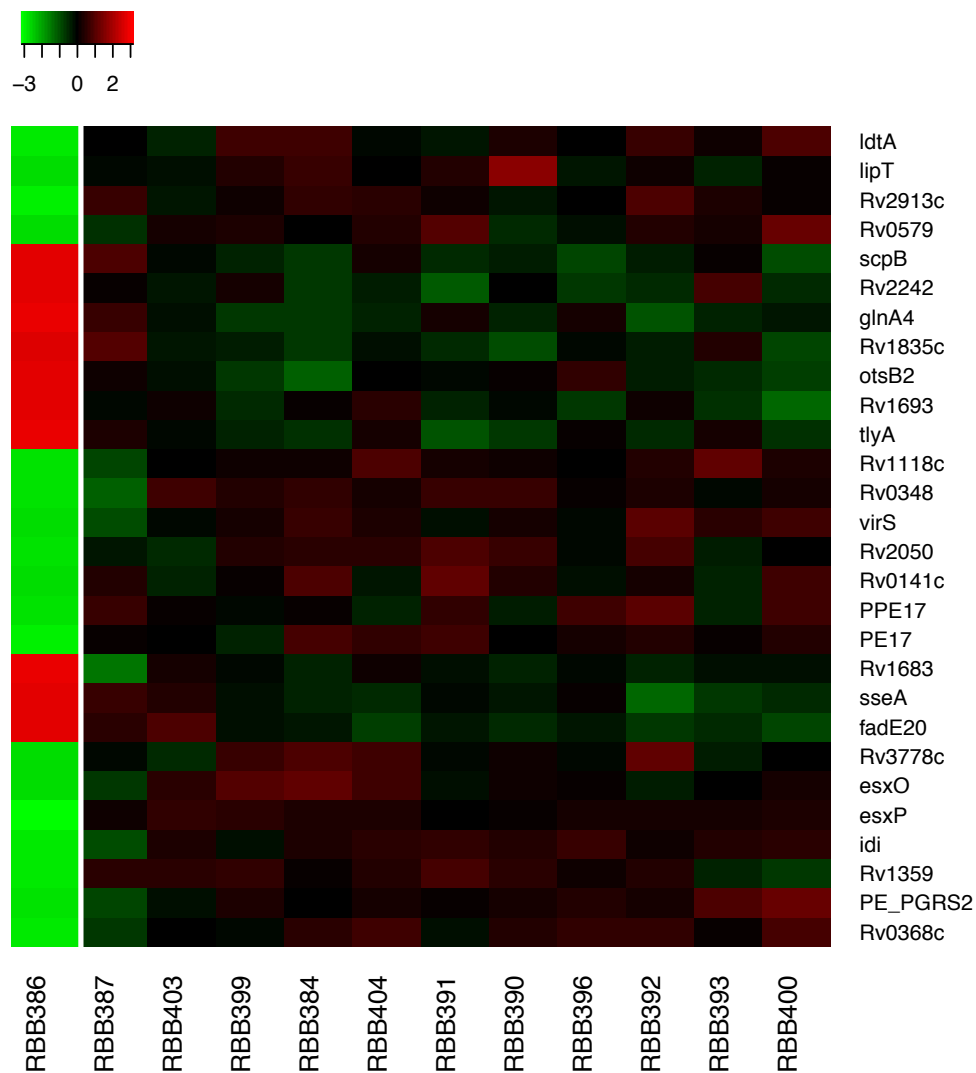
Supplementary Figure S1

The analytical workflow. **(A)** Differential gene expression analysis by clade (between ancient and modern strains). **(B)** eQTL analysis at whole-genome scale, looking for statistical associations between the 9,384 SNPs and 3,987 transcripts in the 22 samples. **(C)** *cis*-eQTL analysis using intragenic or <200 bp upstream SNPs from genes tested for differential transcription. **(D)** *tr*-eQTL analysis looking at the association between transcriptional regulators harbouring potential impairing mutations and differential transcription of genes found within their regulation networks. **(E)** Differential gene expression analysis linked with methylation status (intragenic or in promoter regions).



Supplementary Figure S2

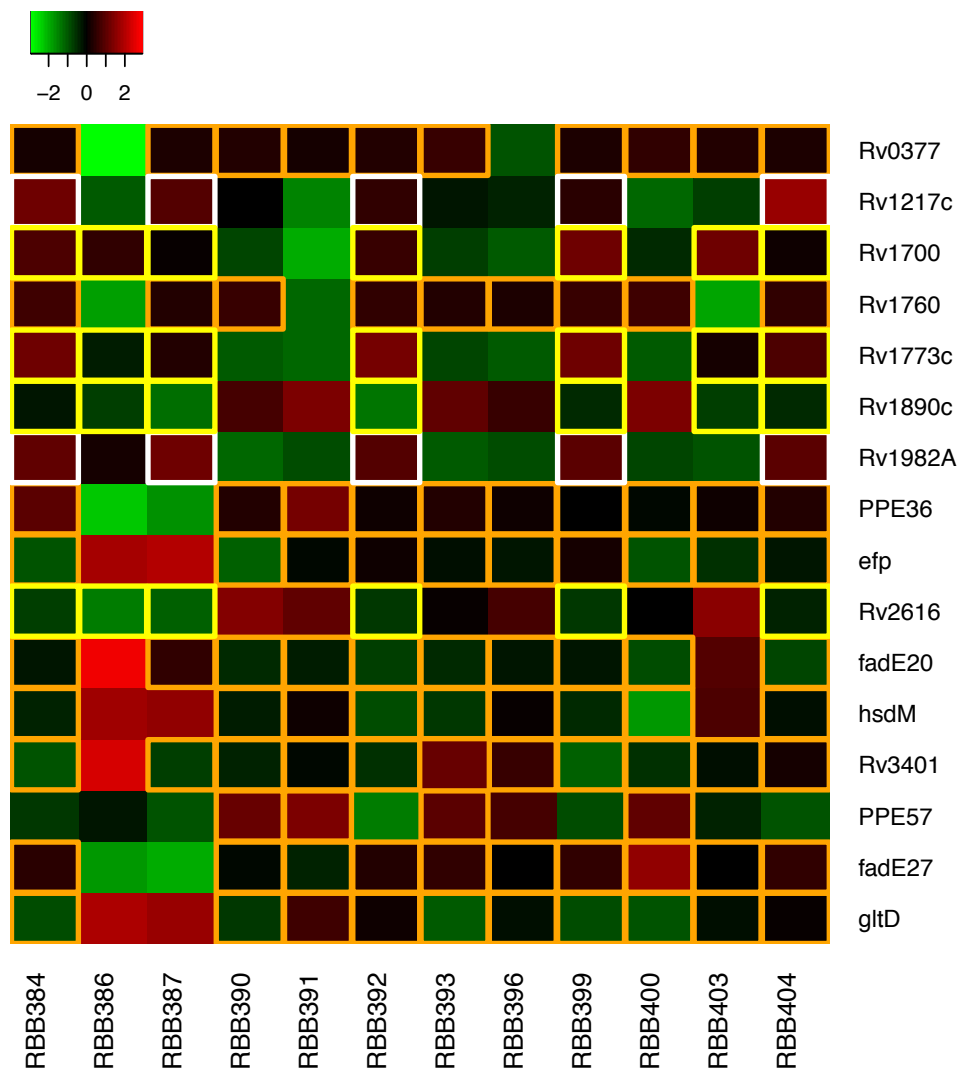
Differential expression of genes non-methylated only in sample with G152S mutation in lineage 4



Heatmap with the 28 genes differentially expressed among L4 isolates, associated with the lack of methylation in the sample harbouring the mutation G152S in *mamA* (RBB386), constructed with the gene expression distances between rows. Over-expressed genes are coloured in red and under-expressed ones in green. The isolate with none of the CTCCAG motifs methylated is bordered on the left of the white vertical line.

Supplementary Figure S3

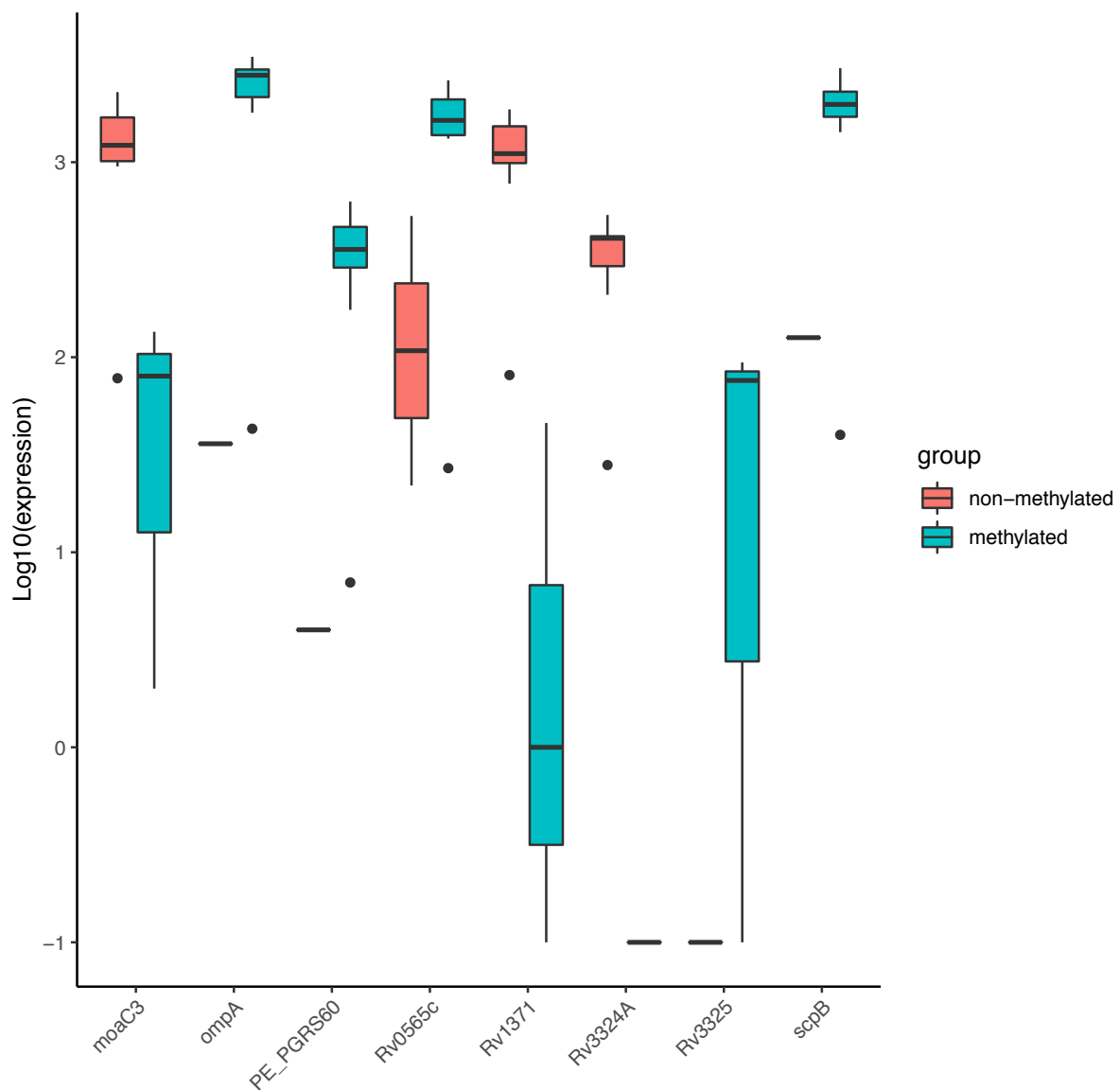
Differential expression of genes that non-methylated in Lineage 4 samples



Heatmap with the 16 genes differentially expressed among Lineage 4 samples associated with the lack of methylation of the different motifs, constructed with the gene expression distances between rows. The 28 genes that were non-methylated only in the strain that contained the G152S mutation are not shown. Over-expressed genes are coloured in red whilst under-expressed ones in green. Bordered cells represent the non-methylated samples for each gene. Bordered in orange are CTCCAG motifs, in yellow are CACGCAG motifs, and in white are GATN₄RTAC motifs.

Supplementary Figure S4

Comparison of expression levels of genes differentially expressed in lineage 4 clinical isolates with methylated and non-methylated motifs in intergenic regions upstream.



Boxplots showed the quartiles and median of the log10 of the expression levels for each gene labelled in the x-axis. Red boxplots represent those clinical isolates where the motif found in the upstream intergenic region is not methylated, whilst blue ones represent the isolates where it is methylated. Black points represent the number of samples falling in each of the two groups. When showing a line instead of a boxplot, only one sample is in the group.

CHAPTER 4

Genetic diversity of candidate loci linked to
Mycobacterium tuberculosis resistance to
bedaquiline, delamanid and pretomanid

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	lsh1704009	Title	
First Name(s)	Paula Josefina		
Surname/Family Name	Gómez González		
Thesis Title	Analysis of Mycobacterium tuberculosis 'omics data to inform on loci linked to drug resistance, pathogenicity and virulence		
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?	2021		
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.

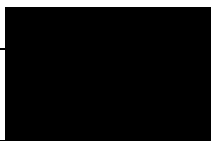
SECTION C – Prepared for publication, but not yet published

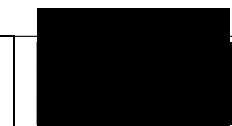
Where is the work intended to be published?	
Please list the paper's authors in the intended authorship order:	
Stage of publication	Choose an item.

SECTION D – Multi-authored work

<p>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)</p>	<p>I received the raw sequencing data and performed the bioinformatic analysis, which involved mapping, variant calling and phylogenetic reconstruction. Statistical analysis and plotting were performed with custom scripts in R. I wrote the first draft of the manuscript and circulated to co-authors. After receiving feedback and comments, I revised the manuscript and submitted to Scientific Reports, and dealt with subsequent revisions.</p>
---	---

SECTION E

Student Signature		
Date	28/01/2022	

Supervisor Signature		
Date	28/01/2022	



OPEN

Genetic diversity of candidate loci linked to *Mycobacterium tuberculosis* resistance to bedaquiline, delamanid and pretomanid

Paula J. Gómez-González¹, Joao Perdigo², Pedro Gomes², Zully M. Puyen³, David Santos-Lazaro³, Gary Napier¹, Martin L. Hibberd¹, Miguel Viveiros⁴, Isabel Portugal², Susana Campino¹, Jody E. Phelan¹ & Taane G. Clark^{1,4,5}✉

Tuberculosis (TB), caused by *Mycobacterium tuberculosis*, is one of the deadliest infectious diseases worldwide. Multidrug and extensively drug-resistant strains are making disease control difficult, and exhausting treatment options. New anti-TB drugs bedaquiline (BDQ), delamanid (DLM) and pretomanid (PTM) have been approved for the treatment of multi-drug resistant TB, but there is increasing resistance to them. Nine genetic loci strongly linked to resistance have been identified (*mmpR5*, *atpE*, and *pepQ* for BDQ; *ddn*, *fgd1*, *fbiA*, *fbiB*, *fbiC*, and *fbiD* for DLM/PTM). Here we investigated the genetic diversity of these loci across >33,000 *M. tuberculosis* isolates. In addition, epistatic mutations in *mmpL5-mmpS5* as well as variants in *ndh*, implicated for DLM/PTM resistance in *M. smegmatis*, were explored. Our analysis revealed 1,227 variants across the nine genes, with the majority (78%) present in isolates collected prior to the roll-out of BDQ and DLM/PTM. We identified phylogenetically-related mutations, which are unlikely to be resistance associated, but also high-impact variants such as frameshifts (e.g. in *mmpR5*, *ddn*) with likely functional effects, as well as non-synonymous mutations predominantly in MDR-/XDR-TB strains with predicted protein destabilising effects. Overall, our work provides a comprehensive mutational catalogue for BDQ and DLM/PTM associated genes, which will assist with establishing associations with phenotypic resistance; thereby, improving the understanding of the causative mechanisms of resistance for these drugs, leading to better treatment outcomes.

Mycobacterium tuberculosis (*Mtb*) remains one of the deadliest single infectious agent, leading to 10 million human tuberculosis (TB) cases and 1.4 million associated deaths in 2019¹. Most TB cases are found in Asia, Africa, and Western Pacific regions. Drug resistance is one of the major threats to control the disease, especially *Mtb* resistant to rifampicin (RR-TB), and multi-drug resistant (MDR-TB; isoniazid and rifampicin). MDR-TB with further resistance to at least one fluoroquinolone and second-line injectable drug has been defined as extensively drug resistant *Mtb* (XDR-TB), but the definition has recently changed, in part due to a need to include bedaquiline (BDQ) and linezolid (LNZ)². More than 3% of new TB cases are RR- or MDR-TB, and among MDR-TB, more than 6% are XDR-TB. In 2019, approximately half a million people developed MDR-TB, and ~12,000 patients had XDR-TB¹.

BDQ, delamanid (DLM) and pretomanid (PTM) comprise the most recent additions to the anti-TB drug armamentarium and therefore constitute alternative effective drugs for resistant cases³. BDQ has been in use since 2013¹, and is a diarylquinoline that inhibits the proton pump ATP synthase, more specifically, the subunit c

¹Department of Infection Biology, Faculty of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine, Keppel Street, London, UK. ²Faculdade de Farmácia, Universidade de Lisboa, Lisbon, Portugal. ³Instituto Nacional de Salud, Lima, Peru. ⁴Global Health and Tropical Medicine, GHTM, Instituto de Higiene E Medicina Tropical, IHMT, Universidade Nova de Lisboa, Lisbon, Portugal. ⁵Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London WC1E 7HT, UK. ✉email: taane.clark@lshtm.ac.uk

encoded by the *atpE* gene (*Rv1305*)⁴. DLM is a nitro-dihydro-imidazooxazole derivative that targets the synthesis of the cell wall mycolic acids. It is a pro-drug that is activated by the enzyme deazaflavin dependent nitroreductase encoded by the *ddn* gene (*Rv3547*)⁵, which requires the F₄₂₀ coenzyme system for its activity. DLM started to be used to treat MDR-TB patients in 2014⁶. By the end of 2018, more than fifty countries were using BDQ and DLM. However, resistance to BDQ and DLM emerged quickly, with reports of resistance *in vitro*^{7,8} and then clinically^{9,10}, as well as reported cross-resistance between BDQ and the repurposed antimycobacterial drug clofazimine (CFZ)¹¹. There are fears for wider emergence and spread of drug-resistant *Mtb* to these new drugs, particularly among MDR-/XDR-TB strains, which will impose new obstacles that threaten global TB control. PTM was introduced in 2019 in a joint regimen with BDQ and LNZ¹.

Acquired drug resistance in *Mtb* is almost exclusively due to spontaneous mutations, including single nucleotide polymorphisms (SNPs) and insertions and deletions (indels), in genes coding for drug-targets or drug-converting enzymes¹². Acquisition and accumulation of resistance conferring mutations sometimes entails fitness loss, which triggers putative compensatory mechanisms^{13,14}. Drug resistance can be determined by phenotypic or genotypic methods, and new mutations are being found using genome-wide association and convergent evolution studies¹⁵. Putative molecular markers of resistance to BDQ include mutations in the drug target *atpE*, and off-target mutations in *mmpR5* (*Rv0678*) and *pepQ* (*Rv2535c*). The *mmpR5* gene encodes for a transcriptional repressor of the MmpS5-MmpL5 efflux pump, whose upregulation has been associated with BDQ resistance⁸. Loss of function of MmpR5 leads to the de-repression of this efflux pump, thereby mediating increased values of minimum inhibitory concentrations (MICs) for BDQ. Some mutations in *mmpR5* have been observed in isolates that pre-date the introduction of BDQ, and may be linked to the use of CFZ or other azoles for fungal infections^{15,16}. Epistatic interactions through loss of function mutations in *mmpL5* that counteract the effect of *mmpR5* mutations have been suggested^{17,18}. Resistance caused by mutations in the peptidase encoded by *pepQ* has also been reported with increased BDQ MIC values¹⁹, but the exact mechanism is unclear. Other off-target genes investigated for BDQ resistance include *Rv1979c*, *atpB* and *ppsC*, but only *mmpR5* and *pepQ* have strong experimental evidence of developing mutations under drug exposure *in vitro* or *in vivo*^{19,20}.

As pro-drugs, the nitroimidazoles DLM and PTM require activation by the deazaflavin (F₄₂₀)-dependent nitroreductase Ddn. Mutations in the essential genes required for the F₄₂₀ cofactor biosynthesis and recycling, including *ddn*, *fgd1*, *fbiA*, *fbiB*, *fbiC*, and *fbiD*, are putative resistance markers that directly hamper DLM/PTM activation or, work indirectly through F₄₂₀ depletion^{5,21–23}. Important residues for the interaction of Ddn-PTM are known, which may differ from those involved in Ddn-DLM activation²⁴. The role of Fgd1 as a F₄₂₀-dependent glucose-6-phosphate dehydrogenase is to reduce F₄₂₀, which is essential for the correct performance of Ddn. FbiA, FbiB and FbiC are also proteins involved in the activation of DLM and PTM through their role in the synthesis of F₄₂₀ cofactor. Mutations in these 3 genes have been shown to alter the production of F₄₂₀²². Similarly, it has been recently demonstrated the essential role of FbiD for the biosynthesis of F₄₂₀ and thereby its participation in DLM and PTM resistance²³. The contribution of *ndh*, a NADH dehydrogenase, in isoniazid and ethionamide resistance involves retaining an appropriate NADH/NAD⁺ ratio that enables the formation of adducts with NAD⁺, necessary for their activity²⁵. The same mechanism of adduct formation has been recently suggested for DLM, with evidence of increased MIC values in *ndh* mutants in a *M. smegmatis* model²⁶.

For phenotypic derived resistance, BDQ and DLM drug susceptibility testing use provisional critical concentration values defined by the WHO or the European Committee on Antimicrobial Susceptibility Testing (EUCAST), where the thresholds are highly variable and/or limited²⁷. There is currently no established MIC cut-offs for PTM and BDQ by the EUCAST reference method, but ongoing work is attempting to establish these^{28,29}. Studies involving genetic-phenotypic functional analysis for resistance have been of limited sample size, and those looking at candidate region genomic variation have considered small numbers of populations. To provide a global view, we perform an analysis of nine candidate genes and their mutations associated with BDQ (*atpE*, *mmpR5* and *pepQ*) and DLM/PTM (*ddn*, *fgd1*, *fbiA*, *fbiB*, *fbiC* and *fbiD*) resistance in > 33,000 clinical *Mtb* isolates, sourced from all WHO regions, and with whole genome sequencing data. In addition, we investigated potential epistatic mutations in *mmpL5* and *mmpS5*, as well as variants in *ndh*. Our goal was to establish the frequency of putative resistance markers across geographical regions and, where possible, rule in or out putative mutations based on source population and date of DLM and BDQ roll-out, individual drug-resistance profiles and phenotypic data, and application of phylogenetic methods and protein structural modelling. *In lieu* of large-scale studies with phenotypic susceptibility testing, we present evidence for mutations involved in BDQ and DLM/PTM putative genotypic resistance, where possible validated by quantitative data on resistance levels. Ultimately, we aim to present a variant catalogue with important mutations that could potentially reduce BDQ, DLM and PTM drug effectiveness globally.

Results

The samples. Our study consists of 33,675 publicly available *Mtb* isolates with complete whole-genome sequencing data, collected between 1991 and 2018 across 114 countries³⁰. These strains represent the main *Mtb* complex lineages, with the majority in lineage 4 (52%), followed by lineages 2 (25%), 3 (11%) and 1 (10%). Using genotypic resistance prediction³¹, the majority of strains (65%) were pan-susceptible, while 22% were at least MDR-TB, with the remainder being non-MDR but resistant to at least one drug (termed “other drug resistance”) (S1 Table). The vast majority (91%) of isolates were collected before the roll out of BDQ and DLM, and we have used the definition of XDR-TB before the recent WHO update. The most represented geographical areas were Europe and Central Asia, followed by Sub-Saharan Africa, East Asia, and Pacific regions. The highest proportion of MDR-TB strains were from the Latin American and Caribbean region (63%) (S1 Table).

Gene	Drug	Gene SNPs [Indels,fs*]	Prom. SNPs [Indels]	Total analysed [# known**]	# samples with 1 [>1] mutations	Lineages	Ave. mut. Susc. samples	Ave. mut. MDR samples	Ave. mut. XDR samples	Ave. mut. DR samples	Diversity $\times 10^{-5***}$
<i>atpE</i>	BDQ	15 [1,0]	5 [5]	26 [1]	48 [1]	1–4, <i>bov</i>	0.002	0	0	0.002	0.87
<i>mmpR5</i>	BDQ	116 [25, 29]	14 [4]	163 [38]	555 [17]	1–6, <i>bov</i>	0.008	0.040	0.079	0.017	3.2
<i>pepQ</i>	BDQ	117 [2, 3]	0 [0]	120 [0]	482 [4]	1–6	0.018	0.010	0.004	0.009	2.4
<i>fgd1</i>	DLM/PTM	118 [4, 9]	11 [1]	139 [4]	4229 [35]	1–7, <i>bov</i>	0.141	0.095	0.075	0.124	23
<i>ddn</i>	DLM/PTM	86 [16, 27]	18 [2]	132 [31]	743 [21]	1–5	0.025	0.019	0.015	0.023	7.6
<i>fbiA</i>	DLM/PTM	113 [2, 3]	3 [0]	119 [4]	991 [0]	1–5, <i>bov</i>	0.037	0.016	0.007	0.019	5.5
<i>fbiB</i>	DLM/PTM	135 [1]	0 [0]	136 [3]	851 [3]	1–6, <i>bov</i>	0.025	0.022	0.012	0.033	3.6
<i>fbiC</i>	DLM/PTM	280 [9, 17]	26 [4]	326 [4]	2413 [45]	1–6, <i>bov</i>	0.079	0.052	0.058	0.091	3.8
<i>fbiD</i>	DLM/PTM	57 [0,0]	9 [0]	66 [0]	223 [2]	1–4, <i>bov</i>	0.008	0.004	0.004	0.007	1.8

Table 1. Number of variants per analysed gene across the 33,675 isolates, with the average number of mutations per sample and by resistance profile. Indels = insertions and deletions; DLM = Delamanid; PTM = Pretomanid; BDQ = Bedaquiline; Prom. = promoter; Susc. = Susceptible; DR = Other drug resistance; fs = frame shifts; *bov* = *M. bovis*; * number of indels that lead to frameshifts; ** see S3 Table; *** Nei's Pi nucleotide diversity per site (only non-synonymous SNPs considered).

Mutational diversity and prevalence across resistance associated genes. Across the three BDQ resistance candidate genes (*atpE*, *mmpR5* and *pepQ*), we observed 467 unique variants, and focused the analysis on the 309 non-synonymous or indel mutations, distributed across 1,085 (3%) isolates representing all geographical regions and lineages (except lineage 7) (Table 1, S1 Table). Synonymous mutations changing the start codon of *mmpR5* or *pepQ* were not identified. Co-occurrence of multiple mutations in the same candidate gene in an isolate was rare (2% of isolates (n = 22) with > 1 mutation; maximum of 3). Similarly, only 2% (n = 22) of isolates had a mutation in 2 of the 3 BDQ candidate genes (Fig. 1). Most mutations were found in *mmpR5* (n = 163, 53%) and *pepQ* (n = 120, 39%) loci, and the majority of indels (29/33) were present in the former and lead to a high proportion of frameshifts (25/29) (Table 1). Nucleotide diversity in the coding regions of *atpE* was slightly lower than in *mmpR5* and *pepQ* (S2 Figure). The distribution of variants along the *mmpR5* and *pepQ* genes was broadly uniform, but the *atpE* promoter region has a high density of mutations (n = 10, 39% of total mutations in *atpE*), especially between 28 and 41 bp upstream (n = 8, 80% of promoter mutations in *atpE*) (Table 1; S2 Table). In the case of *mmpR5*, there was a greater risk of mutations in MDR-/XDR-TB isolates (adjusted odds ratio > 3.7; $P < 0.0001$), as well as those sourced after year 2014 (adjusted odds ratio 2.574, $P = 0.002$) (Table 1, S2 Table). Most of the BDQ candidate variants (n = 180, 58%) were unique mutations, present in single isolates across the whole data set. Only 17 (6%) of the mutations found in BDQ candidate genes occurred in 10 or more samples (Fig. 1). Of 144 mutations identified previously as associated with increments in MICs (S3 Table), 33 (23%) were identified in our ~33,000 *Mtb* dataset.

Across the six DLM/PTM candidate genes (*ddn*, *fgd1*, *fbiA*, *fbiB*, *fbiC* and *fbiD*), we observed 1,595 unique mutations, and focused the analysis on 918 (58%) non-synonymous or indel variants found within 8,622 isolates (26% of the samples, all lineages present) (Table 1). Synonymous mutations changing the start codon of the genes starting with amino acids V or L were not identified among our isolates. The *fbiC* gene, which is the largest of the loci considered, accounted for the highest number of different mutations (n = 326, 36% of the total variants identified), with a high density of variants in the promoter region compared to the rest of the coding area (S3 Figure). However, *fgd1* was the most polymorphic gene per isolate, accounting for the higher nucleotide diversity when compared to the other genes (Table 1). The *ddn*, *fgd1* and *fbiD* genes also harboured more than 8% of their variants in the intergenic promoter region. Both *ddn* and *fbiC* harboured a higher number of indels (44/57) along the whole coding region, compared to the other genes (13/57), where more than half (56%) led to a frameshift. For the six genes, the average number of mutations per sample among susceptible isolates was higher than in MDR- or XDR-TB, which could be due to a higher representation of the different sub-lineages among susceptible samples, or the effects of clonality. For the *ddn* gene, there was a marginally greater risk of mutations in MDR-/XDR-TB isolates (adjusted odds ratio > 1.5; $P < 0.02$; S2 Table). Co-occurrence of variants in genes in the same sample was rare (83 (1%) samples with > 1 mutation; maximum of 3 mutations) (Fig. 1). Likewise, only 828 (10%) isolates with mutations in DLM/PTM candidate genes had at least one mutation in two or more of the genes considered (Fig. 1), where the most prevalent combination of mutations involved *fbiC* with either *ddn* or *fgd1*. A total of 117 (13%) mutations were present at higher frequencies (> 5 samples; note, 62 (7%) mutations with > 10 samples). Of 198 mutations reported previously as associated with some degree of resistance (S3 Table), only 26 associated with DLM or PTM were in our dataset. Co-occurrence of mutations in at least one BDQ and one DLM/PTM candidate gene was also rare, with only less than 2% (n = 153/9,538) of samples harbouring these variants.

Eight mutations in candidate genes (7 DLM/PTM, 1 BDQ) were considered as phylogenetic deep branching variants at high frequency within single sub-lineages (> 50% allele frequency) (S4 Table). Isolates harbouring each of these mutations were collected from > 10 different countries and had a high pairwise SNP distance (> 200). All eight mutations were mostly found in susceptible samples and, where available, date of collection pre-dated the introduction of BDQ and DLM. Seven of these mutations have been previously reported as phylogenetically-related or -informative¹⁷. These strain specific mutations have been incorporated within the TB-Profiler tool³¹.

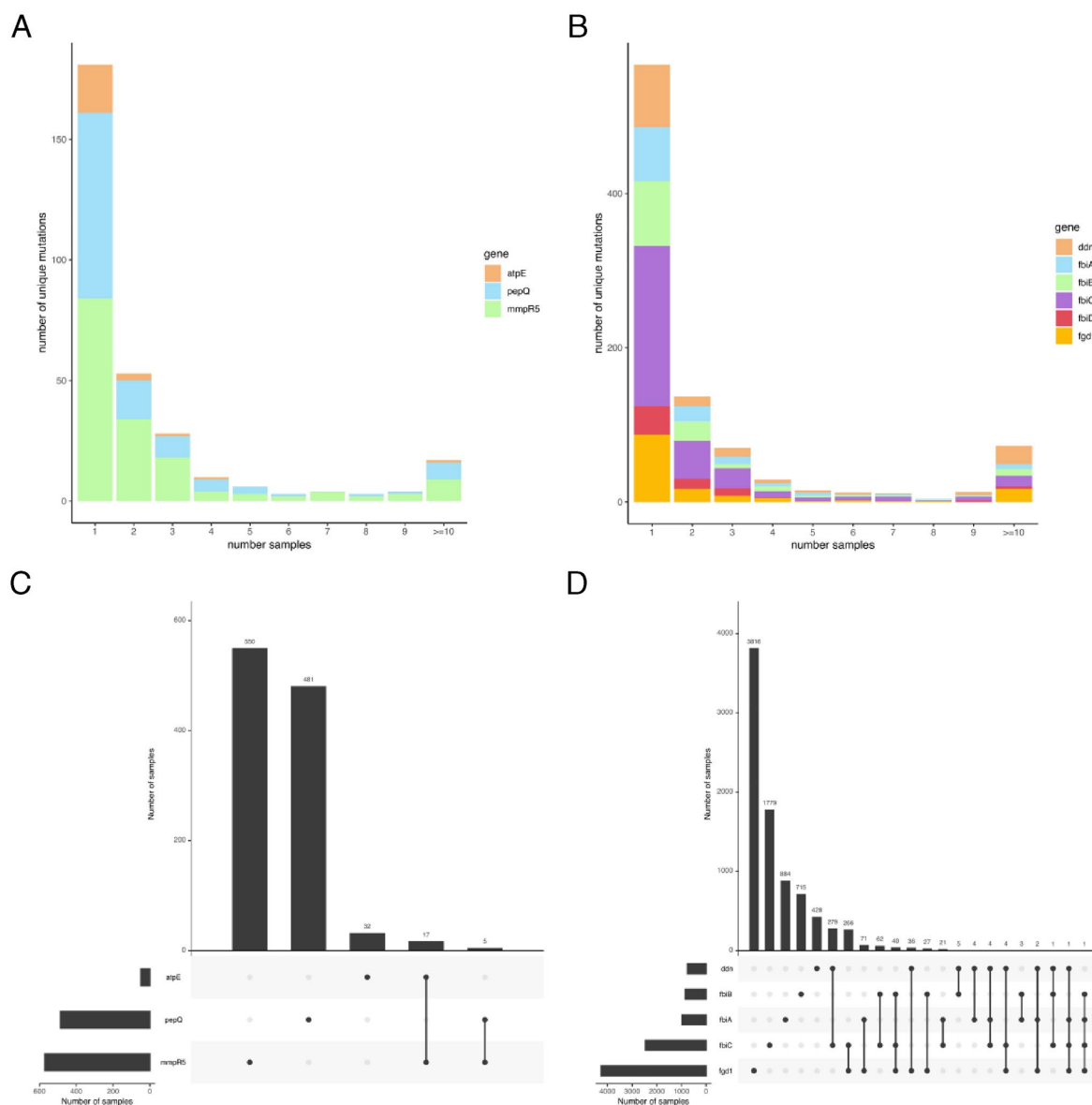


Figure 1. (A), (B) Frequency of mutations identified across data set. The vertical axis is the number of mutations that are found in 1 to 10 or more isolates (horizontal axis). Colours represent the different genes, each bar showing the distribution of those mutations in the candidate genes for each drug (A = Bedaquiline (BDQ), B = Delamanid (DLM)/Pretomanid (PTM)). (C), (D) Intersection of mutations in the different genes by sample. Bars represent the number of samples that hold mutations in each gene, or combination of them (horizontal bars show total samples with mutations in each gene); C = BDQ, D = DLM/PTM.

Nonetheless, the 326 (27%) mutations detected in > 1 isolates and a single homoplasic distribution may denote potentially advantageous polymorphisms with impact at the phenotypic level.

Diversity and phylogenetic distribution of BDQ-associated variants. Twenty-two of the 37 most frequent mutations (> 5 isolates, S5 Table) were present in isolates in a single monophyletic cluster. Two mutations (*pepQ* T354A, *mmpR5* M146T) were present in isolates within potential transmission chains (maximum of 11 SNPs difference) (Fig. 2). The majority (13/15) of mutations that showed evidence of convergent evolution were observed in *mmpR5*, of which 8 have been previously associated with increased MICs (Table 2, S3 Table), including 6 variants in high frequency (> 80%) in MDR-/XDR-TB clinical isolates. Two mutations in *mmpR5* (-11C>A, D5G) not linked to in vitro resistance (S3 Table) were also found in high frequency among our isolates, where intergenic -11C>A was prevalent in MDR-/XDR-TB isolates. The -11C>A mutation has been

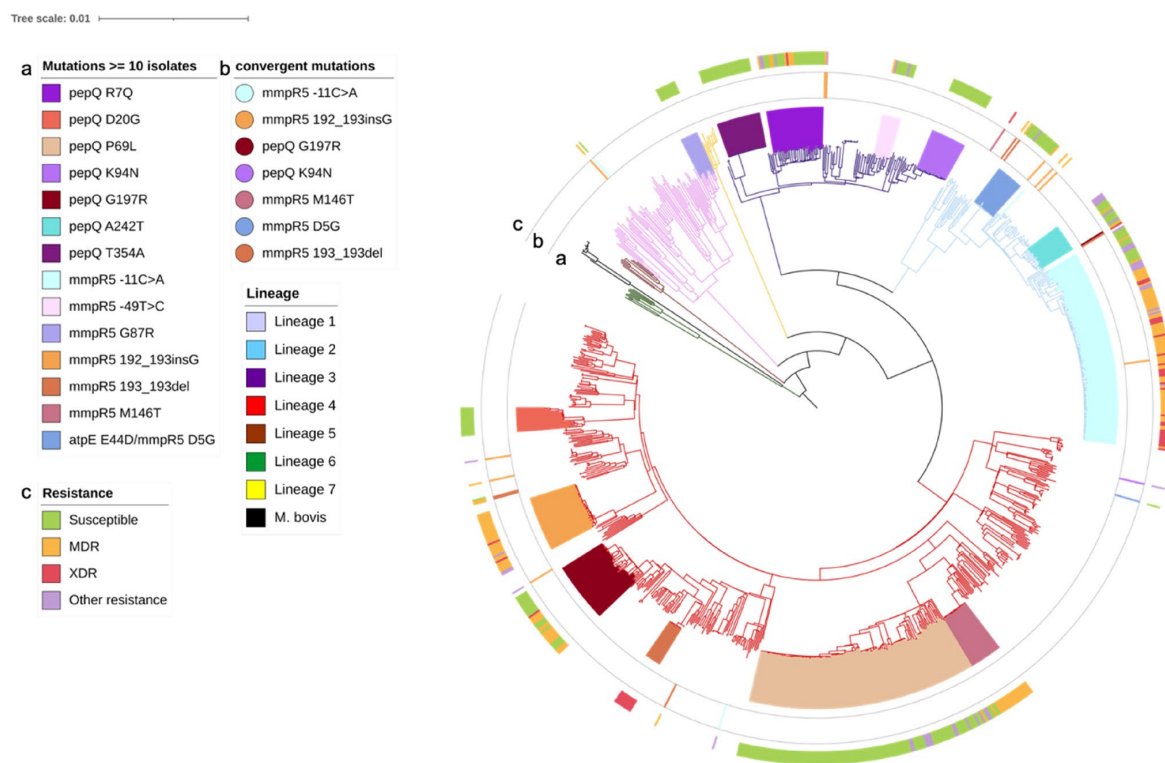


Figure 2. Phylogenetic tree of high frequency (≥ 10 isolates) mutations in bedaquiline candidate genes. The outer track (c) shows the resistance phenotype; the second track (b) shows the convergent mutations that have arisen in more than one clade; the third track (a) shows the clades formed by isolates harbouring the same phylogenetic-related mutations. Branches are coloured by lineage as per legend.

reported in hyper-susceptible strains¹⁵. The two other high frequency mutations (2/15) in multiple lineages were found to occur in the *pepQ* (G197R, K94N) gene, and predominantly in susceptible strains with one (G197R) predicted to have functional effects by Provean and SNAP2 scores.

atpE and pepQ. Most mutations in *atpE* (20/26; 77%) were found in single isolates (S7 Table), and those with higher frequencies did not show evidence of convergent evolution, being part of single clades (S5 Table, Fig. 2). Of twenty-five novel mutations found in *atpE*, 15 were non-synonymous SNPs, of which 9 were predicted to confer resistance using SUSPECT-BDQ software³² (S5 Table, S7 Table). Only the I66V mutation is present in residues involved in BDQ-*atpE* interactions (S4 Figure). The E44D mutation, predicted as conferring resistance, was present in 17 mostly pan-susceptible Beijing (lineage 2.2.1) isolates.

The 120 novel mutations identified in *pepQ* included 117 non-synonymous SNPs (S5 Table) and 3 indels, 2 of them leading to frameshifts found in single isolates (S7 Table, S5 Figure). These frameshifts are likely to be involved in the functional loss of *pepQ*, consistent with others that have been found (see S3 Table). In the absence of a crystal structure of PepQ, SNAP2 and Provean scores revealed 9 mutations with a potential functional effect (S5 Table), and 3 were present in MDR-/XDR-TB isolates.

mmpR5 mutations. Of the 163 mutations (116 non-synonymous SNPs, 29 indels and 18 promoter variants) found in *mmpR5*, 32 and 14 have been previously associated with MIC incrementation or no change, respectively (S3 Table). A high density of variants ($n=64$) in the DNA binding domain was observed, including 14 frameshifts (S6 Figure). In addition, 3 SNPs were translated into stop codons (E13*, W42* and R156*; S5 Table; S7 Table), which are likely to alter the protein function. Three frameshifts (192_193insG, 193_193del, 141_142insC) have a high number of independent occurrences (range: 5–11) in a phylogenetic tree (Table 2, Fig. 2), all previously associated with higher MICs in vitro to BDQ³³. The 192_193 indel (sometimes denoted as I67fs), involving a premature stop codon, appears in 44 isolates through 10 independent acquisitions. The largest subclade (34 isolates) consists of resistant lineage 4 strains, with all except one sourced from Peru and collected between years 2009 and 2012, prior to the introduction of BDQ in that country (Table 2; S7 Figure). A potential epistatic effect involving the 605_605 deletion in *mmpL5* was found in 33 of these isolates, confirming recent work^{17,18}. In addition, two isolates from Malawi belonging to lineage 4.3.4.2.1 with a pan-susceptible profile had the beginning and most of *mmpR5* deleted (778866_779429del), which could have similar epistatic effects.

Mutation	Gene	Freq	Sub-lineage (# isolates)	# sub-lineages	Max SNP dist.*	# Independent Occurrences	Susc. %	MDR or XDR %	Pre-2014%**	Functional support***
-11C>A	<i>mmpR5</i>	124	2.2.1(122); 4.3.2.1(1); 1.1.1(1)	3	207	3	12.1	76.6	93.1	-
192_193insG (I67fs)	<i>mmpR5</i>	44	4(34); 2.2.1(4); 3(2); 4.9(1); 4.8(1); 4.5(1); 1.1.1(1)	7	60	10	0	86.4	100	-
G197R	<i>pepQ</i>	38	4.3.4.1(37); 2.2.1(1)	2	168	2	52.6	47.4	72.2	S,P
K94N	<i>pepQ</i>	23	3.1.1(22); 4.1.2(1)	2	24	2	95.7	0	100	-
M146T	<i>mmpR5</i>	21	4.4.1.1(20); 2.2.2(1)	2	11	2	0	100	-	S,M
D5G	<i>mmpR5</i>	18	2.2.1(17); 4.1.2.1(1)	2	33	2	94.4	0	75.0	-
193_193del (I67fs)	<i>mmpR5</i>	16	4.3.4.2(10); 2.2.1(3); 4.7(2); 4.3.3.1(1)	4	17	5	0	100	83.3	-
141_142insC	<i>mmpR5</i>	15	2.2*(8); 4.1.2*(2); 4.3*(2); 4.4.1.1(1); 3(2)	8	-	11	6.7	86.7	85.7	-
V20A	<i>mmpR5</i>	10	4.1.2.1(8); 4.3.2.1(1); 2.2.1(1)	3	23	3	90	10	83.3	M
L117R	<i>mmpR5</i>	9	3(5); 4.3.4.2(2); 4.2.2(1); 4.1(1)	4	98	5	44.4	44.4	100	S
L32S	<i>mmpR5</i>	8	2.2.1(8)	1	21	3	0	87.5	50	S,M
G121R	<i>mmpR5</i>	7	2.2.2(5); 3(1); 4.4.1.1(1)	3	4	3	0	100	100	S,P
D141H	<i>mmpR5</i>	7	2.2.1(6); 1.1.3(1)	2	130	2	14.3	57.1	100	S,P
R90C	<i>mmpR5</i>	7	2.2.1(6); 4.1.1.3(1)	2	24	4	85.7	0	50	-
N98D	<i>mmpR5</i>	5	4.1.2.1(2); 4.4.1.1(2); 2.2.1(1)	3	5	3	0	80	100	-

Table 2. Mutations in bedaquiline candidate genes occurring in at least 5 samples and more than one independent clade. Sub-lineages: + = more than 1 sub-lineage; # = number; * Maximum SNP distance calculated in clades of ≥ 5 isolates; Drug resistance (%): Susc. = Susceptible; ** % of number of samples pre-2014/total number of samples with available collection date; *** Functional support: S = snap2 score ≥ 50 ; P = Provean Score ≤ -4 ; M = mCSM predicted stability change ($\Delta\Delta G$) below -2 . Mutations associated with increased MIC for BDQ in previous studies in **bold**; mutations associated with susceptibility to BDQ underlined (see S3 Table).

The *mmpR5* 193_193 deletion (I67fs) was present in XDR-TB isolates from Portugal (lineage 4.3.4.2; n = 10) and present in the phylogenetic tree an additional 4 times independently in modern strains within different sub-lineages (Table 2, Fig. 2). To investigate the contribution of this mutation to BDQ resistance levels, we screened for it in a recently published dataset focused on the evolutionary history of MDR-TB in Portugal³⁴. One clinical isolate (MTB1) was available with a BDQ MIC value of 0.25 mg/L, which is at least 6- to 8-fold higher in comparison to wild-type strains, including one isolate from the same phylogenetic clade and *M. tuberculosis* H37Rv (ATCC 27,294) (S9 Table). CFZ MICs determined in parallel showed a 4- to 6-fold increase for the *mmpR5* mutant strain, which corroborates the high impact of this variant on MmpR5 function, and is consistent with previous findings in South Africa¹⁶. Further, our analysis confirmed the presence of the *mmpR5* M146T mutation within a transmission cluster in Eswatini³⁵, as well as in an independent XDR-TB (lineage 2.2.2) strain (Table 2, Fig. 2). Twenty-one of the remaining SNPs in *mmpR5*, including high frequency D5G, V20A, L117R, L32S, G121R, D141H, R90C and N98D, were in the same residue where mutations associated with increments in MIC have been observed; however, mutations V20A and D141H have associated MIC values within a susceptibility range³⁶.

Mutational diversity in Delamanid and Pretomanid associated genes. Thirty four of the 117 mutations were found in > 4 isolates and occurred in at least two sub-lineages, appearing up to 4 times in the phylogenetic tree (Table 3, Fig. 3, S6 Table). Eleven of the mutations (*fgd1* K270M, K296E; *ddn* P45L, G81S, G34R, R72W, D113N; *fbtB* D90N, K448R; *fbtC* T273A, W678G) have been identified previously in susceptible samples (S3 Table). The *ddn* L49P mutation, found to be associated with an increment in DLM and PTM MIC²⁴, was identified in Beijing strains occurring in genomic clusters from Vietnam, the Netherlands and Mexico, highlighting an ability to disseminate with low fitness impact at an epidemiological level. These isolates were mostly assessed genotypically as non-MDR, and all pre-dated the introduction of DLM as a TB treatment (Table 3). L49 is involved in activation of both DLM and PTM, and L49P is thought to confer cross-resistance to both drugs²⁴.

Mutation	Gene	Freq	Sub-lineage (# isolates)	# sub-lineages	Max SNP distance*	# Independent Occurrences	Susc. %	MDR or XDR %	Pre-2014%**	Functional Support***
K270M	<i>fgd1</i>	3136	4.1.2+ (3135); 2.2.1(1)	3	1329	2	70.1	18.1	84.2	-
-32A>G	<i>fbiC</i>	639	5, 6, <i>Bov</i> (634); 2.2.1(2); 4.3.3(1); 4.2(1); 4.9(1)	7	3264	5	60.1	8.3	63.1	-
T273A	<i>fbiC</i>	626	4.8(625); 1.1.1(1)	2	330	2	97.9	0.3	93.6	-
K448R	<i>fbiB</i>	293	3(293)	1	496	3	57.7	30.0	51.1	-
D113N	<i>ddn</i>	267	5(264); 2.2.1(3)	2	1402	2	70.7	15.4	91.7	-
K296E	<i>fgd1</i>	162	6(161); 4.1.2.1(1)	2	933	2	87.0	3.7	85.7	-
I208V	<i>fbiA</i>	122	4.1.2(121); 4.1.2.1(1)	2	524	2	70.5	11.5	96.9	-
W678G	<i>fbiC</i>	96	4.3.3(88); 1.1.1(8)	2	87	2	8.3	81.3	90.9	P
I128V	<i>fbiC</i>	79	2.2.1(79)	0	91	2	0	81.0	100	-
R72W	<i>ddn</i>	75	1.1.2(75)	1	345	2	76.0	10.7	70.2	S,P
A31T	<i>fbiB</i>	71	2.2.1(70); 2.2.2(1)	1	238	3	54.9	9.9	100	-
G34R	<i>ddn</i>	47	4.3.2(44); 4.3.4.2(3)	2	147	2	89.3	8.5	0.0	S,P
-11G>A	<i>fbiC</i>	37	4.1.2.1(31); 4.1.1.3(3); 6(2); 4.4.2(1)	4	244	4	56.8	16.2	100	-
-14G>GA	<i>fbiC</i>	34	2.2.1(25); 4.3.4.2.1(9)	2	30	2	26.5	73.5	94.7	-
G81S	<i>ddn</i>	21	2.2.2(12); 2.1(9)	2	277	2	33.3	52.4	100	S,P
I49P	<i>ddn</i>	21	2.2.1.1(21)	1	226	3	57.1	9.5	94.4	S,P
G139R	<i>fbiA</i>	18	2.2.1(17); 1.1.2(1)	2	30	2	94.4	0	75.0	P
W20*	<i>ddn</i>	17	4.5(11); 5(6)	2	241	2	100	0	75.0	-
K296R	<i>fgd1</i>	16	4.1.2.1(12); 4.8(3); 4.4.1.1(1)	3	75	3	18.8	31.3	37.5	-
D90N	<i>fbiB</i>	14	3(14)	1	419	2	50.0	14.3	14.3	-
R265Q	<i>fbiB</i>	13	2.2.1(12); 1.1.2(1)	2	77	3	30.8	0	100	-
A178T	<i>fbiA</i>	10	1.2.1(8); 4.5(1); 3(1)	3	141	4	70	0	66.7	-
-43G>A	<i>ddn</i>	9	5(4); 4.2.1(3); 2.2.1(2)	3	244	3	77.8	22.2	100	-
P131L	<i>ddn</i>	9	4.8(8); 4.3.4.2.1(1)	2	50	2	88.9	0	100	S,P
G655S	<i>fbiC</i>	9	2.2.1(8); 4.1.2(1)	2	24	2	33.3	0	100	-
R247W	<i>fgd1</i>	8	3(7); 4.5(1)	2	62	2	100	0	50	P
V348I	<i>fbiB</i>	8	2.2.2(7); 4.1(1)	2	17	2	100	0	100	-
R304Q	<i>fbiA</i>	8	3(8)	2	203	2	87.5	0	50	-
G325S	<i>fbiB</i>	7	2.2.1(5); 4.9(1); 4.1.2.1(1)	3	63	3	100	0	83.3	-
P182L	<i>fbiB</i>	6	4.3.4.2.1(3); 6(3)	2	320	2	66.7	16.7	100	-
M93I	<i>fgd1</i>	6	4.9(3); 4.1.2.1(2); 2.2.1(1)	3	3	3	83.3	0	-	-
P45L	<i>ddn</i>	5	4.4.1.1(3); 3(1); 1.1.1(1)	3	63	3	80	0	100	S,P
I326F	<i>fbiB</i>	5	4.6.1.1(2); 4.1.2*(2); 4.8(1)	4	-	4	100	0	-	-
T455A	<i>fbiC</i>	5	3(4); 1.1.1(1)	2	236	2	80	0	100	-

Table 3. Mutations in delamanid candidate genes occurring in at least 5 samples and more than one independent clade. Sub-lineages: + = more than 1 sub-lineage; # = number; * Maximum SNP distance calculated in clades of ≥ 5 isolates; Drug resistance (%): Susc. = Susceptible; ** % of number of samples pre-2014/total number of samples with available collection date; *** Functional support: S = snap2 score ≥ 50 ; P = Provean Score ≤ -4 ; M = mCSM predicted stability change ($\Delta\Delta G$) below -2 ; Mutations associated with increased MIC for DLM/PTM in previous studies in bold; mutations associated with susceptibility to DLM/PTM underlined (see S3 Table).

ddn and fgd1 mutations. Mutations identified in *ddn* included 86 non-synonymous SNPs, 23 small indels, 4 large deletions and 20 mutations in the promoter region. Of these, 10 and 30 have been previously

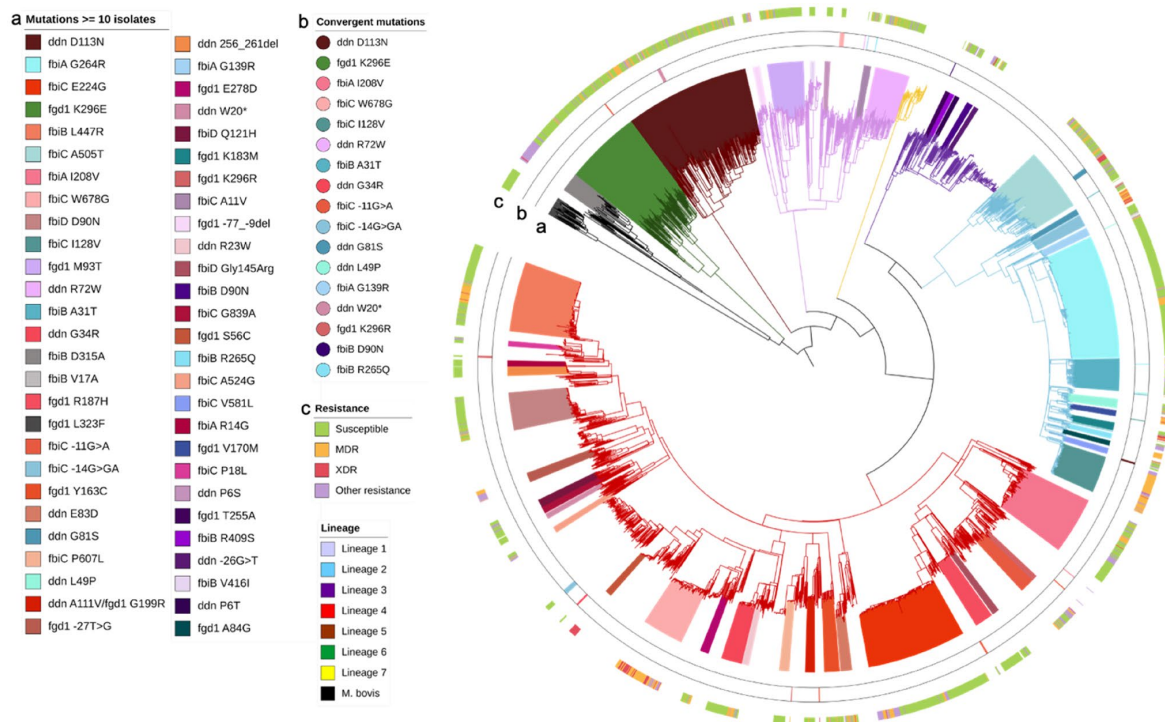


Figure 3. Phylogenetic tree of high frequency mutations (≥ 10 isolates) in delamanid and pretomanid candidate genes (*fgd1* K270M and R64S, *fbiC* -32A>G and T273A, *fbiA* T302M and *fbiB* K448R found in >290 isolates not represented). Clades formed by isolates harbouring the same mutations are differentiated by colour. The outer (c) track shows the resistance phenotype; the second track (b) shows the convergent mutations that have arisen in more than one clade; the third track (a) shows the clades formed by isolates harbouring the same phylogenetic-related mutations. Branches are coloured by lineage as per legend.

associated with DLM/PTM resistance and susceptibility, respectively (S3 Table). In general, *ddn* amino acid changes were dispersed along the coding region (S8 Figure). Twenty-seven of the (100) novel mutations were indels with 16 causing frameshifts along the coding region (S6 Table; S7 Table; S8 Figure). These indels included 4 large deletions (> 100 bp), identified in low frequencies in MDR-TB isolates (except for 1 susceptible isolate) sourced from China in 2007, before the introduction of DLM as a treatment. Most frameshifts and large deletions were identified in single isolates. Moreover, 6 amino acid changes leading to stop codons and the resultant truncated proteins were identified, including 3 reported (W88*, W27* and Q58*; S3 Table) and 3 unreported variants (W20*, W139*, Y133*). W20* was present in clades consisting of lineage 4.5 (n = 11) and 5 (n = 6) isolates (Fig. 3), where all 16 samples were pan-susceptible. The maximum pairwise SNP difference between lineage 4.5 isolates harbouring *ddn* W20* was 241, suggesting that the variant established itself in that population some time ago. The W88* mutation, which has in vitro evidence of resistance to DLM (S3 Table), appeared within a potential transmission cluster of Beijing MDR-/XDR-TB isolates. Other SNPs known to cause an increment in DLM/PTM MIC (M1T, W88R, Y65S and G53D) were found in 3 or less isolates.

Of the 139 mutations identified in the *fgd1* gene (S9 Figure), six SNPs have been described previously, including two phylogenetically-related (K270M lineage 4.1.2; K296E lineage 6) (S4 Table) with no association with resistance, and two known to increase PTM MIC (G71D and E230K) (see S3 Table). Four frameshifts with disruptive functional consequences for the protein were identified in low frequencies. One isolate was found to harbour K259E, which is a residue involved in F₄₂₀ binding³⁷. Of the other mutations, only F79S had a predicted destabilizing effect on the protein (S7 Table).

***fbiA*, *fbiB*, *fbiC* and *fbiD* mutations.** In total, 119, 136 and 326 mutations were identified in *fbiA*, *fbiB* and *fbiC* respectively (S6 Table; S7 Table). Several mutations that are known to increase DLM/PTM MICs in vitro (S3 Table) were identified (*fbiA* K2E, V154I, I208V, I209V, K250*, S126P, R304Q; *fbiB* P361A; *fbiC* C105R, L228F, L377P, A856P, A835V, S762N), some of them in high frequency, including *fbiA* I208V (n = 122)³⁶. Other variants with likely functional impairment of the Fbi proteins comprised one SNP translating into a premature stop codon (*fbiC* G310*) and 12 frameshifts (*fbiA* 2, *fbiB* 1, *fbiC* 9) (S10 Figure; S11 Figure; S12 Figure). In addition, two isolates harboured a 28 amino acid deletion in *fbiA*. One SNP in *fbiA* and 5 SNPs in *fbiC* were found in residues known to be involved in conferring resistance, although different alternate alleles were found compared

to those previously reported (S3 Table). Variants previously associated to susceptibility were identified in *fbia* (Q120R, n = 6; T302M, n = 355), *fbib* (F220L, n = 2; K448R, n = 293), and *fbic* (T273A, n = 626; T681I, n = 9) (see S3 Table). Some of these are phylogenetically related (e.g., *fbia* T302M, *fbic* T681I). Protein structural modelling revealed predicted deleterious novel mutations in *fbia* (6), *fbic* (31), and *fbib* (4), which may have an impact on the function of their proteins, but not necessarily an association with resistance. For *fbid*, 66 variants were found, but all are absent in strains from lineages 5, 6 or 7. No deletions or SNPs leading to stop codons were identified in our analysis (S6 Table; S7 Table), including an absence of the 79_80insC indel, which leads to loss of function of the protein and an increase in DLM and PTM MIC values (S3 Table).

ndh mutations. Three non-synonymous SNPs in *ndh* demonstrated to increase DLM MIC values in *M. smegmatis* (G84V, A175T and M221R)²⁶, were not identified in the corresponding residues of our *Mtb* isolates. Five amino acid changes leading to premature stop codon were identified in the data set, and 20 indels leading to frameshifts and 7 large deletions with potential deleterious effects were found. Only the 304_304 deletion was identified in high frequency, namely in 82 MDR-TB isolates from Australia and Papua New Guinea, collected between 2010 and 2015 (S8 Table).

Discussion

BDQ and DLM are among the last anti-TB drugs approved for the treatment of MDR- and XDR-TB, and have been in use since 2013. Soon after the introduction of BDQ and DLM, resistance to both drugs emerged, and concerns about intrinsic resistance have been raised through the identification of mutations in isolates pre-introduction of both drugs. Similarly, spontaneous resistance-associated variants have been found in BDQ/DLM naïve isolates^{15,16,22,38,39}. Recently, PTM has been introduced in combination therapy with BDQ and LNZ for the treatment of XDR-TB cases. A 6-month regimen of PTM, BDQ, and LNZ for XDR-TB or MDR-intolerant TB has been demonstrated to be 90% effective up to 6 months post-treatment, with no event of acquired resistance to PTM⁴⁰. However, the potential for cross-resistance between DLM and PTM exists.

Our study, consisting of > 33,000 isolates, is the largest study to date, and characterised 1,227 variants in nine drug resistance candidate genes for BDQ and DLM. Most mutations (78%), including frameshifts with likely functional effects, were present in isolates collected prior to roll-out of BDQ and DLM. Our analysis has identified phylogenetically related mutations that are unlikely to be drug resistance associated, including in large clades mostly encompassing sensitive profiles to first- and second-line drugs, as well as several mutations that were not considered strain-specific (e.g., *fbia* G264R, *fbib* L558R or *fbic* E224G). As resistance to BDQ and DLM/PTM is relatively rare, newly associated mutations are likely to be discovered through sequencing of resistant isolates in studies of small samples sizes. A potential pitfall of this approach is the spurious association of lineage-defining mutations to drug resistance in candidate genes. An example of this is the G269S mutation in *kasA*, which was initially suggested to cause isoniazid resistance⁴¹, but in subsequent large studies is associated with T family isolates rather than resistance⁴². To aid researchers in tackling this issue, a list of mutations at high frequency in lineages is provided, and automated detection and annotation of these mutations has now been built into TB-Profler software³¹. One limitation of our analysis is the relatively low number of sequenced isolates from lineages 5 to 7.

We found mutations known to increase BDQ or DLM MICs in isolates predating the introduction of the three drugs as TB treatments. These included 192_193insG, 193_193del (I67fs) and M146T mutations in *mmpR5* and L49P in *ddn*, with all four variants found in > 20 isolates. Although some studies have observed a correlation between the length of BDQ treatment and the acquisition of mutations in *atpE* or *mmpR5*⁴³, the pre-existence of such mutations in BDQ/DLM/PTM naïve isolates has also been described^{8,16,22,38,39,44}. The use of CFZ, which is known to cause cross-resistance through mutations in *mmpR5*⁸, has been proposed as a potential explanation. The M146T mutation in *mmpR5* has been identified in a transmission cluster from Eswatini in 2009, where the use of CFZ by some patients could have selected for this variant³⁵. Similarly, in Portugal the use of CFZ in the treatment of MDR-/XDR-TB patients may have selected for the *mmpR5* frameshift detected¹⁴. In the absence of a previous history of CFZ or BDQ use, the treatment of fungal respiratory infections with azoles (i.e., fluconazole or voriconazole) may explain the presence of *mmpR5* mutations³⁸. The *mmpR5* 192_193 insertion (I67fs) appears in 10 independent clades, with the largest cluster involving lineage 4 Peruvian samples. High pairwise SNP distances within this clade suggest that this mutation became fixed in this strain pre-2013. The suggested epistatic effect of a *mmpL5* deletion identified in these Peruvian strains¹⁷ could counteract the potential associated resistance due to I67fs, although there is currently no supporting phenotypic DST data accounting for the 2 mutations (*mmpR5* 192_193ins-*mmpL5* 605_605del). The I67fs frameshift has also been reported in South Africa¹⁶. A high density of indels were identified along the DNA binding domain of *mmpR5*, which could increase the production of the MmpS5-MmpL5 efflux pump. Fourteen frameshifts were found in the *mmpR5* DNA binding domain, including 2 within the known 192–198 bp hotspot³³.

For the cross-resistance of DLM and PTM, although both pro-drugs are nitroimidazole derivatives that share the activation pathway, the binding of DLM to Ddn might differ from PTM²⁴. However, alteration of specific residues in *ddn*, such as L49P, found in 21 isolates in this study, seemed to confer cross-resistance to both drugs²⁴. Nevertheless, as the introduction of PTM in TB treatment regimens is very recent, its use does not provide an explanation for the acquisition of DLM resistance mutations in pre-2014 isolates, but there is evidence of pre-exposure resistance and naturally occurring polymorphisms⁴⁴.

Frameshifts and nonsense non-synonymous mutations are more likely to have a higher functional impact. We have identified several SNPs causing premature stop codons that have already been associated with increments in MIC (*mmpR5* W42*, *ddn* W88* and *fbia* K250*), as well as others unreported, including one present in eleven lineage 4.5 isolates collected between 2013 and 2015 (*ddn* W20*). Considering the drug susceptibility

profile of these isolates and the high SNP distance within the cluster, it seems unlikely that *ddn* W20* emerged from the use of DLM. The *ddn* locus harboured 16 frameshifts mostly in single isolates, likely associated with loss of function. Ddn may have an essential role in recovery from hypoxia, and mutations that keep its native activity would be favoured over those leading to a loss of function²⁴.

Protein stability predictions can help to elucidate whether the function of these genes might be altered by non-synonymous SNPs. By using the SUSPECT-BDQ prediction tool, we identified 9 mutations in *atpE* predicted to confer resistance. Among these mutations, E44D was present in a clade of Beijing strains with collection years ranging from 2016 to 2019. However, the sensitive profile of the samples and the monophyletic distribution of the substitution, mean that the acquisition of E44D is unlikely to be a consequence of drug selective pressure, although it could be a naturally occurring polymorphism potentially leading to intrinsic elevated MICs to BDQ for this clade. Moreover, all isolates with the E44D variant also had a SNP in *mmpR5* (D5G) which was predicted not to alter protein stability. Using conservative SNAP2, Provean and mCSM software tools and available crystal structures, we found 51 SNPs with predicted alteration of protein function due to their associated amino acid changes. However, further advanced protein modelling analysis or DST data is required to establish evidence of association with BDQ or DLM/PTM resistance. Similarly, a significant number of SNPs in *mmpR5*, *ddn*, *fgd1*, *fbiA* and *fbiC* were found in residues where amino acid changes leading to increments in MICs have been detected. However, the alternate amino acids identified in this analysis were different. Since differing amino acid changes lead to different values of MIC^{24,33}, further investigation is necessary to establish their drug resistance links.

Co-occurrence of mutations in the same gene by isolate was rare. This finding matches previous studies that observed combinations of mutations in *atpE* and *mmpR5* for isolates selected in vitro, whilst clinical isolates tend to harbour unique mutations³⁸. For DLM candidate genes, the combination of variants in *fbiC* and *ddn* or *fbiC* and *fgd1* were the most common, potentially due to the greater diversity of these genes, especially *fbiC* and *fgd1*. Since, only one mutation per sample across the nine genes considered was the most prevalent scenario, any additive effects of mutations to reach BDQ and DLM/PTM resistance maybe unlikely. Nevertheless, one limitation of the study is the higher number of samples with a pre-2014 collection date, and therefore the lack of isolates that may have undergone selective pressure under BDQ or DLM/PTM drug regimens. Some of the variants linked to phenotypic drug susceptibility are considered to confer low-level resistance (0.25–0.75 mg/L) or decreases in susceptibility that reach the MIC breakpoint value established by EUCAST (i.e., some frameshifts in *mmpR5*)^{15,33} for MIC determination using the agar proportion method on Middlebrook 7H10/7H11 medium. Noteworthy, evaluation of MIC values by other studies have shown discrepancies between the methods used^{33,43,45}. Even assuming that a significant number of these known variants elevate the MICs, some values remain within susceptible ranges, their clinical importance is yet unknown, and they could lead to suboptimal treatment regimens⁴³. Moreover, a higher risk of relapse was observed in patients with isolates holding increased MICs but below standard resistance breakpoints for rifampicin and isoniazid⁴⁶. Finally, for *mmpR5*, we observed an elevated risk of mutations among MDR- and XDR-TB isolates, which together with the high proportion of pre-2014 strains, could pose a significant complication for the treatment of BDQ naïve infections.

In summary, we have shown that there are highly frequent resistance-associated variants pre-dating the introduction of BDQ, DLM and PTM, suggesting an intrinsic resistance of these strains, which could constitute a problem for the treatment of MDR-/XDR-TB patients. The use of CFZ and other azoles before the introduction of BDQ could explain the presence of mutations in *mmpR5* in MDR-/XDR-TB isolates. However, the treatment history of some patients is unavailable, including missing sampling dates, making the phylogenetic-based inference of the ages of mutations inaccurate, and the evolutionary pressure by which these mutations have been selected is unclear. Moreover, several frameshifts and nonsense mutations with likely resistance effects have been identified. Since one limitation of the study was the lack of drug susceptibility test data, further investigation is necessary to establish the association between these candidate variants and the phenotypic resistance profiles; ultimately, to elucidate the causative mechanisms of resistance for these new drugs and to achieve better treatment outcomes.

Methods

Candidate genes for BDQ, DLM and PTM drug resistance were selected based on a review of the literature. Only those genes with experimental evidence of developing mutations under drug exposure either in vitro, in vivo or in *M. tuberculosis* clinical isolates were considered. Specifically, we included 3 genes for BDQ (the target *atpE* and off-targets *mmpR5* (Rv0678) and *pepQ*), and 6 genes for DLM/PTM (*ddn*, *fgd1*, *fbiA*, *fbiB*, *fbiC* and *fbiD*) for genetic analysis. Loss of function mutations in the *ndh* gene were considered, as well as in *mmpL5*-*mmpS5* for epistatic effects with *mmpR5*. Phenotypic drug resistance to CFZ and BDQ was assessed for Portuguese clinical isolates by broth microdilution in Middlebrook 7H9 medium supplemented with oleic acid, albumin, dextrose, catalase (OADC) as per the guidelines of the European Committee on Antimicrobial Susceptibility Testing (EUCAST)⁴⁷. BDQ and CFZ concentrations tested ranged between 4 and 0.016 µg/mL. These Portuguese clinical isolates were retrospectively selected from the Faculty of Pharmacy of the University of Lisbon TB strain bank by screening for isolates with available whole genome sequencing (WGS) data and bearing the *mmpR5* I67fs mutation. Only one isolate met these criteria and four additional *mmpR5* wild-type isolates were included for comparative purposes, including one isolate from the same phylogenetic clade as the mutant isolate (L4.3.4.2/SIT20/LAM1/Lisboa3; SNP distance of 34)³⁴. *M. tuberculosis* H37Rv ATCC 27,294 was included as a susceptible reference strain for quality control purpose. Work involving the manipulation of viable *M. tuberculosis* strains and cultures was performed under strict Biosafety Level 3 containment facilities and processed using methods in accordance with the relevant WHO guidelines and institutional regulations.

Publicly available Illumina WGS data for 33,675 *Mtb* isolates spanning 114 countries and all seven main lineages were analysed (see³⁰ for raw data accession numbers). Only WGS data with a minimum average coverage

of 30, >90% of reads mapping to H37Rv and >90% of the genome covered were included. Metadata including collection date and geographical region were incorporated where available. The bioinformatics pipeline for processing raw sequence data is described previously³⁰. In brief, raw sequences were aligned with bwa-mem (v0.7.17) software to the H37Rv reference sequence (Genbank accession: NC_000962.3). SNPs and small indels with an allele frequency >0.95 were identified using GATK HaplotypeCaller (v4.1.4.1). Bcftools csq was used to call amino acid changes. This software handles multiple mutations in the same codon better than alternatives, and in the case of *mmpR5*, some codon numbers differ slightly to previously used nomenclature, and we highlight these (e.g., 193_193del being the same as previously reported I67fs). Large deletions were detected using Delly (v0.8.3, -T DEL) software, and confirmed manually using the IGV (v2.4.9) visualisation tool. TB-Profler (v3.0) software was used to predict lineage and drug resistance to first and second line drugs^{31,48,49}. All high-quality variants identified in the nine candidate genes were extracted. Phylogenetic trees were constructed using concatenated SNP alignments using IQ-Tree (v1.6.12, -m GTR + G + ASC) and visualised together with annotations in iTOL (v5) software. The number of independent acquisitions of variants was calculated by phylogenetic reconstruction followed by ancestral state reconstruction implemented in IQ-Tree (v1.6.12) software.

The R (v3.4.3) statistical package was used to generate the maps. It was also used to perform all statistical analysis, including the fitting of logistic regression models to assess the association of the presence of mutations in candidate genes with the sample collection period, drug resistance status and lineage, where odds ratios and P-values were estimated. The functional effect of SNPs was assessed using SNAP2 and Provean score calculators, and where crystal structures of the *Mtb* proteins were available (PDB: 4NB5, 3R5P, 3B4Y, 4XOM, 6BWG) the mCSM stability predictor was used. For *atpE* SNPs, SUSPECT-BDQ³² was used. The protein structures were visualised and annotated using UCSC chimera (<https://www.cgl.ucsf.edu/chimera/>).

Data availability

Raw sequencing data is available from the ENA short read archive (see³⁰ for a list of accession numbers).

Received: 17 June 2021; Accepted: 16 September 2021

Published online: 30 September 2021

References

1. WHO. *Global Tuberculosis Report 2020*. (2020).
2. World Health Organization. *Meeting report of the WHO expert consultation on the definition of extensively drug-resistant tuberculosis*. (2020).
3. Zumla, A. I. *et al.* New antituberculosis drugs, regimens, and adjunct therapies: Needs, advances, and future prospects. *Lancet Infect. Dis.* **14**, 327–340 (2014).
4. Andries, K. *et al.* A diarylquinoline drug active on the ATP synthase of *Mycobacterium tuberculosis*. *Science* (80-) **307**, 223–227 (2005).
5. Matsumoto, M. *et al.* OPC-67683, a nitro-dihydro-imidazooxazole derivative with promising action against tuberculosis in vitro and in mice. *PLoS Med.* **3**, 2131–2144 (2006).
6. WHO. *Global Tuberculosis Report 2019*. (2019).
7. Choi, K. P., Kendrick, N. & Daniels, L. Demonstration that fbiC is required by *Mycobacterium bovis* BCG for coenzyme F420 and FO biosynthesis. *J. Bacteriol.* **184**, 2420–2428 (2002).
8. Andries, K. *et al.* Acquired resistance of *Mycobacterium tuberculosis* to bedaquiline. *PLoS ONE* **9**, e102135 (2014).
9. Bloemberg, G. V., Gagneux, S. & Böttger, E. C. Acquired resistance to bedaquiline and delamanid in therapy for tuberculosis: To the editor. *N. Engl. J. Med.* **373**, 1986–1988 (2015).
10. Hoffmann, H. *et al.* Delamanid and bedaquiline resistance in *Mycobacterium tuberculosis* ancestral Beijing genotype causing extensively drug-resistant tuberculosis in a tibetan refugee. *Am. J. Respir. Crit. Care Med.* **193**, 337–340 (2016).
11. Hartkoorn, R. C., Uplekar, S. & Cole, S. T. Cross-resistance between clofazimine and bedaquiline through upregulation of *mmp15* in *Mycobacterium tuberculosis*. *Antimicrob. Agents Chemother.* **58**, 2979–2981 (2014).
12. da Silva, P. E. A. & Palomino, J. C. Molecular basis and mechanisms of drug resistance in *Mycobacterium tuberculosis*: Classical and new drugs. *J. Antimicrob. Chemother.* **66**, 1417–1430 (2011).
13. Coll, F. *et al.* Genome-wide analysis of multi- and extensively drug-resistant *Mycobacterium tuberculosis*. *Nat. Genet.* **50**, 307–316 (2018).
14. Perdigão, J. *et al.* Unraveling *Mycobacterium tuberculosis* genomic diversity and evolution in Lisbon, Portugal, a highly drug resistant setting. *BMC Genomics* **15**, 991 (2014).
15. Villellas, C. *et al.* Unexpected high prevalence of resistance-associated Rv0678 variants in MDR-TB patients without documented prior use of clofazimine or bedaquiline. *J. Antimicrob. Chemother.* **72**, dkw502 (2016).
16. Nimmo, C. *et al.* Population-level emergence of bedaquiline and clofazimine resistance-associated variants among patients with drug-resistant tuberculosis in southern Africa: A phenotypic and phylogenetic analysis. *Lancet Microbe* **1**, e165–e174 (2020).
17. Merker, M. *et al.* Phylogenetically informative mutations in genes implicated in antibiotic resistance in *Mycobacterium tuberculosis* complex. *Genome Med.* **12**, 1–8 (2020).
18. Vargas, R. *et al.* The role of epistasis in amikacin, kanamycin, bedaquiline, and clofazimine resistance in *Mycobacterium tuberculosis* complex. *Antimicrob. Agents Chemother.* <https://doi.org/10.1128/aac.01164-21> (2021).
19. Almeida, D. *et al.* Mutations in *pepQ* confer low-level resistance to bedaquiline and clofazimine in *Mycobacterium tuberculosis*. *Antimicrob. Agents Chemother.* **60**, 4590–4599 (2016).
20. Ismail, N., Omar, S. V., Ismail, N. A. & Peters, R. P. H. In vitro approaches for generation of *Mycobacterium tuberculosis* mutants resistant to bedaquiline, clofazimine or linezolid and identification of associated genetic variants. *J. Microbiol. Methods* **153**, 1–9 (2018).
21. Haver, H. L. *et al.* Mutations in genes for the F420 biosynthetic pathway and a nitroreductase enzyme are the primary resistance determinants in spontaneous in vitro-selected PA-824-resistant mutants of *Mycobacterium tuberculosis*. *Antimicrob. Agents Chemother.* **59**, 5316–5323 (2015).
22. Fujiwara, M., Kawasaki, M., Hariguchi, N., Liu, Y. & Matsumoto, M. Mechanisms of resistance to delamanid, a drug for *Mycobacterium tuberculosis*. *Tuberculosis* **108**, 186–194 (2018).
23. Rifat, D. *et al.* Mutations in *fbiD* (Rv2983) as a novel determinant of resistance to pretomanid and delamanid in *Mycobacterium tuberculosis*. *Antimicrob. Agents Chemother.* **65**, e01948–e020 (2020).

24. Lee, B. M. *et al.* Predicting nitroimidazole antibiotic resistance mutations in *Mycobacterium tuberculosis* with protein engineering. *PLoS Pathog.* **16**, 1–27 (2020).
25. Vilchèze, C. *et al.* Altered NADH/NAD⁺ ratio mediates coresistance to isoniazid and ethionamide in mycobacteria. *Antimicrob. Agents Chemother.* **49**, 708–720 (2005).
26. Hayashi, M. *et al.* Adduct formation of delamanid with NAD in mycobacteria. *Antimicrob. Agents Chemother.* **64**, e01755–e1819 (2020).
27. Ramirez, L. M. N., Vargas, K. Q. & Diaz, G. Whole genome sequencing for the analysis of drug resistant strains of *Mycobacterium tuberculosis*: A systematic review for bedaquiline and delamanid. *Antibiotics* **9**, 133 (2020).
28. World Health Organization, (WHO). Technical report on critical concentrations for TB drug susceptibility testing of medicines used in the treatment of drug-resistant TB. *Who* 1–106 (2018).
29. EUCAST. Breakpoint tables for interpretation of MICs and zone diameters. Version 11.0, 2021. <http://www.eucast.org/>.
30. Napier, G. *et al.* Robust barcoding and identification of *Mycobacterium tuberculosis* lineages for epidemiological and clinical studies. *Genome Med.* **12**, 1–10 (2020).
31. 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).
32. Karmakar, M. *et al.* Empirical way to identify novel Bedaquiline resistance mutations in AtpE. *PLoS ONE* **14**, e0217169 (2019).
33. Kadura, S. *et al.* Systematic review of mutations associated with resistance to the new and repurposed *Mycobacterium tuberculosis* drugs bedaquiline, clofazimine, linezolid, delamanid and pretomanid. *J. Antimicrob. Chemother.* **75**, 2031–2043 (2020).
34. Perdigão, J. *et al.* Using genomics to understand the origin and dispersion of multidrug and extensively drug resistant tuberculosis in Portugal. *Sci. Rep.* **10**, 1–17 (2020).
35. Beckert, P. *et al.* MDR *M. tuberculosis* outbreak clone in Eswatini missed by Xpert has elevated bedaquiline resistance dated to the pre-treatment era. *Genome Med.* **12**, 1–11 (2020).
36. Battaglia, S. *et al.* Characterization of genomic variants associated with resistance to bedaquiline and delamanid in naive *Mycobacterium tuberculosis* clinical strains. *J. Clin. Microbiol.* **58**, 1–16 (2020).
37. Nguyen, Q. T., Tringo, G., Binda, C., Mattevi, A. & Fraaije, M. W. Discovery and characterization of an F420-dependent glucose-6-phosphate dehydrogenase (Rh-FGD1) from *Rhodococcus jostii* RHA1. *Appl. Microbiol. Biotechnol.* **101**, 2831–2842 (2017).
38. Zimenkov, D. V. *et al.* Examination of bedaquiline- and linezolid-resistant *Mycobacterium tuberculosis* isolates from the Moscow region. *J. Antimicrob. Chemother.* **72**, 1901–1906 (2017).
39. Xu, J. *et al.* Primary clofazimine and bedaquiline resistance among isolates from patients with multidrug-resistant tuberculosis. *Antimicrob. Agents Chemother.* **61**, 1–8 (2017).
40. Conradie, F. *et al.* Treatment of highly drug-resistant pulmonary tuberculosis. *N. Engl. J. Med.* **382**, 893–902 (2020).
41. Mdluli, K. *et al.* Inhibition of a *Mycobacterium tuberculosis* β -Ketoacyl ACP synthase by isoniazid. *Science (80-)* **280**, 1607–1610 (1998).
42. Sun, Y. J., Lee, A. S. G., Wong, S. Y. & Paton, N. I. Analysis of the role of *Mycobacterium tuberculosis* kasA gene mutations in isoniazid resistance. *Clin. Microbiol. Infect.* **13**, 833–835 (2007).
43. Peretokina, I. V. *et al.* Reduced susceptibility and resistance to bedaquiline in clinical *M. tuberculosis* isolates. *J. Infect.* **80**, 527–535 (2020).
44. Reichmuth, M. L. *et al.* Natural polymorphisms in *Mycobacterium tuberculosis* conferring resistance to delamanid in drug-naïve patients. *Antimicrob. Agents Chemother.* **64**, 1–5 (2020).
45. Ruesen, C. *et al.* Linking minimum inhibitory concentrations to whole genome sequence-predicted drug resistance in *Mycobacterium tuberculosis* strains from Romania. *Sci. Rep.* **8**, 1–8 (2018).
46. Colangeli, R. *et al.* Bacterial factors that predict relapse after tuberculosis therapy. *N. Engl. J. Med.* **379**, 823–833 (2018).
47. Schön, T. *et al.* Antimicrobial susceptibility testing of *Mycobacterium tuberculosis* complex isolates—The EUCAST broth micro-dilution reference method for MIC determination. *Clin. Microbiol. Infect.* **26**, 1488–1492 (2020).
48. Coll, F. *et al.* A robust SNP barcode for typing *Mycobacterium tuberculosis* complex strains. *Nat. Commun.* **5**, 1–5 (2014).
49. Coll, F. *et al.* SpolPred: Rapid and accurate prediction of *Mycobacterium tuberculosis* spoligotypes from short genomic sequences. *Bioinformatics* **28**, 2991–2993 (2012).

Acknowledgements

PJG-G is funded by an MRC-LID 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) and BBSRC (Grant no. BB/R013063/1). SC is funded by Medical Research Council UK grants (ref. MR/M01360X/1, MR/R025576/1, and MR/R020973/1). JP is supported by the Portuguese FCT (ref. CEECIND/00394/2017). PG is the recipient of a PhD studentship from the Portuguese FCT (ref. 2020.05942.BD). The authors declare no conflicts of interest.

Author contributions

J.E.P. and T.G.C. conceived and directed the project. J.P., P.G., Z.P.G., D.S.L., G.N., M.V., I.P. and S.C. contributed data. P.J.G.-G. performed bioinformatic and statistical analyses under the supervision of M.L.H., S.C., J.E.P. and T.G.C. P.J.G.-G., S.C., J.E.P. and T.G.C. interpreted results. P.J.G.-G. wrote the first draft of the manuscript with inputs from J.P., J.E.P. and T.G.C. All authors commented and edited on various versions of the draft manuscript and approved the final manuscript. P.J.G.-G., J.P., 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/s41598-021-98862-4>.

Correspondence and requests for materials should be addressed to T.G.C.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021

Genetic diversity of candidate loci linked to *Mycobacterium tuberculosis* resistance to bedaquiline, delamanid and pretomanid

Paula J. Gómez-González¹, Joao Perdigao², Pedro Gomes², Zully M. Puyen³, David Santos-Lazaro³, Gary Napier¹, Martin L. Hibberd¹, Miguel Viveiros⁴, Isabel Portugal², Susana Campino¹, Jody E. Phelan¹, Taane G. Clark^{1,5}

¹ Faculty of Infectious and Tropical Diseases, London School of Hygiene & Tropical Medicine, WC1E 7HT London, UK; ² Faculdade de Farmácia, Universidade de Lisboa, Portugal; ³ Instituto Nacional de Salud, Lima, Peru; ⁴ Global Health and Tropical Medicine, GHTM, Instituto de Higiene e Medicina Tropical, IHMT, Universidade Nova de Lisboa, Lisbon, Portugal; ⁵ Faculty of Epidemiology and Population Health, London School of Hygiene & Tropical Medicine, WC1E 7HT London, UK

* Correspondence

taane.clark@lshtm.ac.uk, Department of Infection Biology,
Faculty of Infectious and Tropical Diseases
London School of Hygiene & Tropical Medicine, Keppel Street, London, UK

Scientific Reports

S1 Table. Geographical region breakdown summary of isolates analysed.

Region	# countries	# samples	Susc. # (%)	MDR # (%)	XDR # (%)	DR # (%)	Lineages	# pre-2014*
South Asia	6	941	327(34.8)	456(48.5)	23(2.4)	135(14.4)	1-4	305
Europe & Central Asia	36	11323	7414(65.5)	2240(19.8)	427(3.8)	1242(11.0)	1-6	3202
Middle East & N. Africa	9	239	108(45.2)	83(34.7)	23(9.6)	25(10.5)	1-4, 6-7	149
Sub-Saharan Africa	34	8118	6011(74.1)	1175(14.5)	259(3.2)	673(8.3)	1-4, 6-7	5784
Latin America*	13	1463	209(14.3)	923(63.1)	78(5.3)	253(17.3)	1-4	800
East Asia & Pacific	14	6068	3214(53.0)	1371(22.6)	130(2.1)	1353(22.3)	1-4, 7	3874
North America	2	1962	1730(88.2)	27(1.4)	0(0)	205(10.5)	1-5	1658
Unknown	-	3561	2762(77.6)	228(6.4)	23(0.7)	548(15.4)	1-6	-
Overall	113	33675	21775(64.7)	6503(19.3)	963(2.9)	4434(13.2)	1-7	15772

* and Caribbean; # = number, Susc. = Susceptible; MDR = multidrug resistant; XDR = extensively drug resistant; DR = Other resistance; ** Number of isolates with date of collection data before 2014.

S2 Table. Analysis of the odds of gene mutations.

Gene	Variable	Odds ratio*	95% Lower confidence limit	95% Upper confidence limit	P-value
<i>mmpr5</i>	Sensitive	1.000			
	Other DR**	2.040	1.367	3.044	<0.0001
	MDR	3.781	2.765	5.171	<0.0001
	XDR	9.937	6.626	14.904	<0.0001
<i>ddn</i>	Sensitive	1.000			
	Other DR**	1.019	0.684	1.517	0.926
	MDR	1.559	1.104	2.202	0.012
	XDR	2.268	1.150	4.474	0.018

* adjusted for lineage and year of collection; ** non-MDR; MDR multi-drug resistant; XDR extensively drug resistant

S3 Table. Mutations reported in the literature

Drug	Phenotype	Gene	Mutation*	PMID	Author
BDQ	Resistant	<i>atpE</i>	G25S	PMID:31981638	Peretokina et al., 2020
BDQ	Resistant	<i>atpE</i>	D28A	PMID:30165087	Ismail et al., 2018
BDQ	Resistant	<i>atpE</i>	D28G	PMID:30165087	Ismail et al., 2018
BDQ	Resistant	<i>atpE</i>	D28P	PMID:20038615	Huitric et al., 2010
BDQ	Resistant	<i>atpE</i>	D28V	PMID:30165087	Ismail et al., 2018
BDQ	Resistant	<i>atpE</i>	D28N	PMID:31981638	Peretokina et al., 2020
BDQ	Resistant	<i>atpE</i>	E32V	PMID:20038615	Huitric et al., 2010
BDQ	Resistant	<i>atpE</i>	L59V	PMID:22354303	Segala et al., 2012
BDQ	Resistant	<i>atpE</i>	E61D	PMID:30165087	Ismail et al., 2018
BDQ	Resistant	<i>atpE</i>	A63V	PMID:31981638	Peretokina et al., 2020
BDQ	Resistant	<i>atpE</i>	A63P	PMID:15591164	Andries et al., 2005
BDQ	Resistant	<i>atpE</i>	I66M	PMID:17496888	Koul et al., 2007
BDQ	Resistant	<i>atpE</i>	I66V	PMID:30029911	Martinez et al., 2018
BDQ	Resistant	<i>mmpR5</i>	-13insIS6110	PMID:28031270	Villellas et al., 2017
BDQ	Resistant	<i>mmpR5</i>	V1A	PMID:26559594	Bloemberg et al., 2015
BDQ	Resistant	<i>mmpR5</i>	S2I	PMID:28320727	Xu et al., 2017
BDQ	Resistant	<i>mmpR5</i>	16_16del	PMID:31981638	Peretokina et al., 2020
BDQ	Resistant	<i>mmpR5</i>	16_17del	PMID:28182568	Veziris et al., 2017
BDQ	Resistant	<i>mmpR5</i>	27_28insC	PMID:32907992	Battaglia et al., 2020
BDQ	Resistant	<i>mmpR5</i>	30_30del	PMID:31981638	Peretokina et al., 2020
BDQ	Resistant	<i>mmpR5</i>	37_38insA	PMID:25010492	Andries et al., 2014
BDQ	Resistant	<i>mmpR5</i>	43_44insA	PMID:31981638	Peretokina et al., 2020
BDQ	Resistant	<i>mmpR5</i>	V20G	PMID:31262765	Ghodousi et al., 2019
BDQ	Resistant	<i>mmpR5</i>	E21D	PMID:30165087	Ismail et al., 2018
BDQ	Resistant	<i>mmpR5</i>	E21Q	PMID:31981638	Peretokina et al., 2020
BDQ	Resistant	<i>mmpR5</i>	Q22L	PMID:30642938	Ismail et al., 2019
BDQ	Resistant	<i>mmpR5</i>	65_66insIS6110	PMID:30833432	Xu et al., 2019
BDQ	Resistant	<i>mmpR5</i>	M23L	PMID:31981638	Peretokina et al., 2020
BDQ	Resistant	<i>mmpR5</i>	71_72insGC	PMID:30833432	Xu et al., 2019
BDQ	Resistant	<i>mmpR5</i>	G25D	PMID:30833432	Xu et al., 2019
BDQ	Resistant	<i>mmpR5</i>	G25C	PMID:30833432	Xu et al., 2019
BDQ	Resistant	<i>mmpR5</i>	Y26*	PMID:28387862	Zimenkov et al., 2017
BDQ	Resistant	<i>mmpR5</i>	94insIS6110	PMID:30833432	Xu et al., 2019
BDQ	Resistant	<i>mmpR5</i>	T33A	PMID:30642938	Ismail et al., 2019
BDQ	Resistant	<i>mmpR5</i>	A36T	PMID:31138569	Ismail et al., 2019
BDQ	Resistant	<i>mmpR5</i>	A36V	PMID:31981638	Peretokina et al., 2020
BDQ	Resistant	<i>mmpR5</i>	L40S	PMID:31981638	Peretokina et al., 2020
BDQ	Resistant	<i>mmpR5</i>	Ismail et al., 2019	PMID:31138569	Ismail et al., 2019
BDQ	Resistant	<i>mmpR5</i>	L43P	PMID:30833432	Xu et al., 2019
BDQ	Resistant	<i>mmpR5</i>	133_134delGT	PMID:30833432	Xu et al., 2019

BDQ	Resistant	<i>mmpR5</i>	133_134insTG	PMID:31981638	Peretokina et al., 2020
BDQ	Resistant	<i>mmpR5</i>	C46R	PMID:31138569	Ismail et al., 2019
BDQ	Resistant	<i>mmpR5</i>	136_137insG	PMID:29337135	Ismail et al., 2018
BDQ	Resistant	<i>mmpR5</i>	138_139insG (D47fs)	PMID:29337135	Ismail et al., 2018
BDQ	Resistant	<i>mmpR5</i>	138_139insGA (D47fs)	PMID:31141643	de Vos et al., 2019
BDQ	Resistant	<i>mmpR5</i>	139_140insTG	PMID:28387862	Zimenkov et al., 2017
BDQ	Resistant	<i>mmpR5</i>	140_141insG	PMID:28182568	Veziris et al., 2017
BDQ	Resistant	<i>mmpR5</i>	141_142insC	PMID:29337135	Ismail et al., 2018
BDQ	Resistant	<i>mmpR5</i>	E49*	PMID:28387862	Zimenkov et al., 2017
BDQ	Resistant	<i>mmpR5</i>	R50W	PMID:28031270	Villellas et al., 2017
BDQ	Resistant	<i>mmpR5</i>	S52F	PMID:28031270	Villellas et al., 2017
BDQ	Resistant	<i>mmpR5</i>	S53L	PMID:28320727	Xu et al., 2017
BDQ	Resistant	<i>mmpR5</i>	S53P	PMID:28320727	Xu et al., 2017
BDQ	Resistant	<i>mmpR5</i>	168_168del	PMID:30833432	Xu et al., 2019
BDQ	Resistant	<i>mmpR5</i>	T58P	PMID:31981638	Peretokina et al., 2020
BDQ	Resistant	<i>mmpR5</i>	A59V	PMID:31981638	Peretokina et al., 2020
BDQ	Resistant	<i>mmpR5</i>	175_176insCG	PMID:31981638	Peretokina et al., 2020
BDQ	Resistant	<i>mmpR5</i>	184_185insC	PMID:31981638	Peretokina et al., 2020
BDQ	Resistant	<i>mmpR5</i>	185_186insCAG	PMID:30933266	Polsfuss et al., 2019
BDQ	Resistant	<i>mmpR5</i>	A62V	PMID:28031270	Villellas et al., 2017
BDQ	Resistant	<i>mmpR5</i>	S63R	PMID:24590481	Hartkoorn et al., 2014
BDQ	Resistant	<i>mmpR5</i>	S63G	PMID:30642938	Ismail et al., 2019
BDQ	Resistant	<i>mmpR5</i>	192_193insG (I67fs)	PMID:25010492	Andries et al., 2014
BDQ	Resistant	<i>mmpR5</i>	193_193del (I67fs)	PMID:30248414	Chawla et al., 2018
BDQ	Resistant	<i>mmpR5</i>	G65R	PMID:31981638	Peretokina et al., 2020
BDQ	Resistant	<i>mmpR5</i>	G66E	PMID:30833432	Xu et al., 2019
BDQ	Resistant	<i>mmpR5</i>	G66W	PMID:31981638	Peretokina et al., 2020
BDQ	Resistant	<i>mmpR5</i>	I67S	PMID:29337135	Ismail et al., 2018
BDQ	Resistant	<i>mmpR5</i>	S68G	PMID:25010492	Andries et al., 2014
BDQ	Resistant	<i>mmpR5</i>	201_206del	PMID:29337135	Ismail et al., 2018
BDQ	Resistant	<i>mmpR5</i>	212_212del	PMID:28387862	Zimenkov et al., 2017
BDQ	Resistant	<i>mmpR5</i>	R72W	PMID:31138569	Ismail et al., 2019
BDQ	Resistant	<i>mmpR5</i>	R72Q	PMID:29038265	Xu et al., 2017
BDQ	Resistant	<i>mmpR5</i>	L74P	PMID:31138569	Ismail et al., 2019
BDQ	Resistant	<i>mmpR5</i>	L74V	PMID:31981638	Peretokina et al., 2020
BDQ	Resistant	<i>mmpR5</i>	224_225insA	PMID:28031270	Villellas et al., 2017
BDQ	Resistant	<i>mmpR5</i>	G78A	PMID:31981638	Peretokina et al., 2020
BDQ	Resistant	<i>mmpR5</i>	F79S	PMID:31981638	Peretokina et al., 2020
BDQ	Resistant	<i>mmpR5</i>	I80M	PMID:31981638	Peretokina et al., 2020
BDQ	Resistant	<i>mmpR5</i>	L83P	PMID:31138569	Ismail et al., 2019
BDQ	Resistant	<i>mmpR5</i>	L83V	PMID:31981638	Peretokina et al., 2020
BDQ	Resistant	<i>mmpR5</i>	V85A	PMID:31981638	Peretokina et al., 2020
BDQ	Resistant	<i>mmpR5</i>	258_259insG	PMID:30833432	Xu et al., 2019

BDQ	Resistant	<i>mmpR5</i>	262_263insA	PMID:31981638	Peretokina et al., 2020
BDQ**	Resistant	<i>mmpR5</i>	R90C	PMID:30833432	Xu et al., 2019
BDQ	Resistant	<i>mmpR5</i>	272insIS6110	PMID:25010492	Andries et al., 2014
BDQ	Resistant	<i>mmpR5</i>	274_278del	PMID:31981638	Peretokina et al., 2020
BDQ	Resistant	<i>mmpR5</i>	Y92*	PMID:29941636	Rancoita et al., 2018
BDQ	Resistant	<i>mmpR5</i>	274_275insA	PMID:32907992	Battaglia et al., 2020
BDQ	Resistant	<i>mmpR5</i>	274_283del	PMID:28387862	Zimenkov et al., 2017
BDQ	Resistant	<i>mmpR5</i>	R94Q	PMID:25010492	Andries et al., 2014
BDQ	Resistant	<i>mmpR5</i>	R96Q	PMID:30029911	Martinez et al., 2018
BDQ	Resistant	<i>mmpR5</i>	R96W	PMID:30833432	Xu et al., 2019
BDQ	Resistant	<i>mmpR5</i>	289_289del	PMID:28387862	Zimenkov et al., 2017
BDQ	Resistant	<i>mmpR5</i>	291_292insA (N98fs)	PMID:31981638	Peretokina et al., 2020
BDQ	Resistant	<i>mmpR5</i>	N98D	PMID:33239092	Beckert et al., 2020
BDQ	Resistant	<i>mmpR5</i>	A98V	PMID:28387862	Zimenkov et al., 2017
BDQ	Resistant	<i>mmpR5</i>	A99V	PMID:31981638	Peretokina et al., 2020
BDQ	Resistant	<i>mmpR5</i>	A102P	PMID:31138569	Ismail et al., 2019
BDQ	Resistant	<i>mmpR5</i>	R105C	PMID:30833432	Xu et al., 2019
BDQ	Resistant	<i>mmpR5</i>	314_315delGT	PMID:30833432	Xu et al., 2019
BDQ	Resistant	<i>mmpR5</i>	318_319insCG	PMID:31981638	Peretokina et al., 2020
BDQ	Resistant	<i>mmpR5</i>	R107C	PMID:30833432	Xu et al., 2019
BDQ	Resistant	<i>mmpR5</i>	A112S	PMID:28031270	Villellas et al., 2017
BDQ	Resistant	<i>mmpR5</i>	334_335insIS6110	PMID:30833432	Xu et al., 2019
BDQ	Resistant	<i>mmpR5</i>	335_335del	PMID:28387862	Zimenkov et al., 2017
BDQ	Resistant	<i>mmpR5</i>	E113K	PMID:31981638	Peretokina et al., 2020
BDQ	Resistant	<i>mmpR5</i>	L114P	PMID:31981638	Peretokina et al., 2020
BDQ	Resistant	<i>mmpR5</i>	345_345del	PMID:29941636	Rancoita et al., 2018
BDQ**	Resistant	<i>mmpR5</i>	L117R	PMID:29038265	Xu et al., 2017
BDQ	Resistant	<i>mmpR5</i>	349insIS6110	PMID:30833432	Xu et al., 2019
BDQ	Resistant	<i>mmpR5</i>	359_360insG	PMID:31981638	Peretokina et al., 2020
BDQ	Resistant	<i>mmpR5</i>	G120E	PMID:30833432	Xu et al., 2019
BDQ	Resistant	<i>mmpR5</i>	G121E	PMID:28387862	Zimenkov et al., 2017
BDQ	Resistant	<i>mmpR5</i>	G121V	PMID:30833432	Xu et al., 2019
BDQ	Resistant	<i>mmpR5</i>	G121R	PMID:33239092	Beckert et al., 2020
BDQ	Resistant	<i>mmpR5</i>	L122P	PMID:28387862	Zimenkov et al., 2017
BDQ	Resistant	<i>mmpR5</i>	382_383insC	PMID:28387862	Zimenkov et al., 2017
BDQ	Resistant	<i>mmpR5</i>	R134*	PMID:24590481	Hartkoorn et al., 2014
BDQ	Resistant	<i>mmpR5</i>	R135G	PMID:30165087	Ismail et al., 2018
BDQ	Resistant	<i>mmpR5</i>	R135W	PMID:31981638	Peretokina et al., 2020
BDQ	Resistant	<i>mmpR5</i>	L136P	PMID:31138569	Ismail et al., 2019
BDQ	Resistant	<i>mmpR5</i>	E138G	PMID:25010492	Andries et al., 2014
BDQ	Resistant	<i>mmpR5</i>	E138fs	PMID:31138569	Ismail et al., 2019
BDQ	Resistant	<i>mmpR5</i>	M139I	PMID:31981638	Peretokina et al., 2020
BDQ	Resistant	<i>mmpR5</i>	M139T	PMID:28182568	Veziris et al., 2017

BDQ	Resistant	<i>mmpR5</i>	418_419insG	PMID:28387862	Zimenkov et al., 2017
BDQ	Resistant	<i>mmpR5</i>	L142R	PMID:31981638	Peretokina et al., 2020
BDQ	Resistant	<i>mmpR5</i>	425_425del	PMID:30833432	Xu et al., 2019
BDQ	Resistant	<i>mmpR5</i>	435_435del	PMID:28387862	Zimenkov et al., 2017
BDQ	Resistant	<i>mmpR5</i>	M146T	PMID:29038265	Xu et al., 2017
BDQ	Resistant	<i>mmpR5</i>	435_436insA	PMID:30833432	Xu et al., 2019
BDQ	Resistant	<i>mmpR5</i>	A153P	PMID:30833432	Xu et al., 2019
BDQ	Resistant	<i>mmpR5</i>	L154P	PMID:30165087	Ismail et al., 2018
BDQ	Resistant	<i>mmpR5</i>	465_466insC (R156fs)	PMID:31981638	Peretokina et al., 2020
BDQ	Resistant	<i>mmpR5</i>	Y157D	PMID:28739779	Pang et al., 2017
BDQ	Resistant	<i>mmpR5</i>	466_467insGA	PMID:28387862	Zimenkov et al., 2017
BDQ	Resistant	<i>pepQ</i>	A14fs	PMID:27185800	Almedia et al., 2016
BDQ	Resistant	<i>pepQ</i>	M23T	PMID:30833432	Xu et al., 2019
BDQ	Resistant	<i>pepQ</i>	L44P	PMID:27185800	Almedia et al., 2016
BDQ	Resistant	<i>pepQ</i>	E139K	PMID:30833432	Xu et al., 2019
BDQ	Resistant	<i>pepQ</i>	812_813insG	PMID:30833432	Xu et al., 2019
BDQ	Resistant	<i>pepQ</i>	R271fs	PMID:30833432	Xu et al., 2019
BDQ	Resistant	<i>pepQ</i>	G299V	PMID:30833432	Xu et al., 2019
BDQ	Susceptible	<i>mmpR5</i>	-59T>C	PMID:28031270	Villellas et al., 2017
BDQ	Susceptible	<i>mmpR5</i>	-53C>A	PMID:32907992	Battaglia et al., 2020
BDQ	Susceptible	<i>mmpR5</i>	-47T>C	PMID:28031270	Villellas et al., 2017
BDQ**	Susceptible	<i>mmpR5</i>	-44T>C	PMID:28031270	Villellas et al., 2017
BDQ	Susceptible	<i>mmpR5</i>	-20T>A	PMID:32907992	Battaglia et al., 2020
BDQ	Susceptible	<i>mmpR5</i>	-11C>A	PMID:30029911	Martinez et al., 2018
BDQ	Susceptible	<i>mmpR5</i>	-4A>T	PMID:32907992	Battaglia et al., 2020
BDQ	Susceptible	<i>mmpR5</i>	V3I	PMID:29941636	Rancoita et al., 2018
BDQ	Susceptible	<i>mmpR5</i>	N4T	PMID:32907992	Battaglia et al., 2020
BDQ	Susceptible	<i>mmpR5</i>	D5G	PMID:30029911	Martinez et al., 2018
BDQ**	Susceptible	<i>mmpR5</i>	43_44insA	PMID:28387862	Zimenkov et al., 2017
BDQ	Susceptible	<i>mmpR5</i>	46_47insTCATGGAATTCG	PMID:32907992	Battaglia et al., 2020
BDQ	Susceptible	<i>mmpR5</i>	V20A	PMID:32907992	Battaglia et al., 2020
BDQ**	Susceptible	<i>mmpR5</i>	M23V	PMID:30029911	Martinez et al., 2018
BDQ	Susceptible	<i>mmpR5</i>	G37S	PMID:32907992	Battaglia et al., 2020
BDQ**	Susceptible	<i>mmpR5</i>	L39S	PMID:28031270	Villellas et al., 2017
BDQ**	Susceptible	<i>mmpR5</i>	W42R	PMID:28031270	Villellas et al., 2017
BDQ	Susceptible	<i>mmpR5</i>	D44G	PMID:28031270	Villellas et al., 2017
BDQ	Susceptible	<i>mmpR5</i>	M49L	PMID:28031270	Villellas et al., 2017
BDQ	Susceptible	<i>mmpR5</i>	R50P	PMID:31981638	Peretokina et al., 2020
BDQ**	Susceptible	<i>mmpR5</i>	E55D	PMID:30029911	Martinez et al., 2018
BDQ**	Susceptible	<i>mmpR5</i>	212_212del	PMID:28387862	Zimenkov et al., 2017
BDQ	Susceptible	<i>mmpR5</i>	225_225del	PMID:28031270	Villellas et al., 2017
BDQ	Susceptible	<i>mmpR5</i>	R82Q	PMID:31981638	Peretokina et al., 2020
BDQ	Susceptible	<i>mmpR5</i>	A84V	PMID:32907992	Battaglia et al., 2020

BDQ	Susceptible	<i>mmpR5</i>	V85G	PMID:31981638	Peretokina et al., 2020
BDQ**	Susceptible	<i>mmpR5</i>	A86T	PMID:28031270	Villellas et al., 2017
BDQ**	Susceptible	<i>mmpR5</i>	G87R	PMID:30029911	Martinez et al., 2018
BDQ**	Susceptible	<i>mmpR5</i>	D88G	PMID:28031270	Villellas et al., 2017
BDQ	Susceptible	<i>mmpR5</i>	R90L	PMID:28031270	Villellas et al., 2017
BDQ	Susceptible	<i>mmpR5</i>	Y92D	PMID:31981638	Peretokina et al., 2020
BDQ	Susceptible	<i>mmpR5</i>	A101T	PMID:28031270	Villellas et al., 2017
BDQ	Susceptible	<i>mmpR5</i>	A110V	PMID:33239092	Beckert et al., 2020
BDQ	Susceptible	<i>mmpR5</i>	M111V	PMID:32907992	Battaglia et al., 2020
BDQ	Susceptible	<i>mmpR5</i>	D116N	PMID:28031270	Villellas et al., 2017
BDQ	Susceptible	<i>mmpR5</i>	V120M	PMID:32907992	Battaglia et al., 2020
BDQ	Susceptible	<i>mmpR5</i>	L136V	PMID:32907992	Battaglia et al., 2020
BDQ	Susceptible	<i>mmpR5</i>	D141H	PMID:32907992	Battaglia et al., 2020
BDQ	Susceptible	<i>mmpR5</i>	Y145N	PMID:28031270	Villellas et al., 2017
BDQ	Susceptible	<i>mmpR5</i>	M146R	PMID:32907992	Battaglia et al., 2020
BDQ	Susceptible	<i>mmpR5</i>	457_458insC	PMID:32907992	Battaglia et al., 2020
BDQ	Susceptible	<i>mmpR5</i>	S157E	PMID:28031270	Villellas et al., 2017
BDQ	Susceptible	<i>pepQ</i>	-31C>T	PMID:32907992	Battaglia et al., 2020
BDQ	Susceptible	<i>pepQ</i>	-12G>C	PMID:32907992	Battaglia et al., 2020
BDQ	Susceptible	<i>pepQ</i>	H3Y	PMID:32907992	Battaglia et al., 2020
BDQ	Susceptible	<i>pepQ</i>	R7Q	PMID:32907992	Battaglia et al., 2020
BDQ	Susceptible	<i>pepQ</i>	P69L	PMID:32907992	Battaglia et al., 2020
BDQ	Susceptible	<i>pepQ</i>	A78V	PMID:32907992	Battaglia et al., 2020
BDQ	Susceptible	<i>pepQ</i>	A90V	PMID:32907992	Battaglia et al., 2020
BDQ	Susceptible	<i>pepQ</i>	V92M	PMID:32907992	Battaglia et al., 2020
BDQ	Susceptible	<i>pepQ</i>	D93E	PMID:32907992	Battaglia et al., 2020
BDQ	Susceptible	<i>pepQ</i>	D136E	PMID:32907992	Battaglia et al., 2020
BDQ	Susceptible	<i>pepQ</i>	A152T	PMID:32907992	Battaglia et al., 2020
BDQ	Susceptible	<i>pepQ</i>	R167L	PMID:32907992	Battaglia et al., 2020
BDQ	Susceptible	<i>pepQ</i>	M180V	PMID:32907992	Battaglia et al., 2020
BDQ	Susceptible	<i>pepQ</i>	V214A	PMID:32907992	Battaglia et al., 2020
BDQ	Susceptible	<i>pepQ</i>	V214F	PMID:32907992	Battaglia et al., 2020
BDQ	Susceptible	<i>pepQ</i>	T236A	PMID:32907992	Battaglia et al., 2020
BDQ	Susceptible	<i>pepQ</i>	A305V	PMID:32907992	Battaglia et al., 2020
BDQ	Susceptible	<i>pepQ</i>	G309R	PMID:32907992	Battaglia et al., 2020
BDQ	Susceptible	<i>pepQ</i>	T341A	PMID:32907992	Battaglia et al., 2020
BDQ	Susceptible	<i>pepQ</i>	A370T	PMID:32907992	Battaglia et al., 2020
BDQ	Susceptible	<i>pepQ</i>	L372V	PMID:32907992	Battaglia et al., 2020
DLM	Resistant	<i>ddn</i>	M1T	PMID:32032366	Lee et al., 2019
DLM	Resistant	<i>ddn</i>	2_2del	PMID:32907992	Battaglia et al., 2020
DLM	Resistant	<i>ddn</i>	P2Q	PMID:32907992	Battaglia et al., 2020
PTM	Resistant	<i>ddn</i>	24_24del	PMID:26100695	Haver et al., 2015
PTM	Resistant	<i>ddn</i>	S11*	PMID:26100695	Haver et al., 2015

PTM	Resistant	<i>ddn</i>	38_38del	PMID:21930879	Feuerriegel et al., 2011
PTM	Resistant	<i>ddn</i>	L13fs	PMID:16387854	Manjunatha et al., 2006
DLM	Resistant	<i>ddn</i>	41_41del	PMID:32907992	Battaglia et al., 2020
DLM/PTM	Resistant	<i>ddn</i>	S22L	PMID:32032366	Lee et al., 2019
DLM	Resistant	<i>ddn</i>	68_69insGATTAATACCT	PMID:27076101	Schena et al., 2016
PTM	Resistant	<i>ddn</i>	73_73del	PMID:26100695	Haver et al., 2015
PTM	Resistant	<i>ddn</i>	N25I	PMID:26100695	Haver et al., 2015
DLM	Resistant	<i>ddn</i>	W27*	PMID:32907992	Battaglia et al., 2020
PTM	Resistant	<i>ddn</i>	Y29*	PMID:26100695	Haver et al., 2015
DLM	Resistant	<i>ddn</i>	R30H	PMID:32907992	Battaglia et al., 2020
DLM/PTM	Resistant	<i>ddn</i>	117_117del	PMID:33077652	Rifat et al., 2020
PTM	Resistant	<i>ddn</i>	Q42*	PMID:26100695	Haver et al., 2015
PTM	Resistant	<i>ddn</i>	L48P	PMID:26100695	Haver et al., 2015
DLM/PTM	Resistant	<i>ddn</i>	L49P	PMID:32032366	Lee et al., 2019
DLM	Resistant	<i>ddn</i>	G53D	PMID:30933266	Polzfuss et al., 2019
PTM	Resistant	<i>ddn</i>	163_164insCGC	PMID:26100695	Haver et al., 2015
PTM	Resistant	<i>ddn</i>	163_164ins21bp	PMID:26100695	Haver et al., 2015
DLM	Resistant	<i>ddn</i>	Q58*	PMID:32907992	Battaglia et al., 2020
DLM	Resistant	<i>ddn</i>	180_181insGGTCA	PMID:27076101	Schena et al., 2016
DLM/PTM	Resistant	<i>ddn</i>	L64P	PMID:32032366	Lee et al., 2019
PTM	Resistant	<i>ddn</i>	Y65L	PMID:32032366	Lee et al., 2019
PTM	Resistant	<i>ddn</i>	Y65M	PMID:32032366	Lee et al., 2019
PTM	Resistant	<i>ddn</i>	Y65C	PMID:32032366	Lee et al., 2019
DLM/PTM	Resistant	<i>ddn</i>	Y65S	PMID:32032366	Lee et al., 2019
DLM	Resistant	<i>ddn</i>	215_215del	PMID:29523322	Fujiwara et al., 2018
PTM	Resistant	<i>ddn</i>	A76E	PMID:16387854	Manjunatha et al., 2006
PTM	Resistant	<i>ddn</i>	S78Y	PMID:32032366	Lee et al., 2019
PTM	Resistant	<i>ddn</i>	S78A	PMID:32032366	Lee et al., 2019
PTM	Resistant	<i>ddn</i>	S78C	PMID:32032366	Lee et al., 2019
PTM	Resistant	<i>ddn</i>	S78T	PMID:32032366	Lee et al., 2019
PTM	Resistant	<i>ddn</i>	S78V	PMID:32032366	Lee et al., 2019
PTM	Resistant	<i>ddn</i>	S78P	PMID:26100695	Haver et al., 2015
DLM/PTM	Resistant	<i>ddn</i>	K79Q	PMID:32032366	Lee et al., 2019
PTM	Resistant	<i>ddn</i>	G81D	PMID:26100695	Haver et al., 2015
DLM	Resistant	<i>ddn</i>	252_253del	PMID:29523322	Fujiwara et al., 2018
PTM	Resistant	<i>ddn</i>	P86L	PMID:26100695	Haver et al., 2015
DLM/PTM	Resistant	<i>ddn</i>	W88R	PMID:32032366	Lee et al., 2019
DLM	Resistant	<i>ddn</i>	W88*	PMID:27076101	Schena et al., 2016
PTM	Resistant	<i>ddn</i>	Y89*	PMID:16387854	Manjunatha et al., 2006
DLM	Resistant	<i>ddn</i>	L91P	PMID:29523322	Fujiwara et al., 2018
PTM	Resistant	<i>ddn</i>	289_289del	PMID:26100695	Haver et al., 2015
DLM	Resistant	<i>ddn</i>	307_307del	PMID:27076101	Schena et al., 2016
DL:M	Resistant	<i>ddn</i>	L107P	PMID:27076101	Schena et al., 2016

DLM/PTM	Resistant	<i>ddn</i>	324_325insIS6110	PMID:33077652	Rifat et al., 2020
DLM	Resistant	<i>ddn</i>	328_329insC	PMID:29523322	Fujiwara et al., 2018
DLM/PTM	Resistant	<i>ddn</i>	R112W	PMID:33077652	Rifat et al., 2020
PTM	Resistant	<i>ddn</i>	E121K	PMID:26100695	Haver et al., 2015
PTM	Resistant	<i>ddn</i>	Y133C	PMID:32032366	Lee et al., 2019
PTM	Resistant	<i>ddn</i>	Y133D	PMID:26100695	Haver et al., 2015
PTM	Resistant	<i>ddn</i>	Y133L	PMID:32032366	Lee et al., 2019
PTM	Resistant	<i>ddn</i>	Y133W	PMID:32032366	Lee et al., 2019
PTM	Resistant	<i>ddn</i>	Y133M	PMID:32032366	Lee et al., 2019
PTM	Resistant	<i>ddn</i>	Y136E	PMID:32032366	Lee et al., 2019
DLM	Resistant	<i>ddn</i>	Y136S	PMID:32032366	Lee et al., 2019
PTM	Resistant	<i>ddn</i>	Y136T	PMID:32032366	Lee et al., 2019
PTM	Resistant	<i>ddn</i>	Q137*	PMID:26100695	Haver et al., 2015
DLM	Resistant	<i>ddn</i>	432_432del	PMID:29523322	Fujiwara et al., 2018
PTM	Resistant	<i>ddn</i>	C149Y	PMID:32032366	Lee et al., 2019
DLM/PTM	Resistant	<i>fgd1</i>	K9N	PMID:33077652	Rifat et al., 2020
PTM	Resistant	<i>fgd1</i>	P43R	PMID:26100695	Haver et al., 2015
PTM	Resistant	<i>fgd1</i>	G71D	PMID:26100695	Haver et al., 2015
DLM	Resistant	<i>fgd1</i>	227_228del	PMID:29523322	Fujiwara et al., 2018
PTM	Resistant	<i>fgd1</i>	Q88E	PMID:21930879	Feuerriegel et al., 2011
DLM	Resistant	<i>fgd1</i>	A89P	PMID:29523322	Fujiwara et al., 2018
PTM	Resistant	<i>fgd1</i>	G106V	PMID:26100695	Haver et al., 2015
PTM	Resistant	<i>fgd1</i>	N112K	PMID:26100695	Haver et al., 2015
PTM	Resistant	<i>fgd1</i>	146_151del	PMID:26100695	Haver et al., 2015
PTM	Resistant	<i>fgd1</i>	W143*	PMID:26100695	Haver et al., 2015
PTM	Resistant	<i>fgd1</i>	496_496del	PMID:26100695	Haver et al., 2015
PTM	Resistant	<i>fgd1</i>	G169A	PMID:26100695	Haver et al., 2015
DLM/PTM	Resistant	<i>fgd1</i>	G191D	PMID:33077652	Rifat et al., 2020
DLM	Resistant	<i>fgd1</i>	629_630insG	PMID:29523322	Fujiwara et al., 2018
PTM	Resistant	<i>fgd1</i>	678_678del	PMID:26100695	Haver et al., 2015
PTM	Resistant	<i>fgd1</i>	E230K	PMID:26100695	Haver et al., 2015
DLM	Resistant	<i>fgd1</i>	G314E	PMID:32907992	Battaglia et al., 2020
DLM	Resistant	<i>fbIA</i>	K2E	PMID:32907992	Battaglia et al., 2020
PTM	Resistant	<i>fbIA</i>	Q21P	PMID:26100695	Haver et al., 2015
DLM/PTM	Resistant	<i>fbIA</i>	Q27*	PMID:33077652	Rifat et al., 2020
PTM	Resistant	<i>fbIA</i>	D43Y	PMID:26100695	Haver et al., 2015
DLM/PTM	Resistant	<i>fbIA</i>	141_141del	PMID:33077652	Rifat et al., 2020
DLM	Resistant	<i>fbIA</i>	D49T	PMID:26559594	Bloemberg et al., 2015
DLM	Resistant	<i>fbIA</i>	D49Y	PMID:26829425	Hoffmann et al., 2016
DLM/PTM	Resistant	<i>fbIA</i>	D49G	PMID:33077652	Rifat et al., 2020
PTM	Resistant	<i>fbIA</i>	L56P	PMID:26100695	Haver et al., 2015
PTM	Resistant	<i>fbIA</i>	D63G	PMID:26100695	Haver et al., 2015
PTM	Resistant	<i>fbIA</i>	C65W	PMID:26100695	Haver et al., 2015

PTM	Resistant	<i>fbiA</i>	211_211del	PMID:26100695	Haver et al., 2015
PTM	Resistant	<i>fbiA</i>	222_223del	PMID:26100695	Haver et al., 2015
PTM	Resistant	<i>fbiA</i>	227_228insC	PMID:26100695	Haver et al., 2015
PTM	Resistant	<i>fbiA</i>	W79*	PMID:26100695	Haver et al., 2015
PTM	Resistant	<i>fbiA</i>	242_243insC	PMID:26100695	Haver et al., 2015
PTM	Resistant	<i>fbiA</i>	A88D	PMID:26100695	Haver et al., 2015
DLM	Resistant	<i>fbiA</i>	272_273insCAGG	PMID:29523322	Fujiwara et al., 2018
PTM	Resistant	<i>fbiA</i>	337_338insT	PMID:26100695	Haver et al., 2015
PTM	Resistant	<i>fbiA</i>	347_347del	PMID:26100695	Haver et al., 2015
PTM	Resistant	<i>fbiA</i>	L119P	PMID:26100695	Haver et al., 2015
DLM/PTM	Resistant	<i>fbiA</i>	Q120P	PMID:33077652	Rifat et al., 2020
PTM	Resistant	<i>fbiA</i>	S126P	PMID:26100695	Haver et al., 2015
PTM	Resistant	<i>fbiA</i>	W136R	PMID:26100695	Haver et al., 2015
PTM	Resistant	<i>fbiA</i>	T146A	PMID:26100695	Haver et al., 2015
DLM	Resistant	<i>fbiA</i>	452_452del	PMID:29523322	Fujiwara et al., 2018
DLM	Resistant	<i>fbiA</i>	V154I	PMID:32907992	Battaglia et al., 2020
DLM	Resistant	<i>fbiA</i>	P159Q	PMID:32907992	Battaglia et al., 2020
DLM	Resistant	<i>fbiA</i>	G164fs	PMID:26100695	Haver et al., 2015
PTM	Resistant	<i>fbiA</i>	W172R	PMID:26100695	Haver et al., 2015
PTM	Resistant	<i>fbiA</i>	562_563insT	PMID:26100695	Haver et al., 2015
PTM	Resistant	<i>fbiA</i>	571_572insA	PMID:26100695	Haver et al., 2015
DLM	Resistant	<i>fbiA</i>	I208V	PMID:32907992	Battaglia et al., 2020
DLM	Resistant	<i>fbiA</i>	I209V	PMID:32907992	Battaglia et al., 2020
PTM	Resistant	<i>fbiA</i>	A238E	PMID:26100695	Haver et al., 2015
DLM	Resistant	<i>fbiA</i>	K250*	PMID:27076101	Schena et al., 2016
PTM	Resistant	<i>fbiA</i>	C259R	PMID:26100695	Haver et al., 2015
PTM	Resistant	<i>fbiA</i>	G283R	PMID:26100695	Haver et al., 2015
DLM/PTM	Resistant	<i>fbiA</i>	D286A	PMID:33077652	Rifat et al., 2020
DLM	Resistant	<i>fbiA</i>	C287*	PMID:32907992	Battaglia et al., 2020
DLM	Resistant	<i>fbiA</i>	R304Q	PMID:32907992	Battaglia et al., 2020
DLM/PTM	Resistant	<i>fbiA</i>	L308P	PMID:33077652	Rifat et al., 2020
PTM	Resistant	<i>fbiA</i>	G323V	PMID:26100695	Haver et al., 2015
PTM	Resistant	<i>fbiB</i>	36_37ins17bp	PMID:26100695	Haver et al., 2015
PTM	Resistant	<i>fbiB</i>	W39*	PMID:26100695	Haver et al., 2015
PTM	Resistant	<i>fbiB</i>	G153V	PMID:26100695	Haver et al., 2015
DLM	Resistant	<i>fbiB</i>	G221S	PMID:32907992	Battaglia et al., 2020
DLM	Resistant	<i>fbiB</i>	D224N	PMID:32907992	Battaglia et al., 2020
DLM	Resistant	<i>fbiB</i>	G273R	PMID:32907992	Battaglia et al., 2020
PTM	Resistant	<i>fbiB</i>	P361A	PMID:26100695	Haver et al., 2015
DLM	Resistant	<i>fbiB</i>	1148_1155del	PMID:29523322	Fujiwara et al., 2018
DLM	Resistant	<i>fbiB</i>	1263_1264del	PMID:29523322	Fujiwara et al., 2018
PTM	Resistant	<i>fbiC</i>	52_52del	PMID:26100695	Haver et al., 2015
DLM/PTM	Resistant	<i>fbiC</i>	60_60del	PMID:33077652	Rifat et al., 2020

PTM	Resistant	<i>fbiC</i>	A50P	PMID:26100695	Haver et al., 2015
DLM	Resistant	<i>fbiC</i>	154_154del	PMID:26100695	Haver et al., 2015
PTM	Resistant	<i>fbiC</i>	E54M	PMID:26100695	Haver et al., 2015
DLM	Resistant	<i>fbiC</i>	N58T	PMID:29523322	Fujiwara et al., 2018
PTM	Resistant	<i>fbiC</i>	Y86*	PMID:26100695	Haver et al., 2015
PTM	Resistant	<i>fbiC</i>	F91V	PMID:26100695	Haver et al., 2015
DLM	Resistant	<i>fbiC</i>	C98Y	PMID:29523322	Fujiwara et al., 2018
DLM	Resistant	<i>fbiC</i>	Y104C	PMID:32907992	Battaglia et al., 2020
PTM	Resistant	<i>fbiC</i>	C105R	PMID:26100695	Haver et al., 2015
DLM	Resistant	<i>fbiC</i>	G112A	PMID:32907992	Battaglia et al., 2020
PTM	Resistant	<i>fbiC</i>	491_496del	PMID:26100695	Haver et al., 2015
PTM	Resistant	<i>fbiC</i>	H190R	PMID:26100695	Haver et al., 2015
DLM/PTM	Resistant	<i>fbiC</i>	G194D	PMID:33077652	Rifat et al., 2020
PTM	Resistant	<i>fbiC</i>	S202P	PMID:26100695	Haver et al., 2015
PTM	Resistant	<i>fbiC</i>	L204P	PMID:26100695	Haver et al., 2015
PTM	Resistant	<i>fbiC</i>	S210P	PMID:26100695	Haver et al., 2015
PTM	Resistant	<i>fbiC</i>	E216A	PMID:26100695	Haver et al., 2015
DLM	Resistant	<i>fbiC</i>	R220*	PMID:29523322	Fujiwara et al., 2018
DLM	Resistant	<i>fbiC</i>	L228F	PMID:32907992	Battaglia et al., 2020
DLM	Resistant	<i>fbiC</i>	699_699del	PMID:29523322	Fujiwara et al., 2018
DLM	Resistant	<i>fbiC</i>	811_811del	PMID:29523322	Fujiwara et al., 2018
DLM	Resistant	<i>fbiC</i>	812_812del	PMID:29523322	Fujiwara et al., 2018
PTM	Resistant	<i>fbiC</i>	T273R	PMID:26100695	Haver et al., 2015
DLM	Resistant	<i>fbiC</i>	S280L	PMID:32907992	Battaglia et al., 2020
PTM	Resistant	<i>fbiC</i>	830_831insA	PMID:26100695	Haver et al., 2015
PTM	Resistant	<i>fbiC</i>	845_846insG	PMID:26100695	Haver et al., 2015
DLM	Resistant	<i>fbiC</i>	V318I	PMID:28739779	Pang et al., 2017
PTM	Resistant	<i>fbiC</i>	N336K	PMID:26100695	Haver et al., 2015
PTM	Resistant	<i>fbiC</i>	G356C	PMID:26100695	Haver et al., 2015
PTM	Resistant	<i>fbiC</i>	S358A	PMID:26100695	Haver et al., 2015
PTM	Resistant	<i>fbiC</i>	P372S	PMID:26100695	Haver et al., 2015
DLM/PTM	Resistant	<i>fbiC</i>	L377P	PMID:33077652	Rifat et al., 2020
PTM	Resistant	<i>fbiC</i>	G385V	PMID:26100695	Haver et al., 2015
PTM	Resistant	<i>fbiC</i>	D387Y	PMID:26100695	Haver et al., 2015
DLM	Resistant	<i>fbiC</i>	1337_1337del	PMID:29523322	Fujiwara et al., 2018
DLM	Resistant	<i>fbiC</i>	P523L	PMID:32907992	Battaglia et al., 2020
DLM/PTM	Resistant	<i>fbiC</i>	C562W	PMID:33077652	Rifat et al., 2020
DLM	Resistant	<i>fbiC</i>	R563L	PMID:29941636	Rancoita et al., 2018
PTM	Resistant	<i>fbiC</i>	V630E	PMID:16387854	Manjunatha et al., 2006
PTM	Resistant	<i>fbiC</i>	H631Y	PMID:26100695	Haver et al., 2015
DLM/PTM	Resistant	<i>fbiC</i>	K684T	PMID:33077652	Rifat et al., 2020
PTM	Resistant	<i>fbiC</i>	2127_2128del	PMID:26100695	Haver et al., 2015
PTM	Resistant	<i>fbiC</i>	M708I	PMID:26100695	Haver et al., 2015

PTM	Resistant	<i>fbiC</i>	G711W	PMID:26100695	Haver et al., 2015
PTM	Resistant	<i>fbiC</i>	2131_2131del	PMID:26100695	Haver et al., 2015
PTM	Resistant	<i>fbiC</i>	S715R	PMID:26100695	Haver et al., 2015
PTM	Resistant	<i>fbiC</i>	W719L	PMID:26100695	Haver et al., 2015
PTM	Resistant	<i>fbiC</i>	V720I	PMID:26100695	Haver et al., 2015
PTM	Resistant	<i>fbiC</i>	H722R	PMID:26100695	Haver et al., 2015
DLM	Resistant	<i>fbiC</i>	N724S	PMID:32907992	Battaglia et al., 2020
PTM	Resistant	<i>fbiC</i>	2274_2275insG	PMID:26100695	Haver et al., 2015
DLM	Resistant	<i>fbiC</i>	S762N	PMID:32907992	Battaglia et al., 2020
DLM	Resistant	<i>fbiC</i>	792insQTSWVKL	PMID:26100695	Haver et al., 2015
DLM	Resistant	<i>fbiC</i>	L800R	PMID:26100695	Haver et al., 2015
DLM/PTM	Resistant	<i>fbiC</i>	A827G	PMID:33077652	Rifat et al., 2020
DLM	Resistant	<i>fbiC</i>	2548_2549insC	PMID:26100695	Haver et al., 2015
DLM	Resistant	<i>fbiC</i>	A835V	PMID:32907992	Battaglia et al., 2020
DLM	Resistant	<i>fbiC</i>	A855fs	PMID:32907992	Battaglia et al., 2020
DLM	Resistant	<i>fbiC</i>	A856P	PMID:32907992	Battaglia et al., 2020
PTM	Resistant	<i>fbiC</i>	2734_2735insAACTT	PMID:26100695	Haver et al., 2015
PTM	Resistant	<i>fbiC</i>	1304052_1304452del	PMID:16387854	Manjunatha et al., 2006
DLM/PTM	Resistant	<i>fbiD</i>	79_80insC	PMID:33077652	Rifat et al., 2020
PTM	Resistant	<i>fbiD</i>	A132V	PMID:33077652	Rifat et al., 2020
PTM	Resistant	<i>fbiD</i>	G147C	PMID:33077652	Rifat et al., 2020
DLM	Susceptible	<i>ddn</i>	-32T>C	PMID:32907992	Battaglia et al., 2020
DLM	Susceptible	<i>ddn</i>	-26G>A	PMID:32907992	Battaglia et al., 2020
DLM	Susceptible	<i>ddn</i>	-24C>A	PMID:32907992	Battaglia et al., 2020
DLM	Susceptible	<i>ddn</i>	-15G>A	PMID:32907992	Battaglia et al., 2020
DLM/PTM	Susceptible	<i>ddn</i>	P6S	PMID:32032366	Lee et al., 2019
DLM/PTM	Susceptible	<i>ddn</i>	P6T	PMID:32032366	Lee et al., 2019
DLM/PTM	Susceptible	<i>ddn</i>	P6L	PMID:32032366	Lee et al., 2019
DLM/PTM	Susceptible	<i>ddn</i>	M21T	PMID:32032366	Lee et al., 2019
DLM/PTM	Susceptible	<i>ddn</i>	R23L	PMID:32032366	Lee et al., 2019
DLM/PTM	Susceptible	<i>ddn</i>	R23W	PMID:32032366	Lee et al., 2019
DLM/PTM	Susceptible	<i>ddn</i>	T26P	PMID:32032366	Lee et al., 2019
DLM/PTM	Susceptible	<i>ddn</i>	W27C	PMID:32032366	Lee et al., 2019
DLM/PTM	Susceptible	<i>ddn</i>	Y29H	PMID:32032366	Lee et al., 2019
DLM/PTM	Susceptible	<i>ddn</i>	Y29S	PMID:32032366	Lee et al., 2019
DLM/PTM	Susceptible	<i>ddn</i>	R30S	PMID:32032366	Lee et al., 2019
DLM/PTM	Susceptible	<i>ddn</i>	G34E	PMID:32032366	Lee et al., 2019
DLM/PTM	Susceptible	<i>ddn</i>	G34R	PMID:32032366	Lee et al., 2019
DLM/PTM	Susceptible	<i>ddn</i>	G36V	PMID:32032366	Lee et al., 2019
DLM/PTM	Susceptible	<i>ddn</i>	P45L	PMID:32032366	Lee et al., 2019
DLM/PTM	Susceptible	<i>ddn</i>	T50P	PMID:32032366	Lee et al., 2019
DLM	Susceptible	<i>ddn</i>	T50I	PMID:32907992	Battaglia et al., 2020
DLM/PTM	Susceptible	<i>ddn</i>	T51P	PMID:32032366	Lee et al., 2019

DLM/PTM	Susceptible	<i>ddn</i>	T52N	PMID:32032366	Lee et al., 2019
DLM/PTM	Susceptible	<i>ddn</i>	T52P	PMID:32032366	Lee et al., 2019
DLM/PTM	Susceptible	<i>ddn</i>	T56P	PMID:32032366	Lee et al., 2019
DLM/PTM	Susceptible	<i>ddn</i>	G57A	PMID:32032366	Lee et al., 2019
DLM/PTM**	Susceptible	<i>ddn</i>	V61G	PMID:32032366	Lee et al., 2019
DLM/PTM	Susceptible	<i>ddn</i>	N62D	PMID:32032366	Lee et al., 2019
DLM	Susceptible	<i>ddn</i>	Y65L	PMID:32032366	Lee et al., 2019
DLM	Susceptible	<i>ddn</i>	Y65M	PMID:32032366	Lee et al., 2019
DLM	Susceptible	<i>ddn</i>	Y65C	PMID:32032366	Lee et al., 2019
DLM/PTM	Susceptible	<i>ddn</i>	Y65F	PMID:32032366	Lee et al., 2019
DLM/PTM	Susceptible	<i>ddn</i>	L67P	PMID:32032366	Lee et al., 2019
DLM/PTM	Susceptible	<i>ddn</i>	D69N	PMID:32032366	Lee et al., 2019
DLM/PTM	Susceptible	<i>ddn</i>	G71R	PMID:32032366	Lee et al., 2019
DLM/PTM	Susceptible	<i>ddn</i>	R72Q	PMID:32032366	Lee et al., 2019
DLM/PTM	Susceptible	<i>ddn</i>	R72W	PMID:27076101	Schena et al., 2016
DLM	Susceptible	<i>ddn</i>	S78A	PMID:32032366	Lee et al., 2019
DLM	Susceptible	<i>ddn</i>	S78C	PMID:32032366	Lee et al., 2019
DLM	Susceptible	<i>ddn</i>	S78T	PMID:32032366	Lee et al., 2019
DLM	Susceptible	<i>ddn</i>	S78V	PMID:32032366	Lee et al., 2019
DLM	Susceptible	<i>ddn</i>	S78Y	PMID:32032366	Lee et al., 2019
DLM/PTM	Susceptible	<i>ddn</i>	G81S	PMID:32032366	Lee et al., 2019
DLM/PTM	Susceptible	<i>ddn</i>	E83D	PMID:32032366	Lee et al., 2019
DLM	Susceptible	<i>ddn</i>	E83Q	PMID:32907992	Battaglia et al., 2020
DLM/PTM	Susceptible	<i>ddn</i>	L90V	PMID:32032366	Lee et al., 2019
DLM/PTM**	Susceptible	<i>ddn</i>	N91T	PMID:32032366	Lee et al., 2019
DLM	Susceptible	<i>ddn</i>	K93Q	PMID:32907992	Battaglia et al., 2020
DLM/PTM	Susceptible	<i>ddn</i>	I102V	PMID:32032366	Lee et al., 2019
DLM/PTM	Susceptible	<i>ddn</i>	E105Q	PMID:32032366	Lee et al., 2019
DLM	Susceptible	<i>ddn</i>	T110I	PMID:32907992	Battaglia et al., 2020
DLM/PTM	Susceptible	<i>ddn</i>	A111V	PMID:32032366	Lee et al., 2019
DLM/PTM	Susceptible	<i>ddn</i>	D113N	PMID:32032366	Lee et al., 2019
DLM/PTM	Susceptible	<i>ddn</i>	E117K	PMID:32032366	Lee et al., 2019
DLM/PTM	Susceptible	<i>ddn</i>	P124S	PMID:32032366	Lee et al., 2019
DLM/PTM	Susceptible	<i>ddn</i>	Y130C	PMID:32032366	Lee et al., 2019
DLM/PTM	Susceptible	<i>ddn</i>	Y130D	PMID:32032366	Lee et al., 2019
DLM/PTM	Susceptible	<i>ddn</i>	Y130F	PMID:32032366	Lee et al., 2019
DLM/PTM	Susceptible	<i>ddn</i>	Y130H	PMID:32032366	Lee et al., 2019
DLM/PTM	Susceptible	<i>ddn</i>	Y130N	PMID:32032366	Lee et al., 2019
DLM/PTM	Susceptible	<i>ddn</i>	Y130S	PMID:32032366	Lee et al., 2019
DLM/PTM	Susceptible	<i>ddn</i>	Y130W	PMID:32032366	Lee et al., 2019
DLM	Susceptible	<i>ddn</i>	Y133C	PMID:32032366	Lee et al., 2019
DLM/PTM	Susceptible	<i>ddn</i>	Y133F	PMID:32032366	Lee et al., 2019
DLM	Susceptible	<i>ddn</i>	Y133L	PMID:32032366	Lee et al., 2019

DLM	Susceptible	<i>ddn</i>	Y133M	PMID:32032366	Lee et al., 2019
DLM	Susceptible	<i>ddn</i>	Y133W	PMID:32032366	Lee et al., 2019
DLM	Susceptible	<i>ddn</i>	Y136E	PMID:32032366	Lee et al., 2019
DLM	Susceptible	<i>ddn</i>	Y136T	PMID:32032366	Lee et al., 2019
DLM/PTM	Susceptible	<i>ddn</i>	Y136F	PMID:32032366	Lee et al., 2019
DLM/PTM**	Susceptible	<i>ddn</i>	T140I	PMID:32032366	Lee et al., 2019
DLM/PTM	Susceptible	<i>ddn</i>	V147M	PMID:32032366	Lee et al., 2019
DLM	Susceptible	<i>ddn</i>	C149Y	PMID:32032366	Lee et al., 2019
DLM	Susceptible	<i>fgd1</i>	R18G	PMID:32907992	Battaglia et al., 2020
DLM	Susceptible	<i>fgd1</i>	R18S	PMID:32907992	Battaglia et al., 2020
DLM	Susceptible	<i>fgd1</i>	E19K	PMID:32907992	Battaglia et al., 2020
DLM	Susceptible	<i>fgd1</i>	A60G	PMID:32907992	Battaglia et al., 2020
DLM	Susceptible	<i>fgd1</i>	M93T	PMID:32907992	Battaglia et al., 2020
DLM	Susceptible	<i>fgd1</i>	P98L	PMID:32907992	Battaglia et al., 2020
DLM	Susceptible	<i>fgd1</i>	R187H	PMID:32907992	Battaglia et al., 2020
DLM	Susceptible	<i>fgd1</i>	I225V	PMID:32907992	Battaglia et al., 2020
DLM	Susceptible	<i>fgd1</i>	K270M	PMID:27076101	Schena et al., 2016
DLM	Susceptible	<i>fgd1</i>	A287V	PMID:32907992	Battaglia et al., 2020
DLM	Susceptible	<i>fgd1</i>	K296E	PMID:27076101	Schena et al., 2016
DLM	Susceptible	<i>fgd1</i>	Q299E	PMID:32907992	Battaglia et al., 2020
DLM	Susceptible	<i>fbIA</i>	A43T	PMID:32907992	Battaglia et al., 2020
DLM	Susceptible	<i>fbIA</i>	V47I	PMID:32907992	Battaglia et al., 2020
DLM	Susceptible	<i>fbIA</i>	V58I	PMID:32907992	Battaglia et al., 2020
DLM	Susceptible	<i>fbIA</i>	D74E	PMID:32907992	Battaglia et al., 2020
DLM	Susceptible	<i>fbIA</i>	Q120R	PMID:27076101	Schena et al., 2016
DLM	Susceptible	<i>fbIA</i>	R175H	PMID:26559594	Bloemberg et al., 2015
DLM	Susceptible	<i>fbIA</i>	S184T	PMID:32907992	Battaglia et al., 2020
DLM	Susceptible	<i>fbIA</i>	S219G	PMID:33077652	Rifat et al., 2020
DLM	Susceptible	<i>fbIA</i>	I247V	PMID:32907992	Battaglia et al., 2020
DLM	Susceptible	<i>fbIA</i>	T302M	PMID:27076101	Schena et al., 2016
DLM	Susceptible	<i>fbIA</i>	T302P	PMID:32907992	Battaglia et al., 2020
DLM	Susceptible	<i>fbIA</i>	D312G	PMID:32907992	Battaglia et al., 2020
DLM	Susceptible	<i>fbIA</i>	M319I	PMID:32907992	Battaglia et al., 2020
DLM	Susceptible	<i>fbIB</i>	L15P	PMID:33077652	Rifat et al., 2020
DLM	Susceptible	<i>fbIB</i>	L15R	PMID:32907992	Battaglia et al., 2020
DLM	Susceptible	<i>fbIB</i>	P16R	PMID:32907992	Battaglia et al., 2020
DLM	Susceptible	<i>fbIB</i>	V17I	PMID:32907992	Battaglia et al., 2020
DLM	Susceptible	<i>fbIB</i>	V48A	PMID:32907992	Battaglia et al., 2020
DLM	Susceptible	<i>fbIB</i>	D66E	PMID:32907992	Battaglia et al., 2020
DLM	Susceptible	<i>fbIB</i>	A82T	PMID:32907992	Battaglia et al., 2020
DLM	Susceptible	<i>fbIB</i>	D90N	PMID:32907992	Battaglia et al., 2020
DLM	Susceptible	<i>fbIB</i>	A155T	PMID:32907992	Battaglia et al., 2020
DLM	Susceptible	<i>fbIB</i>	L173P	PMID:33077652	Rifat et al., 2020

DLM	Susceptible	<i>fbiB</i>	F220L	PMID:27076101	Schena et al., 2016
DLM	Susceptible	<i>fbiB</i>	R230Q	PMID:32907992	Battaglia et al., 2020
DLM	Susceptible	<i>fbiB</i>	G236D	PMID:32907992	Battaglia et al., 2020
DLM	Susceptible	<i>fbiB</i>	D315A	PMID:32907992	Battaglia et al., 2020
DLM	Susceptible	<i>fbiB</i>	R333C	PMID:32907992	Battaglia et al., 2020
DLM	Susceptible	<i>fbiB</i>	W397R	PMID:33077652	Rifat et al., 2020
DLM	Susceptible	<i>fbiB</i>	G399S	PMID:32907992	Battaglia et al., 2020
DLM	Susceptible	<i>fbiB</i>	R409S	PMID:32907992	Battaglia et al., 2020
DLM	Susceptible	<i>fbiB</i>	L447R	PMID:27076101	Schena et al., 2016
DLM	Susceptible	<i>fbiB</i>	K448R	PMID:27076101	Schena et al., 2016
DLM	Susceptible	<i>fbiC</i>	-28T>C	PMID:32907992	Battaglia et al., 2020
DLM	Susceptible	<i>fbiC</i>	-27A>G	PMID:32907992	Battaglia et al., 2020
DLM	Susceptible	<i>fbiC</i>	-11G>A	PMID:32907992	Battaglia et al., 2020
DLM	Susceptible	<i>fbiC</i>	V16I	PMID:32907992	Battaglia et al., 2020
DLM	Susceptible	<i>fbiC</i>	R25G	PMID:33077652	Rifat et al., 2020
DLM	Susceptible	<i>fbiC</i>	V41M	PMID:32907992	Battaglia et al., 2020
DLM	Susceptible	<i>fbiC</i>	D168E	PMID:32907992	Battaglia et al., 2020
DLM	Susceptible	<i>fbiC</i>	V181M	PMID:32907992	Battaglia et al., 2020
DLM	Susceptible	<i>fbiC</i>	D235N	PMID:32907992	Battaglia et al., 2020
DLM	Susceptible	<i>fbiC</i>	D272G	PMID:32907992	Battaglia et al., 2020
DLM	Susceptible	<i>fbiC</i>	T273A	PMID:27076101	Schena et al., 2016
DLM	Susceptible	<i>fbiC</i>	M329I	PMID:32907992	Battaglia et al., 2020
DLM	Susceptible	<i>fbiC</i>	A333V	PMID:32907992	Battaglia et al., 2020
DLM	Susceptible	<i>fbiC</i>	V389L	PMID:32907992	Battaglia et al., 2020
DLM	Susceptible	<i>fbiC</i>	R463C	PMID:32907992	Battaglia et al., 2020
DLM	Susceptible	<i>fbiC</i>	D465H	PMID:32907992	Battaglia et al., 2020
DLM	Susceptible	<i>fbiC</i>	D465A	PMID:32907992	Battaglia et al., 2020
DLM	Susceptible	<i>fbiC</i>	T519I	PMID:32907992	Battaglia et al., 2020
DLM	Susceptible	<i>fbiC</i>	A524G	PMID:32907992	Battaglia et al., 2020
DLM	Susceptible	<i>fbiC</i>	T555I	PMID:32907992	Battaglia et al., 2020
DLM	Susceptible	<i>fbiC</i>	V581L	PMID:32907992	Battaglia et al., 2020
DLM	Susceptible	<i>fbiC</i>	E608A	PMID:32907992	Battaglia et al., 2020
DLM	Susceptible	<i>fbiC</i>	A620T	PMID:32907992	Battaglia et al., 2020
DLM	Susceptible	<i>fbiC</i>	E658D	PMID:32907992	Battaglia et al., 2020
DLM	Susceptible	<i>fbiC</i>	D674H	PMID:32907992	Battaglia et al., 2020
DLM	Susceptible	<i>fbiC</i>	W678G	PMID:32907992	Battaglia et al., 2020
DLM	Susceptible	<i>fbiC</i>	T681I	PMID:27076101	Schena et al., 2016
DLM	Susceptible	<i>fbiC</i>	I693V	PMID:32907992	Battaglia et al., 2020
DLM	Susceptible	<i>fbiC</i>	M776T	PMID:33077652	Rifat et al., 2020
DLM	Susceptible	<i>fbiC</i>	T850I	PMID:32907992	Battaglia et al., 2020
DLM	Susceptible	<i>fbiC</i>	A856S	PMID:32907992	Battaglia et al., 2020
DLM	Susceptible	<i>fbiD</i>	R25S	PMID:33077652	Rifat et al., 2020
DLM	Susceptible	<i>fbiD</i>	A68E	PMID:33077652	Rifat et al., 2020

DLM	Susceptible	<i>fbtD</i>	Q114R	PMID:33077652	Rifat et al., 2020
DLM	Susceptible	<i>fbtD</i>	385_387del	PMID:33077652	Rifat et al., 2020
DLM	Susceptible	<i>fbtD</i>	C152R	PMID:33077652	Rifat et al., 2020
DLM	Susceptible	<i>fbtD</i>	A198P	PMID:33077652	Rifat et al., 2020

* In Bold: present in our data; ** where there are discrepancies between different studies in MIC; Delamanid (DLM); Pretomanid (PTM)

S4 Table. Phylogenetic mutations with >50% of allele frequency within a sub-lineage.

Mutation	Gene	Freq	Sub-lineage (# isolates)	Freq (%) in sub-lineage	#sub-lin.	Max SNP dist.	# Indep. Occur.	Susc. %	MDR /XDR %	Pre-2014 % *
<u>K270M</u>	<i>fgd1</i>	3136	4.1.2 ⁺ (3135); 2.2.1(1)	96.0	3	1329	2	70.1	18.1	84.2
-32A>G	<i>fbiC</i>	639	5,6, <i>Bov</i> (634); 2.2.1(2); 4.3.3(1); 4.2(1); 4.9(1)	98.4	7	3264	5	60.1	8.3	63.1
R64S	<i>fgd1</i>	471	1.1.1 ⁺ (471)	54.4	2	515	1	77.9	2.1	99.1
<u>T302M</u>	<i>fbiA</i>	355	4.1.1.1(355)	99.7	1	337	1	82.8	9.9	84.8
<u>D113N</u>	<i>ddn</i>	267	5(264); 2.2.1(3)	100	2	1402	2	70.7	15.4	91.7
<u>K296E</u>	<i>fgd1</i>	162	6(161); 4.1.2.1(1)	98.2	2	933	2	87.0	3.7	85.7
A505T	<i>fbiC</i>	135	2.1(135)	100	1	486	1	61.5	20.0	95.1
<u>P69L</u>	<i>pepQ</i>	141	4.4.1.2(141)	100	1	284	1	89.4	1.4	92.1

Drug resistance (%): Susc. = Susceptible; * % of number of samples pre-2014/total number of samples with available collection date; mutations associated with no significant change in minimum inhibitory concentration are underlined (with MIC usually <0.06 mg/L for BDQ and <0.2 mg/L for DLM/PTM; see **S3 Table**); Bedaquiline (BDQ), delamanid (DLM); pretomanid (PTM).

S5 Table. All mutations (in >1 isolate) in bedaquiline (BDQ) candidate genes found in the 33k isolates

Mutation	Gene	Freq	Sub-lineage(# isolates)	# sub-lin.	# Indep. Occur.	Susc .%	MDR /XDR %	Pre-2014 % *	Functional Support **
<u>P69L</u>	<i>pepQ</i>	141	4.4.1.2(141)	1	1	89.4	1.4	92.1	P
<u>-11C>A</u>	<i>mmpR5</i>	124	2.2.1(122); 4.3.2.1(1); 1.1.1(1)	3	3	12.1	76.6	93.1	-
192_193insG (I67fs)	<i>mmpR5</i>	44	4(34); 2.2.1(4); 3(2); 4.9(1); 4.8(1); 4.5(1); 1.1.1(1)	7	10	0	86.4	100	-
G197R	<i>pepQ</i>	38	4.3.4.1(37); 2.2.1(1)	2	2	52.6	47.4	72.2	S,P
<u>R7Q</u>	<i>pepQ</i>	35	3(35)	1	1	68.6	22.9	66.7	-
T354A	<i>pepQ</i>	27	3(27)	1	1	100	0	0	-
K94N	<i>pepQ</i>	23	3.1.1(22); 4.1.2(1)	2	2	95.7	0	100	-
M146T	<i>mmpR5</i>	21	4.4.1.1(20); 2.2.2(1)	2	2	0	100	-	S,M
<u>D5G</u>	<i>mmpR5</i>	18	2.2.1(17); 4.1.2.1(1)	2	2	94.4	0	75.0	-
E44D	<i>atpE</i>	17	2.2.1(17)	1	1	94.1	0	75.0	B,S
A242T	<i>pepQ</i>	17	2.2.1.1(17)	1	1	58.8	5.9	100	-
193_193del (I67fs)	<i>mmpR5</i>	16	4.3.4.2(10); 2.2.1(3); 4.7(2); 4.3.3.1(1)	4	5	0	100	83.3	-
D20G	<i>pepQ</i>	15	4.6(15)	1	1	100	0	20.0	P
141_142insC	<i>mmpR5</i>	15	2.2 ⁺ (8); 4.1.2 ⁺ (2); 4.3 ⁺ (2); 4.4.1.1(1); 3(2)	8	11	6.7	86.7	85.7	-
-49T>C	<i>mmpR5</i>	12	3.1.2.1(12)	1	1	75.0	8.3	0	-
<u>G87R</u>	<i>mmpR5</i>	11	1.1.2(11)	1	1	100	0	80.0	S,P
<u>V20A</u>	<i>mmpR5</i>	10	4.1.2.1(8); 4.3.2.1(1); 2.2.1(1)	3	3	90.0	10.0	83.3	M
L117R	<i>mmpR5</i>	9	3(5); 4.3.4.2(2); 4.2.2(1); 4.1(1)	4	5	44.4	44.4	100	S
<u>N4T</u>	<i>mmpR5</i>	9	3(9)	1	1	55.5	33.3	-	-
<u>V3I</u>	<i>mmpR5</i>	9	4.3.4.2(9)	1	1	22.2	66.7	-	-
V211A	<i>pepQ</i>	9	3.1.1(9)	1	1	100	0	100	-
E115A	<i>pepQ</i>	8	4.1.2.1(8)	1	1	100	0	0	-
L32S	<i>mmpR5</i>	8	2.2.1(8)	1	3	0	87.5	50.0	S,M
138_139insG	<i>mmpR5</i>	7	2.2.1(7)	1	1	0	100	100	-
<u>D141H</u>	<i>mmpR5</i>	7	2.2.1(6); 1.1.3(1)	2	2	14.3	57.1	100	S,P
R90C	<i>mmpR5</i>	7	2.2.1(6); 4.1.1.3(1)	2	4	85.7	0	50.0	-
418_419insG	<i>mmpR5</i>	7	4.1.2.1(7)	1	1	0	0	-	-
G121R	<i>mmpR5</i>	7	2.2.2(5); 3(1); 4.4.1.1(1)	3	3	0	100	100	S,P
<u>T341A</u>	<i>pepQ</i>	6	2.1(6)	1	1	50	33.3	100	P
D119E	<i>mmpR5</i>	6	4.9(6)	1	1	100	0	0	-
S2R	<i>mmpR5</i>	6	3(6)	1	1	33.3	66.7	-	-
G126D	<i>mmpR5</i>	5	1.2.1(5)	1	1	100	0	50	-
V298I	<i>pepQ</i>	5	4.8(5)	1	1	100	0	100	-

A153G	<i>pepQ</i>	5	4.7(5)	1	1	100	0	-	-
N98D	<i>mmpR5</i>	5	4.1.2.1(2); 4.4.1.1(2); 2.2.1(1)	3	3	0	80.0	100	-
<u>V85G</u>	<i>mmpR5</i>	5	2.2.1(5)	1	1	0	100	100	-
T363I	<i>pepQ</i>	5	4.8(5)	1	1	100	0	100	S,P
<u>E55D</u>	<i>mmpR5</i>	4	2.2.1(4)	1	1	75.0	25.0	100	-
G162E	<i>mmpR5</i>	4	1.1.2(4)	1	1	100	0	100	-
S99R	<i>pepQ</i>	4	4.2.1(4)	1	1	50.0	0	-	-
V149I	<i>mmpR5</i>	4	1.1.2(4)	1	1	0	0	-	-
A224V	<i>pepQ</i>	4	5(2); 2.2.1.1(1); 1.2.2(1)	3	3	75.0	25.0	50.0	-
D26A	<i>pepQ</i>	4	4.3.3(4)	1	1	0	100	100	P
A196V	<i>pepQ</i>	4	2.2.1(4)	1	2	25.0	75.0	100	-
I193T	<i>pepQ</i>	4	3(4)	1	1	25.0	25.0	-	S,P
N148H	<i>mmpR5</i>	4	1.1.2(4)	1	1	100	0	75.0	-
V39I	<i>atpE</i>	4	4.3.3(4)	1	1	100	0	0	-
-29G>A	<i>mmpR5</i>	3	3(2); 1.2.2(1)	2	2	100	0	66.7	-
R109W	<i>mmpR5</i>	3	3(2); 1.2.2(1)	2	2	66.7	33.3	100	-
M111T	<i>mmpR5</i>	3	1.1.2(3)	1	1	100	0	-	M
T341I	<i>pepQ</i>	3	3(2); 1.1.2(1)	2	2	100	0	-	P
-37T>C	<i>mmpR5</i>	3	4.3.4.1(3)	1	1	100	0	100	-
G41A	<i>mmpR5</i>	3	4.3.2(3)	1	1	33.3	0	-	-
G41C	<i>pepQ</i>	3	4.7(3)	1	1	100	0	-	-
A243V	<i>pepQ</i>	3	5(3)	1	1	100	0	0	-
G126S	<i>mmpR5</i>	3	1.2.1(3)	1	1	0	100	0	-
-3C>CT	<i>mmpR5</i>	3	4.3.3(3)	1	1	0	100	100	-
S53L	<i>mmpR5</i>	3	4.1.1.3(2); 2.2.1(1)	2	2	66.7	0	-	-
<u>V120M</u>	<i>mmpR5</i>	3	4.1.1(1); 1.1.2(1); 1.2.1(1)	3	3	66.7	0	50	-
16_16del	<i>mmpR5</i>	3	2.2.2(1); 2.2.1+(2)	3	3	33.3	66.7	100	-
T56I	<i>pepQ</i>	3	4.5(2); 2.2.1(1)	2	2	33.3	66.7	100	P
-30CG>C	<i>mmpR5</i>	3	4.5(3)	1	1	0	100	100	-
-21T>C	<i>mmpR5</i>	3	4.5(3)	1	1	0	100	100	-
S63N	<i>mmpR5</i>	3	4.4.2(2)	1	1	0	100	100	S
-									
31GGCTACC AGA>G	<i>atpE</i>	3	4.4.2(3)	1	1	0	0	100	-
A59T	<i>mmpR5</i>	3	2.2.1(3)	1	1	0	100	100	-
-									
38ATACCGA ACG>A	<i>mmpR5</i>	3	1.1(3)	1	1	66.7	0	-	-
L163V	<i>pepQ</i>	3	1.1.1(3)	1	1	100	0	-	-
T2K	<i>pepQ</i>	3	4.2.2(3)	1	1	66.7	0	-	-
A128V	<i>mmpR5</i>	3	1.2.2(3)	1	1	100	0	0	-
D283G	<i>pepQ</i>	3	4.1.2.1(3)	1	1	100	0	33.3	-
D26G	<i>pepQ</i>	3	3(2); 2.2.1(1)	2	3	33.3	33.3	100	P

V343L	<i>pepQ</i>	3	1.2.2(3)	1	1	100	0	100	-
<u>M23V</u>	<i>mmpR5</i>	3	2.2.1(3)	1	1	0	0	100	-
-9G>C	<i>mmpR5</i>	3	4.1.2(3)	1	1	100	0	100	-
P97L	<i>mmpR5</i>	2	4.1(1); 3(1)	2	2	100	0	-	P
T78I	<i>atpE</i>	2	3(2)	1	1	100	0	-	B
V80A	<i>atpE</i>	2	3(2)	1	1	100	0	-	B
M139T	<i>mmpR5</i>	2	4.5(1); 2.2.1(1)	2	2	50.0	50.0	100	M
<u>A84V</u>	<i>mmpR5</i>	2	4.3.4.2.1(1); 2.2.1(1)	2	2	50.0	50.0	-	-
Y145H	<i>mmpR5</i>	2	3(2)	1	1	0	100	-	S,M
-41C>G	<i>mmpR5</i>	2	4.1.2.1(2)	1	1	100	0	-	-
V85I	<i>mmpR5</i>	2	4.8(2)	1	1	100	0	-	-
P129S	<i>mmpR5</i>	2	4.8(1); 2.1(1)	2	2	50.0	0	100	-
V158L	<i>pepQ</i>	2	4.4.1.1(2)	1	1	100	0	100	-
T91I	<i>mmpR5</i>	2	5(2)	1	1	0	100	-	-
L74M	<i>mmpR5</i>	2	6(1); 4.2.2.1(1)	2	2	0	100	-	-
Y92C	<i>mmpR5</i>	2	4.3.3(1); 4.4.2(1)	2	2	0	100	100	P,M
F93L	<i>mmpR5</i>	2	4.4.1.1(1); 4.2.2(1)	2	2	50.0	50.0	100	S,P
Y229C	<i>pepQ</i>	2	4.8(2)	1	1	100	0	-	S,P
R105G	<i>mmpR5</i>	2	1.2.1(1); 2.2.1(1)	2	2	0	50.0	100	-
-41T>C	<i>atpE</i>	2	1.1.2(1); Bov(1)	2	2	50.0	0	0	-
R156*	<i>mmpR5</i>	2	5(1); .4.2(1)	2	2	0	100	100	-
R50Q	<i>mmpR5</i>	2	2.2.2(2)	1	1	0	100	0	-
E54A	<i>mmpR5</i>	2	1.1.1(1); 4.1.2.1(1)	2	2	50.0	50.0	100	-
<u>V214F</u>	<i>pepQ</i>	2	4.2.1(2)	1	1	100	0	100	-
R109L	<i>mmpR5</i>	2	3(2)	1	1	100	0	100	-
V101A	<i>pepQ</i>	2	4.4.2(2)	1	1	0	0	100	-
R30S	<i>mmpR5</i>	2	4.5(2)	1	1	100	0	100	M
Q22E	<i>mmpR5</i>	2	2.2.1(2)	1	1	0	100	100	-
L44P	<i>mmpR5</i>	2	2.2.1(2)	1	1	0	100	100	S,P
-7G>GA	<i>mmpR5</i>	2	2.2.1.1(2)	1	1	0	100	100	-
136_137insG	<i>mmpR5</i>	2	2.2.1(2)	1	1	0	0	100	-
V101L	<i>pepQ</i>	2	4.5(2)	1	1	100	0	100	-
R206Q	<i>pepQ</i>	2	1.1.1(2)	1	1	100	0	-	-
P366T	<i>pepQ</i>	2	1.1.1(2)	1	1	100	0	-	-
P359L	<i>pepQ</i>	2	1.1.1(1); 6(1)	2	2	50.0	50.0	100	-
A12T	<i>pepQ</i>	2	1.2.2(2)	1	1	100	0	-	-
<u>A90V</u>	<i>pepQ</i>	2	4.1.1.1(2)	1	1	100	0	-	-
R96G	<i>mmpR5</i>	2	4.2.1(2)	1	1	100	0	0	S
F27V	<i>mmpR5</i>	2	4.3.3(2)	1	1	100	0	-	M
V85A	<i>mmpR5</i>	2	1.2.2(2)	1	1	0	0	0	-
D165N	<i>mmpR5</i>	2	3(2)	1	1	0	50.0	-	-
A153P	<i>mmpR5</i>	2	2.2.1(2)	1	1	0	100	-	S
M17V	<i>mmpR5</i>	2	4.9(2)	1	1	0	100	-	-

G116V	<i>pepQ</i>	2	4.1.1.3(2)	1	1	0	100	-	-
128_137del	<i>mmpR5</i>	2	4.1.1.3(1); 4.1.2.1(1)	2	2	0	100	-	-
A124V	<i>pepQ</i>	2	4.1.1.3(1); 1.1.1(1)	2	2	50.0	50.0	100	-
L117P	<i>mmpR5</i>	2	2.2.1(2)	1	1	0	100	100	S,P
V1A	<i>mmpR5</i>	2	2.2.1(1); 4.2.2(1)	2	2	50.0	50.0	100	-
K241T	<i>pepQ</i>	2	4.1.2.1(2)	1	1	100	0	100	-
465_466insC	<i>mmpR5</i>	2	4.1.2.1(2)	1	1	0	100	-	-
274_275insA	<i>mmpR5</i>	2	4.3.4.2.1(1); 2.2.1(1)	2	2	50.0	50.0	100	-
778866_779 429del	<i>mmpR5</i>	2	4.3.4.2.1(2)	1	1	100	0	100	-
G58S	<i>pepQ</i>	2	4.3.4.2.1(2)	1	1	100	0	100	P
I220L	<i>pepQ</i>	2	4.3.4.2(2)	1	1	0	100	0	-
D151G	<i>pepQ</i>	2	4.4(2)	1	1	0	100	100	P
Q22R	<i>mmpR5</i>	2	4.4.2(2)	1	1	0	0	100	-

Sub-lineages: † = more than 1 sub-lineage; # = number; Drug resistance (%): Susc. = Susceptible; * % of number of samples pre-2014/total number of samples with available collection date; ** Functional support: S = snap2 score \geq 50; P = Provean Score \leq -4; M = mCSM predicted stability change ($\Delta\Delta G$) below -2; B = Predicted as resistant by SUSPECT-BDQ (only available for *atpE*). Mutations associated with increased minimum inhibitory concentration (MIC) for bedaquiline (BDQ) in previous studies in bold; mutations associated with susceptibility to BDQ underlined (see **S3 Table**).

S6 Table. All mutations (seen >1 samples) in Delamanid (DLM) /Pretomanid (PTM) candidate genes found in the 33k isolates.

Mutation	Gene	Freq	Sub-lineage(# isolates)	# sub-lin.	# Ind ep Occur.	Susc. %	MD R/X DR %	Pre-2014 % *	Functional Support **
<u>K270M</u>	<i>fgd1</i>	3136	4.1.2*(3135); 2.2.1(1)	3	2	70.1	18.1	84.2	-
-32A>G	<i>fbiC</i>	639	5, 6, <i>Bov</i> (634); 2.2.1(2); 4.3.3(1); 4.2(1); 4.9(1)	7	5	60.1	8.3	63.1	-
<u>T273A</u>	<i>fbiC</i>	626	4.8(625); 1.1.1(1)	2	2	97.9	0.3	93.6	-
R64S	<i>fgd1</i>	471	1.1.1*(471)	2	1	77.9	2.1	99.1	-
<u>T302M</u>	<i>fbiA</i>	355	4.1.1.1(355)	1	1	82.8	9.9	84.8	-
<u>K448R</u>	<i>fbiB</i>	293	3(293)	1	3	57.7	30	51.1	-
<u>D113N</u>	<i>ddn</i>	267	5(264); 2.2.1(3)	2	2	70.7	15.4	91.7	-
G264R	<i>fbiA</i>	261	2.2.1(261)	1	1	91.9	6.1	100	P
E224G	<i>fbiC</i>	210	4.1.1.3(210)	1	1	74.8	10	80	S
<u>K296E</u>	<i>fgd1</i>	162	6(161); 4.1.2.1(1)	2	2	87	3.7	85.7	-
<u>L447R</u>	<i>fbiB</i>	148	4.8(148)	1	1	73.6	23	95.9	-
A505T	<i>fbiC</i>	135	2.1(135)	1	1	61.5	20	95.1	-
I208V	<i>fbiA</i>	122	4.1.2(121); 4.1.2.1(1)	2	2	70.5	11.5	96.9	-
<u>W678G</u>	<i>fbiC</i>	96	4.3.3(88); 1.1.1(8)	2	2	8.3	81.2	90.9	P
D90N	<i>fbiD</i>	80	4.9(80)	1	1	87.5	6.25	100	-
I128V	<i>fbiC</i>	79	2.2.1(79)	1	2	0	81	100	-
<u>M93T</u>	<i>fgd1</i>	76	1.2.2(76)	1	1	85.5	9.2	100	-
<u>R72W</u>	<i>ddn</i>	75	1.1.2(75)	1	2	76	10.7	70.2	S,P
A31T	<i>fbiB</i>	71	2.2.1(70); 2.2.2(1)	1	3	54.9	9.9	100	-
<u>G34R</u>	<i>ddn</i>	47	4.3.2(44); 4.3.4.2(3)	2	2	89.3	8.5	0	S,P
<u>D315A</u>	<i>fbiB</i>	40	<i>Bov</i> (40)	1	1	0	5	0	-
V17A	<i>fbiB</i>	39	<i>Bov</i> (39)	1	1	100	0	0	-
<u>R187H</u>	<i>fgd1</i>	39	4.1.1.1(39)	1	1	100	0	100	-
L323F	<i>fgd1</i>	38	<i>Bov</i> (38)	1	1	100	0	0	-
-11G>A	<i>fbiC</i>	37	4.1.2.1(31); 4.1.1.3(3); 6(2); 4.4.2(1)	4	4	56.8	16.2	100	-
-14G>GA	<i>fbiC</i>	34	2.2.1(25); 4.3.4.2.1(9)	2	2	26.5	73.5	94.7	-
Y163C	<i>fgd1</i>	32	4(32)	1	1	81.3	15.6	28.6	P
<u>E83D</u>	<i>ddn</i>	24	4.2.1(24)	1	1	33.3	45.9	100	
<u>G81S</u>	<i>ddn</i>	21	2.2.2(12); 2.1(9)	2	2	33.3	52.4	100	S,P

P607L	<i>fbiC</i>	21	4.4.1.1(21)	1	1	100	0	100	P
L49P	<i>ddn</i>	21	2.2.1.1(21)	1	3	57.1	9.5	94.4	S,P
<u>A111V</u>	<i>ddn</i>	20	4.4.2(20)	1	1	90	10	100	S,P
G199R	<i>fgd1</i>	20	4.4.2(20)	1	1	90	10	100	P
-27T>G	<i>fgd1</i>	20	4.6(20)	1	1	0	40	0	-
256_261del	<i>ddn</i>	19	4.8(19)	1	1	100	0	0	-
G139R	<i>fbiA</i>	18	2.2.1(17); 1.1.2(1)	2	2	94.4	0	75	P
E278D	<i>fgd1</i>	18	4.3.3(18)	1	1	66.7	27.8	66.7	-
W20*	<i>ddn</i>	17	4.5(11); 5(6)	2	2	100	0	75	-
Q121H	<i>fbiD</i>	17	4.5(17)	1	1	76.5	0	100	-
K183M	<i>fgd1</i>	16	2.2.1(16)	1	1	43.8	50	100	-
K296R	<i>fgd1</i>	16	4.1.2.1(12); 4.8(3); 4.4.1.1(1)	3	3	18.8	31.3	37.5	-
A11V	<i>fbiC</i>	15	1.1.2(15)	1	1	66.7	13.3	53.8	-
-77_-9del	<i>fgd1</i>	15	1.2.1(15)	1	1	40	20	100	-
<u>R23W</u>	<i>ddn</i>	15	4.3.2(15)	1	1	33.3	46.7	0	S,P
G145R	<i>fbiD</i>	15	4.1.1.1(15)	1	1	86.7	0	80	S,P
<u>D90N</u>	<i>fbiB</i>	14	3(14)	1	2	50	14.3	14.3	-
G839A	<i>fbiC</i>	14	4.5(14)	1	1	85.7	0	85.7	-
S56C	<i>fgd1</i>	14	4.3.4.2(14)	1	1	0	100	0	-
R265Q	<i>fbiB</i>	13	2.2.1(12); 1.1.2(1)	2	3	30.8	0	100	-
<u>A524G</u>	<i>fbiC</i>	13	4(13)	1	1	53.8	7.7	100	-
<u>V581L</u>	<i>fbiC</i>	12	2.2.1(12)	1	1	58.3	16.6	100	-
R14G	<i>fbiA</i>	12	4.8(12)	1	1	100	0	100	S,P
V170M	<i>fgd1</i>	12	2.2.1(12)	1	1	0	100	33.3	-
P18L	<i>fbiC</i>	11	4.8(11)	1	1	72.7	0	100	-
<u>P6S</u>	<i>ddn</i>	11	1.1.1(11)	1	1	100	0	100	-
T255A	<i>fgd1</i>	10	3(10)	1	1	60	10	0	-
<u>R409S</u>	<i>fbiB</i>	10	3(10)	1	1	80	0	100	P
-26G>T	<i>ddn</i>	10	3(10)	1	1	70	20	100	-
V416I	<i>fbiB</i>	10	1.1.3(10)	1	1	100	0	100	-
<u>P6T</u>	<i>ddn</i>	10	3(10)	1	1	80	0	0	P
A178T	<i>fbiA</i>	10	1.2.1(8); 4.5(1); 3(1)	2	4	70	0	66.7	-
A84G	<i>fgd1</i>	10	2.2.1(10)	1	1	0	100	100	-
-43G>A	<i>ddn</i>	9	5(4); 4.2.1(3); 2.2.1(2)	3	3	30.8	0	100	-
A199T	<i>fbiA</i>	9	2.2.2(9)	1	1	0	88.9	100	-
<u>R230Q</u>	<i>fbiB</i>	9	1.1.3(9)	1	1	100	0	-	-
2546_2547insCACAT ACGCCCTGCTTGCG	<i>fbiC</i>	9	4.6(9)	1	1	77.8	0	40	-
W589R	<i>fbiC</i>	9	4.3.4.2(9)	1	1	100	0	-	-
<u>R30S</u>	<i>ddn</i>	9	2.2.1(9)	1	1	11.1	55.6	100	S,P
P131L	<i>ddn</i>	9	4.8(8); 4.3.4.2.1(1)	2	2	88.9	0	100	S,P

<u>T681I</u>	<i>fbiC</i>	9	2.2.1(9)	1	1	77.8	11.1	100	P
363_386del	<i>ddn</i>	9	4.5(9)	1	1	77.8	11.1	-	-
A143V	<i>fbiC</i>	9	4.3.4.2.1(9)	1	1	100	0	100	-
G655S	<i>fbiC</i>	9	2.2.1(8); 4.1.2(1)	2	2	33.3	0	100	-
I13L	<i>fbiC</i>	9	Bov(9)	1	1	0	11.1	0	-
A197V	<i>fbiD</i>	9	1.1.3(9)	1	1	100	0	-	-
G572C	<i>fbiC</i>	8	2.2.1(8)	1	1	0	75	100	-
R247W	<i>fgd1</i>	8	4.5(1); 3(7)	2	2	100	0	50	P
V348I	<i>fbiB</i>	8	4.1(1); 2.2.2(7)	2	2	100	0	100	-
R304Q	<i>fbiA</i>	8	3(8)	1	2	87.5	0	50	-
<u>-24C>A</u>	<i>ddn</i>	7	4.1.1.2(7)	1	1	85.7	0	100	-
T687M	<i>fbiC</i>	7	1.1.2(7)	1	1	42.9	42.9	-	-
A349V	<i>fbiC</i>	7	1.1.1(7)	1	1	85.7	0	100	S
A345G	<i>fbiC</i>	7	4.3.4.2.1(7)	1	1	85.7	0	100	-
A2V	<i>fgd1</i>	7	4.2.2(7)	1	1	100	0	0	-
S762N	<i>fbiC</i>	7	3(7)	1	1	42.8	28.6	0	-
<u>D312G</u>	<i>fbiA</i>	7	4.8(7)	1	1	71.4	14.3	100	P
-13A>G	<i>fbiC</i>	7	2.2.1(7)	1	1	85.7	14.3	100	-
V188F	<i>fbiA</i>	7	1.2.1(7)	1	1	28.6	14.3	33.3	P
G325S	<i>fbiB</i>	7	4.9(1); 4.1.2.1(1); 2.2.1(5)	3	3	100	0	83.3	-
P420L	<i>fbiC</i>	7	2.2.1(7)	1	1	14.3	85.7	83.3	P
W88*	<i>ddn</i>	6	2.2.1(6)	1	1	11.1	88.9	66.7	-
G71D	<i>fgd1</i>	6	3(6)	1	1	66.7	0	0	S,P
P182L	<i>fbiB</i>	6	4.3.4.2.1(3); 6(3)	2	2	66.7	16.7	100	-
M93I	<i>fgd1</i>	6	4.9(3); 4.1.2.1(2); 2.2.1(1)	3	3	83.3	0	-	-
-41G>T	<i>fbiC</i>	6	1.2.2(6)	1	1	83.3	0	50	-
<u>I693V</u>	<i>fbiC</i>	6	3(6)	1	1	83.3	16.7	-	-
<u>Q120R</u>	<i>fbiA</i>	6	4.8(6)	1	1	33.3	33.3	0	-
<u>L67P</u>	<i>ddn</i>	6	4.8(6)	1	1	83.3	0	-	S,P
Y167H	<i>fbiC</i>	6	3(6)	1	1	100	0	0	P
K279E	<i>fbiC</i>	6	4.1.2.1(6)	1	1	100	0	0	-
T695K	<i>fbiC</i>	6	4.1.2.1(6)	1	1	0	100	-	-
D224N	<i>fbiB</i>	6	2.2.1(6)	1	1	33.3	16.7	100	-
<u>P45L</u>	<i>ddn</i>	5	4.4.1.1(3); 3(1); 1.1.1(1)	3	3	80	0	100	S,P
G508S	<i>fbiC</i>	5	1.2.2(5)	1	1	20	80	100	-
-23C>T	<i>ddn</i>	5	4.8(5)	1	1	100	0	100	-
P607A	<i>fbiC</i>	5	4.7(5)	1	1	100	0	-	P
-6A>C	<i>ddn</i>	5	4.8(5)	1	1	40	60	100	-
L326F	<i>fbiB</i>	5	4.6.1.1(2); 4.1.2.1(1); 4.1.2(1); 4.8(1)	4	4	100	0	-	-
A206T	<i>fbiA</i>	5	2.2.1(5)	1	1	80	0	66.7	-

V188I	<i>fbiA</i>	5	2.2.1(5)	1	1	0	80	66.7	-
R409C	<i>fbiB</i>	5	1.2.1(5)	1	1	60	0	100	P
T455A	<i>fbiC</i>	5	3(4); 1.1.1(1)	2	2	80	0	100	-
A132T	<i>fbiC</i>	5	2.2.1(5)	1	1	40	20	-	-
A835V	<i>fbiC</i>	5	1.1.3(5)	1	1	100	0	0	-
K183T	<i>fgd1</i>	5	4.3.3(5)	1	1	80	0	0	-
T302A	<i>fbiA</i>	5	2.1(5)	1	1	40	20	80	-
A201P	<i>fbiA</i>	5	2.1(5)	1	1	80	0	100	-
85_87del	<i>ddn</i>	4	2.2.1(4)	1	1	50	0	50	-
<u>T302P</u>	<i>fbiA</i>	4	2.2.1(4)	1	1	0	100	100	-
508_509insT	<i>fgd1</i>	4	4.1.1.3(4)	1	1	100	0	100	-
A380S	<i>fbiB</i>	4	3(4)	1	1	75	25	-	S,M
Q69R	<i>fbiB</i>	4	3(1); 4.2.2(3)	2	2	100	0	0	-
V61I	<i>fgd1</i>	4	1.1.3(4)	1	1	100	0	100	-
W139*	<i>ddn</i>	4	4.1.2(4)	1	1	100	0	-	-
P438S	<i>fbiC</i>	4	4.3.4.2(4)	1	1	100	0	100	P
D126Y	<i>fbiC</i>	4	4.8(4)	1	1	75	0	100	S,P
G168R	<i>fgd1</i>	4	4.1.1.1(3); 2.2.2(1)	2	2	50	50	0	-
D168E	<i>fbiC</i>	4	4.1.2(4)	1	1	0	25	100	-
<u>R72Q</u>	<i>ddn</i>	4	4.8(4)	1	1	100	0	100	-
R154H	<i>fbiC</i>	4	4.6.1.2(4)	1	1	0	100	100	S,P
I18V	<i>fbiB</i>	4	4.6.1.2(4)	1	1	25	25	-	-
R177H	<i>fbiA</i>	4	4.1.2.1(2); 4.5(2)	2	2	100	0	100	-
P438L	<i>fbiC</i>	4	4.4.1.1(4)	1	1	100	0	0	P
R45C	<i>fgd1</i>	4	4.3.2(4)	1	1	0	75	100	P
V123I	<i>fbiB</i>	4	4.6.2.2(4)	1	1	0	100	100	-
283_303del	<i>ddn</i>	4	4.5(4)	1	1	100	0	100	-
3986845_3987298de I	<i>ddn</i>	4	2.2.1(4)	1	1	0	100	100	-
T218A	<i>fbiC</i>	4	2.2.1.1(4)	1	1	0	100	100	-
R293W	<i>fbiB</i>	4	1.2.1(4)	1	1	75	0	100	S,P
Q170H	<i>fbiA</i>	4	1.1.1(3); 1.2.2(1)	2	2	100	0	-	P
G839D	<i>fbiC</i>	4	1.1.1(4)	1	1	100	0	100	-
D263N	<i>fgd1</i>	4	1.1.3(4)	1	1	50	50	0	-
V301L	<i>fbiA</i>	4	4.4.1.1(4)	1	1	100	0	100	-
D387N	<i>fbiC</i>	4	2.2.1(4)	1	1	75	0	100	-
R334Q	<i>fbiB</i>	4	1.1.1.1(4)	1	1	75	0	100	S
D78N	<i>fbiD</i>	4	4.5(4)	1	1	100	0	100	-
-37T>C	<i>fbiC</i>	3	2.2.1.1(3)	1	1	100	0	100	-
G145A	<i>fgd1</i>	3	5(3)	1	1	33.3	33.3	-	-
-38G>A	<i>fgd1</i>	3	3(3)	1	1	100	0	-	-
-3C>T	<i>fgd1</i>	3	4.1.2.1(3)	1	1	66.7	0	-	-
A111T	<i>ddn</i>	3	1.2.2(1); 4.5(1); 3(1)	3	3	66.7	0	-	S,P

A333V	<i>fbiC</i>	3	3.1.2(3)	1	1	0	66.7	-	-
N556D	<i>fbiC</i>	3	1.2.1(2); 3(1)	2	2	66.7	0	100	S,P
G310*	<i>fbiC</i>	3	1.2.1(3)	1	1	100	0	-	-
G70V	<i>ddn</i>	3	1.2.2(3)	1	1	100	0	-	S,P
D542N	<i>fbiC</i>	3	4.1.2(3)	1	1	100	0	100	P
V621I	<i>fbiC</i>	3	4.8(2); 2.2.1(1)	2	2	66.7	33.3	100	-
V61G	<i>ddn</i>	3	4.1.2(3)	1	1	66.7	0	-	S
-40C>A	<i>ddn</i>	3	3(3)	1	1	0	0	-	-
V241I	<i>fbiA</i>	3	5(3)	1	1	100	0	100	-
-33G>A	<i>fbiC</i>	3	1.1.1(3)	1	1	100	0	-	-
V46G	<i>ddn</i>	3	5(3)	1	1	0	100	-	S,P
V740A	<i>fbiC</i>	3	4.1.2.1(3)	1	1	100	0	0	S
-31T>C	<i>fbiC</i>	3	4.1.2.1(3)	1	1	100	0	0	-
A237V	<i>fbiB</i>	3	1.2.2(3)	1	1	100	0	0	-
G26S	<i>fbiB</i>	3	4.1.2.1(3)	1	1	100	0	0	P
D406A	<i>fbiB</i>	3	4.1.2.1(3)	1	1	100	0	0	P
E474A	<i>fbiC</i>	3	<i>Bov</i> (3)	1	1	100	0	0	-
R137H	<i>fbiB</i>	3	4(3)	1	1	100	0	0	-
V389L	<i>fbiC</i>	3	4.8(3)	1	1	100	0	0	-
-10G>C	<i>fbiC</i>	3	4.3.2(3)	1	1	100	0	0	-
K236N	<i>fbiC</i>	3	4.8(3)	1	1	100	0	0	S,P
A77T	<i>ddn</i>	3	4.8(3)	1	1	100	0	0	S,P
R330P	<i>fbiC</i>	3	1.2.1(3)	1	1	100	0	0	P
-10G>A	<i>fbiC</i>	3	4.3.4.2(3)	1	1	0	100	-	-
A10V	<i>fgd1</i>	3	1.1.2(3)	1	1	100	0	-	-
A206S	<i>fbiA</i>	3	2.2.2(3)	1	1	100	0	66.7	-
L723F	<i>fbiC</i>	3	2.2.1(1); 4.3.4.2.1(2)	2	1	66.7	33.3	66.7	-
S42G	<i>fbiA</i>	3	2.2.2(3)	1	1	100	0	-	-
A404V	<i>fbiC</i>	3	2.2.1(3)	1	1	66.7	0	-	-
A620T	<i>fbiC</i>	3	2.2.1(3)	1	1	66.7	0	100	-
P15S	<i>fbiC</i>	3	4.5(3)	1	1	0	0	100	-
527_534del	<i>fgd1</i>	3	2.2.1(2); 4.3.2.1(1)	2	3	33.3	33.3	100	-
G74C	<i>fbiC</i>	3	1.2.2(3)	1	1	100	0	100	P
S78Y	<i>ddn</i>	3	2.2.1(3)	1	1	100	0	-	S,P
V37G	<i>fgd1</i>	3	<i>Bov</i> (1); 4.1.2(2)	2	2	0	0	100	S,P
A29T	<i>fgd1</i>	3	1.1.1(3)	1	1	0	0	100	-
-48C>T	<i>fbiC</i>	3	3.1.2(3)	1	1	33.3	0	50	-
-40C>T	<i>ddn</i>	3	3.1.2(3)	1	1	100	0	-	-
L228F	<i>fbiC</i>	3	3(3)	1	1	100	0	0	-
K2E	<i>fbiA</i>	3	3(2); 1.1.2(1)	2	2	66.7	0	-	-
W88R	<i>ddn</i>	3	3(1); 4.1.1.3(2)	2	2	33.3	66.7	100	S,P, M
V625A	<i>fbiC</i>	3	1.2.2(3)	1	1	0	66.7	-	-

V155M	<i>fbiA</i>	3	3(1); 1.1.1.1(2)	2	2	100	0	100	-
Q27P	<i>fbiA</i>	3	4.1.2.1(3)	1	1	100	0	-	-
P206L	<i>fbiC</i>	3	4.8(1); 2.2.1(2)	2	2	66.7	0	100	P
E312K	<i>fbiC</i>	3	1.1.3(3)	1	1	0	100	-	-
<u>A43T</u>	<i>fbiA</i>	3	4.4.1.1(2); 1.1.2(1)	2	2	66.7	0	100	-
<u>M319I</u>	<i>fbiA</i>	3	2.2.1(3)	1	1	0	100	100	-
P63S	<i>ddn</i>	3	1.1.2(3)	1	1	100	0	100	S,P
I246T	<i>fbiA</i>	3	4.2.2(3)	1	1	100	0	100	P
I193V	<i>fgd1</i>	3	1.1.2(3)	1	1	66.7	0	100	-
I208M	<i>fbiA</i>	3	1.1.2(3)	1	1	33.3	33.3	100	-
Y65S	<i>ddn</i>	3	4.5(3)	1	1	0	0	100	S
P111L	<i>fbiC</i>	3	2.2.1(3)	1	1	66.7	0	100	P
E65G	<i>fbiB</i>	3	4.8(3)	1	1	100	0	100	-
G8D	<i>fbiD</i>	3	2.2.1(3)	1	1	33.3	66.7	50	-
I10V	<i>fbiD</i>	3	4.2.2(3)	1	1	0	0	100	-
A20V	<i>fbiD</i>	3	2.2.1(3)	1	1	100	0	100	-
T34S	<i>fbiD</i>	3	3.1.1(3)	1	1	100	0	0	-
G76S	<i>fbiD</i>	3	4.2(3)	1	1	100	0	100	P
E127Q	<i>fbiD</i>	3	4.3.4.1(3)	1	1	66.7	33.3	-	-
G155S	<i>fbiD</i>	3	1.2.2(3)	1	1	66.7	0	0	-
V211G	<i>fbiD</i>	3	2.2.1(3)	1	1	100	0	100	-
-45G>C	<i>fbiD</i>	3	4.5(3)	1	1	0	100	-	-
-34G>C	<i>fbiD</i>	3	3.1.2(3)	1	1	100	0	0	-
T302I	<i>fbiB</i>	2	1.2.1(1); 2.2.1(1)	2	2	0	50	100	-
V154I	<i>fbiA</i>	2	4.2.1(2)	1	1	100	0	100	-
E282D	<i>fbiB</i>	2	2.2.1(2)	1	1	100	0	-	-
-17T>TC	<i>ddn</i>	2	4.1.1.3(2)	1	1	100	0	100	-
381_464del	<i>fbiA</i>	2	4.1.2.1(1); 1.2.2(1)	2	2	100	0	-	-
K282N	<i>fbiC</i>	2	1.1.2(2)	1	1	100	0	-	-
V16F	<i>fbiC</i>	2	1.2.2(2)	1	1	0	0	100	-
H364Y	<i>fbiC</i>	2	3(1); 4.1.2.1(1)	2	2	50	0	-	P
G755S	<i>fbiC</i>	2	3(1); 4.1.1.3(1)	2	2	50	0	-	P
I262V	<i>fgd1</i>	2	4.1.1.3(2)	1	1	100	0	-	M
E608K	<i>fbiC</i>	2	3(2)	1	1	0	0	-	-
R780C	<i>fbiC</i>	2	4.8(2)	1	1	100	0	-	P
V25A	<i>fgd1</i>	2	4.9(2)	1	1	0	0	-	M
G277S	<i>fbiA</i>	2	3(2)	1	1	100	0	-	P
<u>V41M</u>	<i>fbiC</i>	2	1.1.2(2)	1	1	100	0	-	-
G293A	<i>fbiA</i>	2	4.1.2.1(2)	1	1	100	0	-	-
E332K	<i>fbiB</i>	2	4.4.1.2(2)	1	1	50	0	-	-
A10T	<i>fgd1</i>	2	3.1.2(2)	1	1	100	0	-	M
273_273del	<i>ddn</i>	2	3(2)	1	1	100	0	-	-
P78S	<i>fbiC</i>	2	4.6.2(1); 2.1(1)	2	2	50	50	100	-

R99W	<i>fbiC</i>	2	4.1.2.1(1); 4.3.4.1(1)	2	2	100	0	-	S,P
D308G	<i>fbiC</i>	2	3(2)	1	1	100	0	-	P
R458H	<i>fbiC</i>	2	1.1.3(1); 6(1)	2	2	100	0	-	P
P370R	<i>fbiC</i>	2	3(1); 4.4(1)	2	2	100	0	-	P
R365G	<i>fbiB</i>	2	3(2)	1	1	100	0	100	P,M
I247N	<i>fbiA</i>	2	4.7(2)	1	1	100	0	-	P
G94R	<i>fgd1</i>	2	4.1.2.1(2)	1	1	100	0	-	P
E127D	<i>fbiC</i>	2	4.1.2.1(2)	1	1	0	0	-	-
M268V	<i>fgd1</i>	2	4.7(2)	1	1	100	0	-	-
<u>S184T</u>	<i>fbiA</i>	2	3(2)	1	1	50	50	-	-
P361A	<i>fbiB</i>	2	4(2)	1	1	100	0	-	P
G541S	<i>fbiC</i>	2	4.1.2.1(1); 1.2.2(1)	2	2	50	50	-	P
L93F	<i>fbiB</i>	2	4.3.2.1(1); 4.8(1)	2	2	50	50	-	-
A404P	<i>fbiC</i>	2	4.3.3(2)	1	1	100	0	-	-
G78S	<i>fbiA</i>	2	4.3.4.2(1); 4.3.4.2.1(1)	2	2	50	0	100	S,P
<u>D66E</u>	<i>fbiB</i>	2	4.3.4.2.1(2)	1	1	100	0	100	-
<u>D465A</u>	<i>fbiC</i>	2	4.6(2)	1	1	100	0	-	P
G159V	<i>fgd1</i>	2	4.1.2.1(2)	1	1	100	0	100	-
H295R	<i>fbiA</i>	2	4.1.2.1(2)	1	1	0	50	100	-
P193S	<i>fbiC</i>	2	4.1.2.1(2)	1	1	100	0	-	P
R845C	<i>fbiC</i>	2	6(1); 1.1.1(1)	2	2	100	0	100	P
T185A	<i>fbiC</i>	2	6(2)	1	1	50	0	100	P
I816V	<i>fbiC</i>	2	5(2)	1	1	100	0	-	-
G445D	<i>fbiC</i>	2	3(2)	1	1	50	50	100	P
T292A	<i>fbiB</i>	2	4.3.3(2)	1	1	0	100	100	P
<u>P16R</u>	<i>fbiB</i>	2	2.2.1(2)	1	1	100	0	100	S,P
A136S	<i>fbiC</i>	2	4.8(1); 4.1.2(1)	2	2	50	50	100	-
<u>D69N</u>	<i>ddn</i>	2	4.3.2.1(1); 4.3.4.2(1)	2	2	0	100	100	-
G264E	<i>fbiA</i>	2	1.2.1(1); 4.4.2(1)	2	2	50	0	50	S,P
<u>F220L</u>	<i>fbiB</i>	2	4.1.2.1(1); 4.8(1)	2	2	100	0	0	-
T268I	<i>fbiB</i>	2	2.2.1(1); 3(1)	2	2	50	50	100	-
R321S	<i>fbiA</i>	2	3(2)	1	1	0	100	100	-
G512C	<i>fbiC</i>	2	4.3.3(2)	1	1	100	0	100	-
D147N	<i>fbiA</i>	2	4.6.1.1(2)	1	1	0	0	-	P
N66Y	<i>fgd1</i>	2	2.2.1(2)	1	1	50	50	100	-
R486H	<i>fbiC</i>	2	4.5(1); 4(1)	2	2	50	50	-	-
R68H	<i>ddn</i>	2	2.2.1(2)	1	1	100	0	0	-
R550C	<i>fbiC</i>	2	2.2.1(2)	1	1	100	0	100	S,P
-46GGTGGGGC>G	<i>fbiC</i>	2	2.2.2(2)	1	1	100	0	-	-
-9T>C	<i>fbiC</i>	2	1.1.1(2)	1	1	100	0	-	-
V390G	<i>fbiB</i>	2	2.2.2(1); 3(1)	2	2	50	0	-	-

-18T>C	<i>fgd1</i>	2	4.8(2)	1	1	100	0	-	-
N32T	<i>ddn</i>	2	4.2.1(2)	1	2	0	100	100	-
*152G	<i>ddn</i>	2	4.5(2)	1	1	100	0	100	-
-22C>T	<i>fbiA</i>	2	4.5(2)	1	1	100	0	100	-
F306V	<i>fbiC</i>	2	4.4.2(2)	1	1	100	0	100	-
V33I	<i>fbiB</i>	2	2.2.1(2)	1	1	100	0	100	-
L374S	<i>fbiC</i>	2	2.2.1(2)	1	1	0	100	100	S,P
-17A>C	<i>fbiC</i>	2	4.4.2(2)	1	1	0	100	100	-
P362S	<i>fbiC</i>	2	2.2.1(2)	1	1	0	100	100	P
E13G	<i>fbiB</i>	2	2.2.1.1(2)	1	1	0	100	100	-
I167V	<i>fbiA</i>	2	2.2.1(2)	1	1	0	100	100	-
H183N	<i>fbiA</i>	2	2.2.1(2)	1	1	0	100	100	P
T796A	<i>fbiC</i>	2	2.2.2(2)	1	1	0	100	100	-
R212Q	<i>fgd1</i>	2	1.1.1.1(1); 7(1)	2	2	100	0	100	-
<u>D74E</u>	<i>fbiA</i>	2	1.2.2(2)	1	1	100	0	-	-
G189D	<i>fbiA</i>	2	2.2.1(2)	1	2	100	0	-	P
M313L	<i>fbiB</i>	2	2.2.1(2)	1	1	100	0	0	-
<u>P6L</u>	<i>ddn</i>	2	2.2.1(2)	1	1	0	50	-	P
L204F	<i>fbiC</i>	2	2.2.1(2)	1	1	0	50	-	-
D203N	<i>fbiB</i>	2	1.1.1(2)	1	1	100	0	-	P
D148N	<i>fbiA</i>	2	1.1.1(2)	1	1	50	0	-	P
A178G	<i>fbiA</i>	2	2.2.1(2)	1	1	100	0	-	-
P60S	<i>fbiA</i>	2	3(2)	1	1	0	100	0	P
E205K	<i>fgd1</i>	2	1.2.2(2)	1	1	0	0	-	-
A63T	<i>fbiC</i>	2	1.1.2(1); 2.2.1(1)	2	1	50	50	100	-
S132C	<i>ddn</i>	2	1.2.1(2)	1	1	50	0	-	-
<u>V147M</u>	<i>ddn</i>	2	3(2)	1	1	100	0	-	-
A856T	<i>fbiC</i>	2	4.8(2)	1	1	100	0	-	-
V581I	<i>fbiC</i>	2	4.1.2.1(1); 2.2.1(1)	2	2	50	50	-	-
H46D	<i>fgd1</i>	2	3(1); 4.8(1)	2	2	50	0	-	P
E83A	<i>ddn</i>	2	4.2.1(2)	1	1	0	0	0	-
-29C>G	<i>fgd1</i>	2	1.1.2(1); 3(1)	2	2	50	0	-	-
A659V	<i>fbiC</i>	2	3(2)	1	2	50	50	-	-
<u>T50I</u>	<i>ddn</i>	2	4.8(2)	1	2	100	0	-	S,P
I638L	<i>fbiC</i>	2	<i>Bov</i> (2)	1	1	100	0	-	-
G839S	<i>fbiC</i>	2	3(1); <i>Bov</i> (1)	2	2	50	0	-	-
A328V	<i>fbiB</i>	2	<i>Bov</i> (2)	1	1	0	0	-	-
T36P	<i>fgd1</i>	2	2.2.1(2)	1	1	100	0	100	-
M709I	<i>fbiC</i>	2	2.2.1(2)	1	1	0	100	-	S
851_939del	<i>fgd1</i>	2	4.7(1); 2.2.1(1)	2	2	0	50	-	-
<u>E105Q</u>	<i>ddn</i>	2	1.1.2(1); 4.1.2.1(1)	2	2	50	50	50	-
V599A	<i>fbiC</i>	2	4.2.2(2)	1	1	50	0	50	-
R134L	<i>fbiC</i>	2	3(1); 2.1(1)	2	2	100	0	100	-

A212P	<i>fbiA</i>	2	4.6(1); 4.3.3(1)	2	2	0	100	-	P
R139Q	<i>fbiB</i>	2	4.3.2.1(2)	1	1	100	0	100	-
V495A	<i>fbiC</i>	2	3.1.1(2)	1	1	100	0	100	-
E526K	<i>fbiC</i>	2	4.3.4.2.1(2)	1	1	100	0	100	-
G114S	<i>fbiB</i>	2	4.3.4.2.1(2)	1	1	100	0	100	-
<u>I102V</u>	<i>ddn</i>	2	1.2.2(1); 4.1.2.1(1)	2	2	50	0	100	-
A34T	<i>fbiB</i>	2	1.1.3(2)	1	1	100	0	100	-
S87L	<i>fbiC</i>	2	4.3.3(2)	1	1	100	0	-	S,P
T185I	<i>fbiC</i>	2	4.7(2)	1	1	0	100	100	S,P
E342Q	<i>fbiC</i>	2	1.2.2(2)	1	1	0	50	100	-
A518G	<i>fbiC</i>	2	1.1.2(2)	1	1	100	0	100	-
P272S	<i>fbiB</i>	2	1.1.2(2)	1	1	100	0	100	-
<u>A82T</u>	<i>fbiB</i>	2	4.3.2.1(1); 1.1.1(1)	2	2	100	0	100	-
E299V	<i>fbiC</i>	2	2.2.1(2)	1	1	0	0	100	-
-6G>T	<i>fgd1</i>	2	2.2.1.1(2)	1	1	100	0	100	-
T374K	<i>fbiB</i>	2	2.2.1(2)	1	1	0	50	100	S,P
T371A	<i>fbiB</i>	2	1.1.1.1(2)	1	1	100	0	100	-
V136M	<i>fgd1</i>	2	2.2.1(2)	1	1	0	100	100	-
490706_490745del	<i>fgd1</i>	2	4.7(2)	1	1	0	100	100	-
P270L	<i>fbiB</i>	2	4.1.2.1(2)	1	1	100	0	-	P
I10T	<i>fbiD</i>	2	<i>Bov</i> (2)	1	1	0	0	0	M
V16I	<i>fbiD</i>	2	1.2.1(2)	1	1	50	0	100	-
A21T	<i>fbiD</i>	2	4.1.1.3(2)	1	1	50	0	100	-
A22T	<i>fbiD</i>	2	4.1.1.3(2)	1	1	100	0	-	S
T48I	<i>fbiD</i>	2	4.8(2)	1	1	100	0	-	-
G106E	<i>fbiD</i>	2	<i>Bov</i> (2)	1	1	0	0	0	-
V111I	<i>fbiD</i>	2	1.1.3(2)	1	1	50	50	100	-
T122P	<i>fbiD</i>	2	2.2.2(2)	1	1	0	100	100	-
I129M	<i>fbiD</i>	2	2.2.1(2)	1	1	0	100	0	-
C187F	<i>fbiD</i>	2	4.3.3(2)	1	1	0	100	100	-
-40C>A	<i>fbiD</i>	2	3(2)	1	1	100	0	0	-
-39G>T	<i>fbiD</i>	2	3(2)	1	1	100	0	-	-

Bedaquiline (BDQ), delamanid (DLM); pretomanid (PTM); Sub-lineages: † = more than 1 sub-lineage; # = number; Drug resistance (%): Susc. = Susceptible; * % of number of samples pre-2014/total number of samples with available collection date; ** Functional support: S = snap2 score \geq 50; P = Provean Score \leq -4; M = mCSM predicted stability change ($\Delta\Delta G$) below -2; mutations associated with increased minimum inhibitory concentration for DLM or PTM in previous studies in bold; mutations associated with susceptibility to MIC underlined (see **S3 Table**).

S7 Table. Mutations observed in single isolates in the 33k dataset.

Drug*	Gene	Mutation
BDQ	<i>atpE</i>	A6V, G13S, I16V, <i>M17I</i> (B), A18S, <i>I26V</i> (B), V30I, <i>E44A</i> (B), <i>F50L</i> (B), <i>P52L</i> (B), I66V (B), -40G>GT, -39C>T, -8A>AT, -9GAT>G, -28TACCAGAGCC>T, -32C>T, -33A>G, -39C>G, 225_226insTTCGCTACACCCGTC AAGTAA
BDQ	<i>mmpR5</i>	S2I , V7F, D8G, E13A, E13*, D15A, D15G, E18D, <i>G24S</i> (S), G25D (M), G25S, Y26C, E28A, S29F, S29C, W42* , E49K, <i>Q51P</i> (S), <i>Q51H</i> (S), A61T, S64R, <i>G65W</i> (S,P), G66V, <i>S68I</i> (S,P), R72Q , M73I, L74P (S,P,M), Q76E, <i>G78V</i> (S,P), V85F, A86V, 87GDRRTYFRLRPN>87GGS AHLFPVAAH, D88G , <i>R89W</i> (S,P), <i>R89L</i> (P), R89Q, <i>L95S</i> (M), R96W (S,P), R96Q , P97T, N98K, A99V , A101S, A102T, G103S, E104G, R107G, R109Q, <u>A110V</u> , M111K (S), A112T, Q115P, R134G, L136P (S,P), R137Q, <i>V149G</i> (M), A153V, L154P (S), R160Q, 17_18insGGT, 30_30del , 70_71insGC, 107_108insG, 113_131del, 138_139insGA , 139_140insATC, 212_212del , 216_309del, 234_235insT, 285_285del, 289_289del , 429_429del, 430_431insCA, 431_432insT 462_462del, 479_480insA, 778997_779279del, -3C>A, -10A>C, -22A>C, -30C>G, -31A>G, -33G>T, -46G>A
BDQ	<i>pepQ</i>	Q13E, S25R, I28V, <i>Y32H</i> (S,P), <i>S39F</i> (P), <i>N40S</i> (P), <i>G41R</i> (S), V45L, F46L, A47G, S66P, L71V, E72D, V73M, <u>A78V</u> , V79A, G80R, A84V, G88S, G91D, G93R, <i>F97V</i> (P), H100R, <i>T103M</i> (P), V104M, V104L, G106S, A109V, K117R, N118D, E120D, L121V, T127S, E148G, A152V, V158M, R160C, R160P, R170W, V172M, R174S, A178D, <u>M180V</u> , D182E, <i>E191V</i> (S,P), <i>E191G</i> (S,P), A196T, A201G, <i>R206W</i> (P), R206L, <i>T208I</i> (P), A227D, <i>M233T</i> (P), V238M, D244N, Y250H, <i>R261G</i> (P), A263V, R271Q, A284G, <i>F290V</i> (S,P), <i>F290L</i> (S,P), Q301R, G309E, T315A, S320F, S320C, R333H, A345V, K350Q, E368K, <u>A370T</u> , A370V, <u>L372V</u> , 138_139insTC, 947_948insG, 2859300_2860417del
DLM/ PTM	<i>ddn</i>	M1T , L13R, S14N, K19R, R23Q, T26I, <i>W27G</i> (S), <u>W27C</u> , W27* , R31S, R31C, <u>G34E</u> , <u>G36V</u> , <i>G38R</i> (S,P), K43E, <i>T51A</i> (P), <u>T52N</u> , G53D , <i>G53S</i> (S,P), <i>R54G</i> (S,P), <i>R54C</i> (S,P), Q58K, Q58P, Q58* , <i>P59Q</i> (P), <i>N62K</i> (S), <u>G71R</u> , <i>V75A</i> (S,M), <i>K79E</i> (S), M87I, <u>N91T</u> , <i>N95K</i> (P), K97N, V98F (S,P), V100I, <i>Q101P</i> (S,P), K104R, E105K, <u>E117K</u> , <u>P124S</u> , L126F, M129T, M129I, Y133C , Y133*, <i>Y133H</i> (S,P), Q137R, <u>T140I</u> , -1C>T, -3G>A, -4C>T, -5G>A, -11G>A, -26G>A, -32T>G, -32T>C, -34C>T, -39G>A, -39GC>G, -44C>T, 6_7insAAATC, 24_29del, 36_36del, 59_101del, 90_90del, 92_92del, 164_165del, 211_211del, 255_260del, 267_267del, 270_281del, 285_285del, 309_310insT, 312_312del, 322_323insA, 323_330del, 367_369del, 451_455del, 3986810_3986932del, 3986856_3987298del, 3986857_3987298del,
DLM/ PTM	<i>fgd1</i>	<i>L4R</i> (P), <i>S11P</i> (P), Q14R, A16T, E19D, V21I, A26T, M32V, V37F, Q47L, G62A, N66S, <i>T76I</i> (P), T78I, <i>F79S</i> (S,P,M), V85I, T92S, C95Y, T107A, T107I, A115S, <i>Y118S</i> (M), E119G, F129V, A130T, A130G, R131Q, G137R, Q141H, <i>D146G</i> (P), D146N, D153A, D153E, S161L, <i>I162T</i> (M), V165L, D167E, V172I, <i>A182V</i> , Y184C (P), A188G, E201G, E201K, L202P (P), E205D, K206T, <i>P209A</i> (P), A210G, E213K, A218T, D219N, R220Q, K227R, E230K , S234A, P237T, P239S, N244E, N245S, N245D, <i>P251L</i> (P), T255I, A256P, Q258K, K259E (S), S261N, E267K, A272T, L275P (P), V286M, <i>P290L</i> (P), A293V, T302R, <i>F320L</i> (P), Q325E, <i>P330L</i> (P), R331S, -4A>C, -22G>A, -33G>A, -40G>C, -42G>T, -45CG>C, 502_504del, 643_648del, 986_986del, 490706_490720del
DLM/ PTM	<i>fbIA</i>	T4N, A7G, <i>G12S</i> (P), <i>R14H</i> (P), L22V, L25M, A30T, S32P, S35P, S35A, A37V, S42C, A43G, I53V, I53L, I53T, <u>V58I</u> , <i>G71S</i> (P), <i>R77H</i> (P), <i>R77L</i> (P), Q81R, D83N, <i>W101R</i> (S,P), A121V, <i>Y123S</i> (P), <i>Y123C</i> (P), <i>P124R</i> (P), L125V, S126P (S,P), T129S, A131T, D134N, <i>P138L</i> (P), <i>G139D</i> (P), D158N, K165T, A166V, A178S, Q179K, <i>P181S</i> (P), <i>G189S</i> (P), S194N, A196V, I209V , V218I, <i>A232E</i> (P), A238T, <i>P245S</i> (S,P), K250* ,

M255T (P), D266N, A269S, A271T, *G277D* (P), A278G, C280R, C287Y, V290M, G293S, A296V, D299A, *M310T* (S,P), V303M, A315V, A322V, A331G, -8C>A, -42C>T, 196_283del, 866_866del

DLM/PTM *fbiB* S8F, E13K, *G19W* (S,P), *E22G* (P), *R24L* (P), *G26R* (P), *P38Q* (P), *P38A* (P), *K51N* (S,P), *E57A* (P), *R59W* (S,P), L60M, *P64A* (P), D66N, Q69P, E79R, A104T, *A105T* (S), G114D, A119T, A127T, T131I, L132V, G135E, *G141S* (P), V142I, A145T, *Q160H* (P), V165I, A171S, R180C, E186A, E186D, V188M, V192I, G221S, V222A, D225N, N238D, L243F, A246S, E247D, R253H, *R260L* (S,P), V263I, *R265W* (S), *P270Q* (P), V280I, *H290Y* (S,P), R293Q, V298M, S323N, D324E, P327L, A328T, A328S, *R334G* (S,P), *R334W* (S,P), *G338S* (P), D343E, A344V, E346Q, I349V, I349M, *A357E* (P), A364T, T366N, T366I, A368T, E369Q, *G394S* (S,P), S395N, *I402T* (P), *R409H* (P), D410G, D410H, *P415L* (P), L435S, *P438S* (P), V439A, P440S, A441T, K448E, 1175_1266del

DLM/PTM *fbiC* V1L, G6S, V16I, V17L, P18A, P19R, A21P, A25T, *R27W* (S,P), R31Q, A33T, A33V, V37A, V37G, A45T, A47T, T49A, C59Y, *R75W* (S,P), *F91C* (S,P), *P93L* (S,P), *P93S* (P), *R96L* (P), *R96G* (S,P), ***C105R*** (S,P), *L114R* (P), T121A, D126E, D130N, D130E, R134Q, A136V, E137Q, *F145L* (P), *T146S* (S), *R150S* (S,P), E152D, A153V, *R159G* (S,P), E160D, *E160G* (P), D168A, *S169F* (S,P), S172C, S172F, A208V, *M211I* (S), *R221Q* (S), *D237G* (P), *P238L* (P), A239E, *R243S* (P), *T257A* (S,P), *E269G* (S,P), *D272G* (S,P), L274F, H275Y, *R278L* (S,P), H281Q, K282R, *R296H* (P), A302V, A305V, F306L, *P307S* (P), I311V, *D313G* (S,P), Y314F, A321T, *P327L* (P), R330S, *P334A* (P), *P334S* (P), G340R, D341E, C343W, *R344W* (S), D375N, *D375G* (P), ***L377P*** (P), M388V, M388L, Q395H, Q400R, A401G, V410M, *R411W* (S), R411S, A418E, *P420S* (P), G423S, D427H, *W435G* (P), P438R, V441A, A442V, S443C, *R446W* (P), Q456E, *R458P* (P), V462L, *R463C* (P), *D472A* (P), V495L, V495M, L496R, A497T, A497P, D511E, A516T, T519N, *T519I*, *G522R* (P), *G522S* (P), *G522A* (P), A527S, V540A, *F554L* (S,P), T560A, K571E, *R587W* (S,P), A588V, M601I, *M601T* (S,P), I605V, D606E, *P610S* (P), T612A, A620V, N640T, N640S, *G646E* (P), S648T, *W652R* (S,P), I654V, *E658D*, T663N, *D674H* (P), *P686T* (P), L689W, T695M, *G711A* (S,P), A721V, N724K, I729V, R732H, *R732L* (S,P), *G734S* (P), *H746Y* (P), Q747H, *P750R* (P), *L753R* (P), A756V, *R758C* (S,P), *P759R* (P), *P759T* (P), *H770Y* (P), *G792A* (S,P), E801N, *G808D* (S,P), M812V, *E813A* (P), E813D, T815A, E823G, E823Q, H824Y, A827S, *G841R* (P), *P843Q* (P), *P843A* (P), *R845H* (P), L853P, A855L, ***A856P***, *857W, 24_25insTCCACCGCTCTGCCGAGTCCC,342_347del, 785_786insCAT, 897_898del, 1132_1133insCTT, 1133_1134insTTT, 1205_1206del, 1223_1252del, 1651_1731del, 1871_1871del, 2162_2162del, 2331_2335del, 2545_2546insTCACATACGCCCTGCTTGC, 2551_2552insACGCC, 2562_2623del, 1305491_1305500del, -3A>G, -13A>AC, -14G>A, -17A>G, -18C>A, -23G>A, -27A>G, -29A>C, -29A>G, -30G>C, -32A>C, -33G>C, -41G>A, -46G>C, -46GGT>G, -49C>G

DLM/PTM *fbiD* P5L, I13V, *P28L* (P), F30L, *S31W* (P), V38L, V39A, V44I, A50T, A51S, G52S, V53G, I62V, E66Q, A70T, A81P, P85A, A98T, A99T, R101C, A104V, E105A, G106V, L144R, T146S, V150I, H159N, *R186C* (S,P), V189I, A206T, A210S, -21C>T, -24G>A, -31C>T, -38A>G, -45G>A

* Bedaquiline (BDQ), Delamanid (DLM); Pretomanid (PTM) has similar resistance mechanisms to DLM. In bold, known resistant mutations (see **S3 Table**); underlined: known susceptible mutations (see **S3 Table**); italic: with one parameter predicting to have a functional effect: B = SUSPECT-BDQ predicted as resistant (only for *atpE*); S = snap2 score ≥ 50 ; P = Provean score ≤ -4 ; M = mCSM predicted stability change ($\Delta\Delta G$) below -1. There is no prediction for indels or variants in the promoter region.

S8 Table. Loss of function mutations in the *ndh* gene.

Mutation	Frequency	Lineage	Resistance profile *
Q4*	1	1 (1.1.2)	S
Y56*	1	3	XDR
Q57*	2	1 (1.1.1)	S
Y112*	1	1 (1.1.2)	S
C273*	1	3	XDR
970_971insGG	8	2(n=5;2.2.1); 4(4.1.2.1(n=2),4.3.4.2(n=1)	MDR(4), XDR(4)
304_304del	82	2(2.2.1.1)	MDR(76), XDR(5), DR(1)
149_158del	1	1(1.2.1)	MDR
972_973inC	1	2(2.2.1)	XDR
970_971insG	2	2(2.2.1),4(4.3.3)	XDR,DR
1120_1120del	1	1(1.1.1)	DR
965_965del	2	2(2.2.1)	MDR
15_15del	2	4(4.2.2)	MDR
838_838del	1	1(1.2.1)	MDR
902_903insG	2	2(2.2.1)	MDR,XDR
1007_1008insGC	1	4(4.4.2)	MDR
941_942insGGGTA	1	2(2.2.1)	XDR
1347_1348insA	1	4(4.1.2.1)	S
293_294insG	6	4(4.1.2.1)	XDR
330_337del	1	4(4.3.1)	XDR
900_901insC	1	2(2.2.1)	MDR
633_634insTG	1	1(1.2.1)	DR
199_200insG	1	4(4.1.2.1)	XDR
760_761insC	1	2(2.2.1)	MDR
1206_1207insCG	1	4(4.9)	MDR
2098715_2102885del	1	2(2.2.1)	XDR
2102284_2103965del	3	2(2.2.1)	XDR
2098094_2101927del	1	4(4.3.3)	DR
2096545_2103436del	1	1(1.2.1)	S
2098085_2104480del	1	2(2.2.1)	XDR
2097074_2107484del	1	2(2.2.1)	MDR
2102559_2103251del	1	1(1.1.2)	S

* Resistance profile: DR = Drug-resistant, S = Susceptible

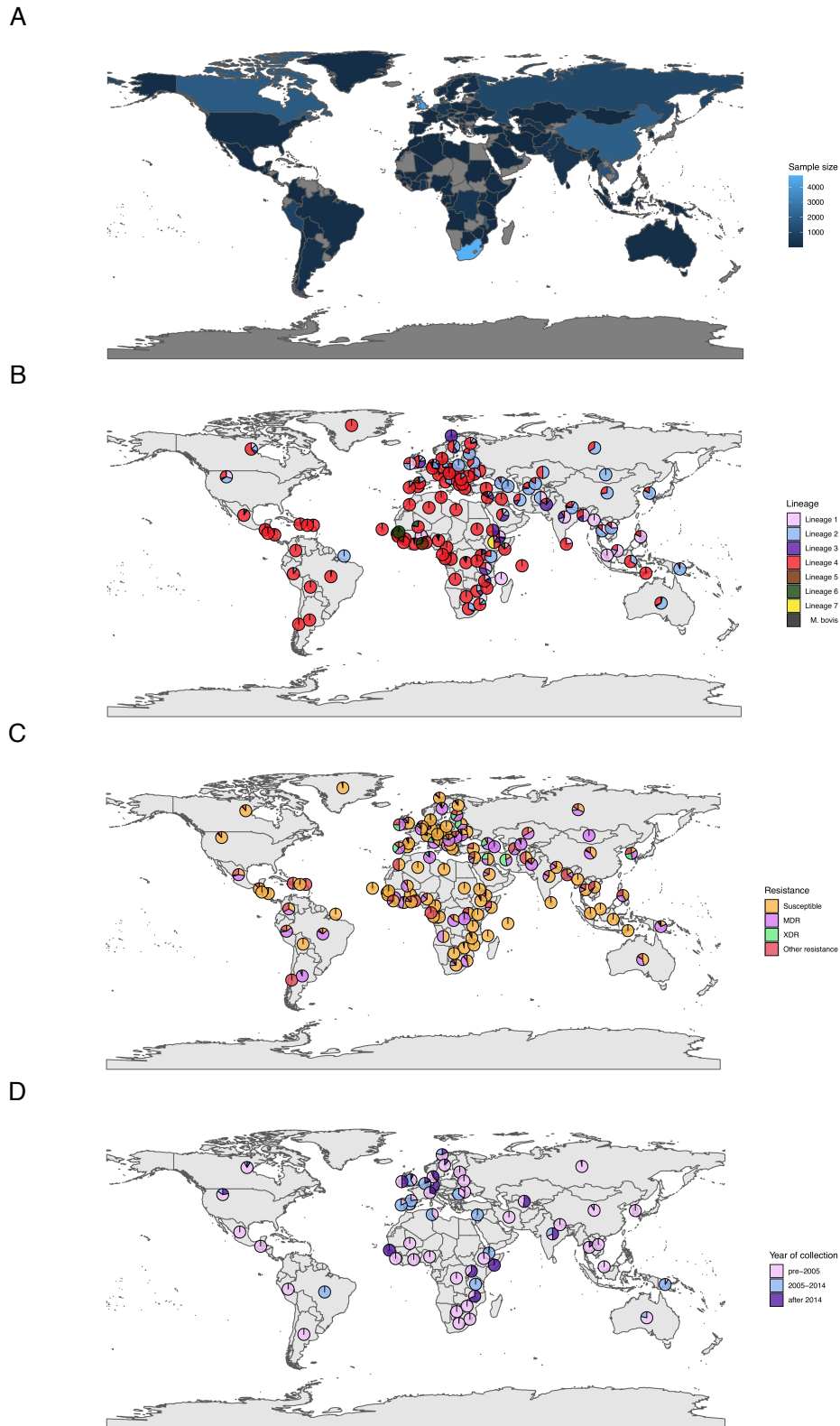
S9 Table. Phenotypic data from the Portuguese *M. tuberculosis* isolates

Isolate	<i>mmpR5/Rv067</i> 8 Mutation ^a	MIC (mg/L)		Resistance Type	Phenotypic Drug Resistance ^b	Genotype ^c
		BDQ	CFZ			
MTB1	lle67fs	0.25	1	XDR	INH ^R RIF ^R STR ^R EMB ^R PZA ^R AMK ^S CAP ^R KAN ^R CIP ^R	L4.3.4.2/SIT20/LAM1/Lisboa3
MTB2	WT	≤0.01 5	0.12 5	XDR	INH ^R RIF ^R STR ^R EMB ^R PZA ^R AMK ^S CAP ^S KAN ^R CIP ^R	L4.3.4.2/SIT20/LAM1/Lisboa3
MTB3	WT	≤0.01 5	0.25	Susceptible	Pan susceptible	L4.3.4.1/SIT17/LAM2/NC
MTB4	WT	≤0.01 5	0.25	MDR	INH ^R RIF ^R STR ^R EMB ^R PZA ^R AMK ^R CAP ^R KAN ^R CIP ^S	L4.3.4.2/SIT1106/LAM4/Q1
MTB5	WT	0.03	0.25	MDR	INH ^R RIF ^R STR ^S EMB ^S PZA ^S AMK ^S CAP ^S KAN ^S CIP ^S	L4.1.2.1/SIT53/T1/NC
H37Rv (ATCC 27294)	WT	0.03	0.25	Susceptible (Reference Strain)	Pan susceptible	-

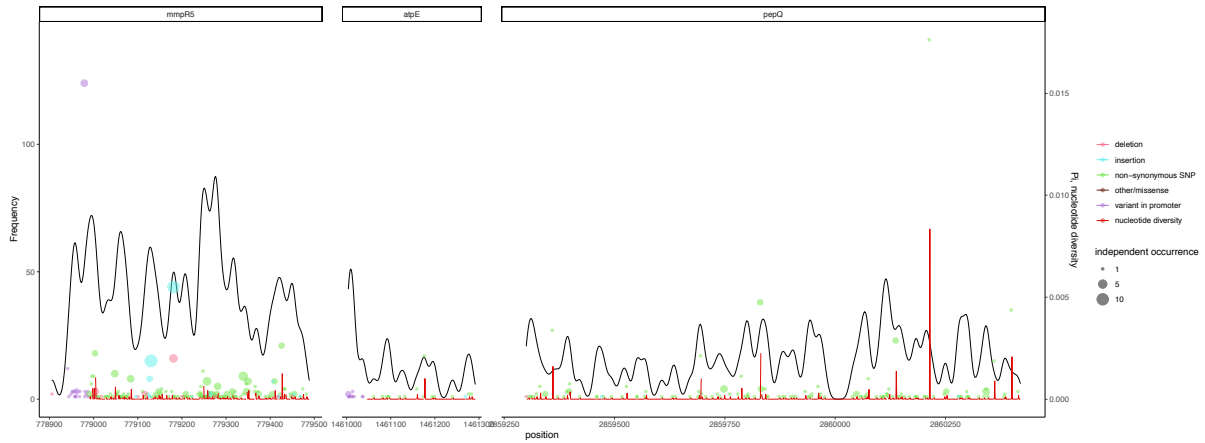
^a WT – wildtype allele for *mmpR5/Rv0678*;

^b R and S in superscript denotes phenotypic resistance or susceptibility to given drug, respectively. INH, isoniazid; RIF, Rifampicin; STR, streptomycin; EMB, ethambutol; PZA, pyrazinamide; AMK, amikacin; CAP, capreomycin; KAN, kanamycin; CIP, ciprofloxacin;

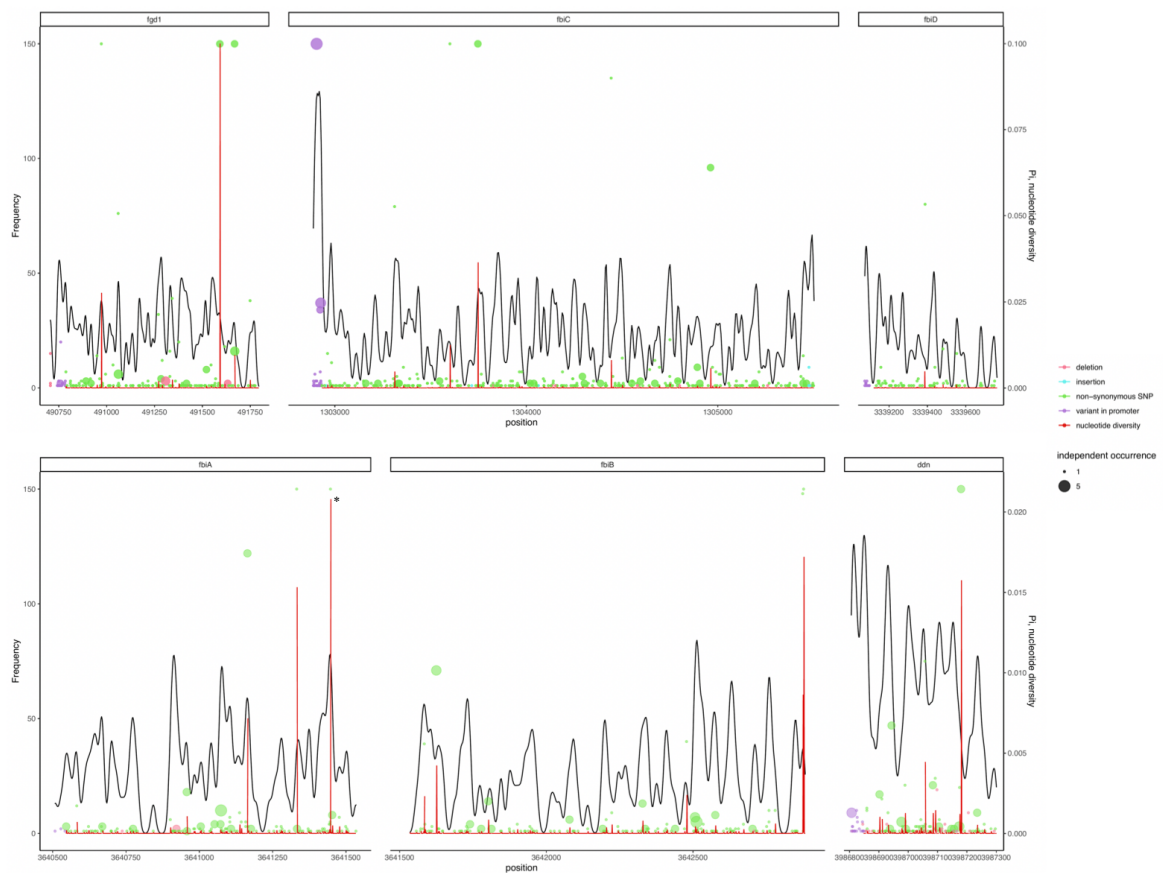
^c NC – non-clustered isolate.



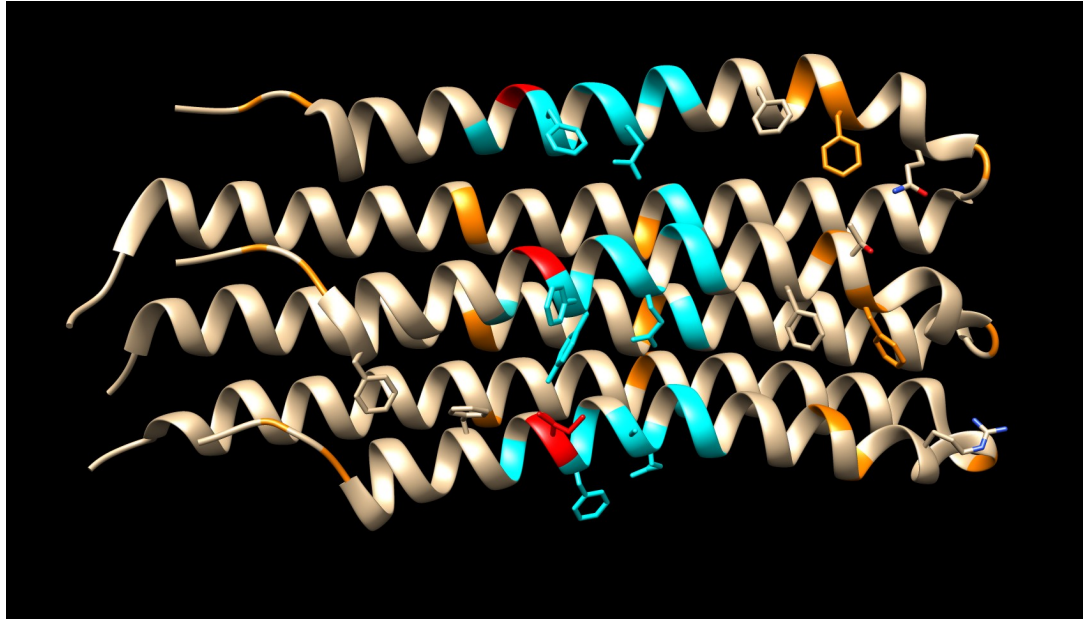
S1 Figure. The isolates analysed by country: **(A)** Sample size; **(B)** Lineage; **(C)** Resistance; **(D)** Year of collection. The R (v3.4.3) statistical package was used to generate the maps (<https://www.r-project.org>).



S2 Figure. Density of mutations and nucleotide diversity (Nei's Pi) along BDQ resistance genes. Density line is represented in black. Nucleotide diversity (only non-synonymous SNPs) by position (Nei's Pi) is represented in red. Left vertical axis is frequency of each mutation represented by a point (type of mutation differ in colour), and size represents the independent occurrence of each mutation in the phylogenetic tree.

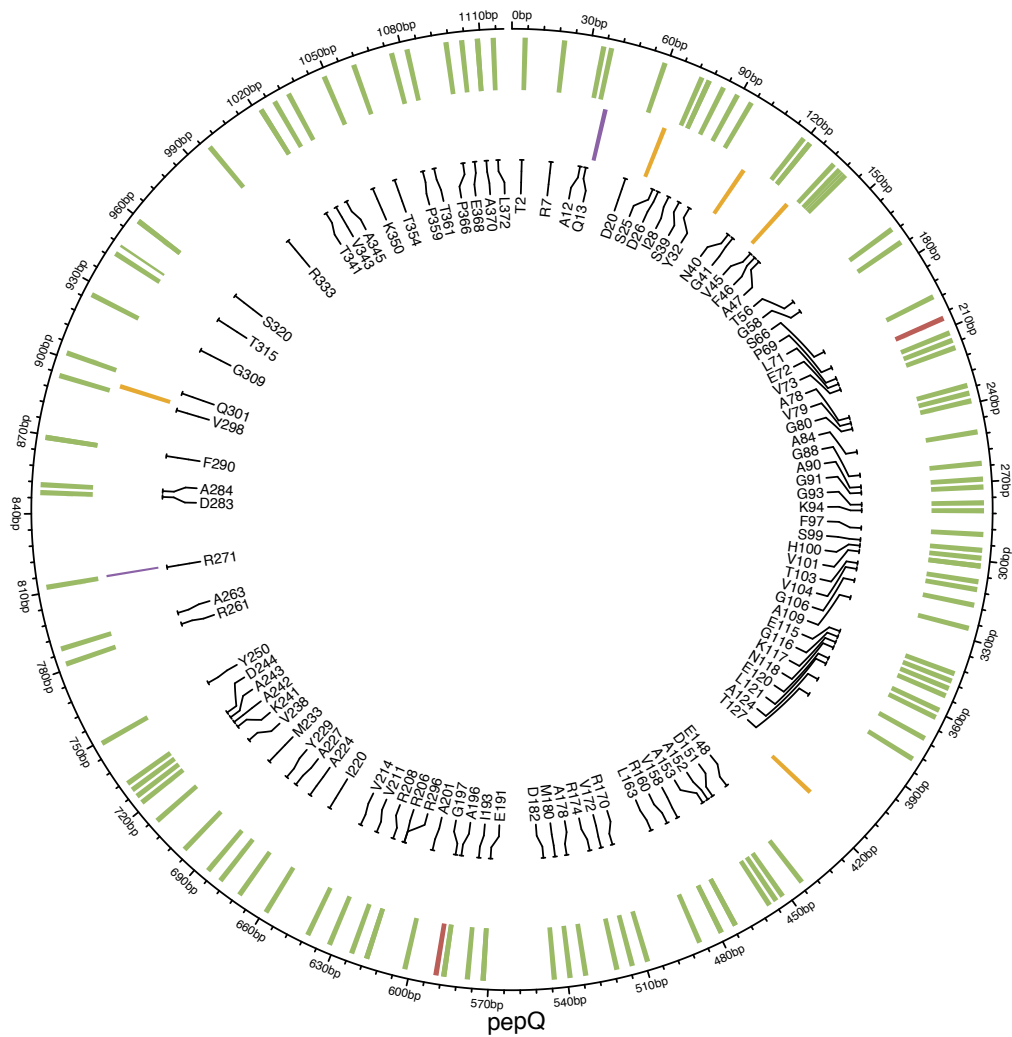


S3 Figure. Density of mutations and nucleotide diversity (Nei's P_i) along Delamanid (DLM) and Pretomanid (PTM) resistance genes. Density line is represented in black. Nucleotide diversity (only non-synonymous SNPs) by position (Nei's P_i) is represented in red. Left vertical axis is frequency of each mutation represented by a point (type of mutation differ in colour), and size represents the independent occurrence of each mutation in the phylogenetic tree; * Nucleotide diversity at position 491592 in *fgd1* is 0.168; ** Mutations with frequency >150 have been represented at 150.

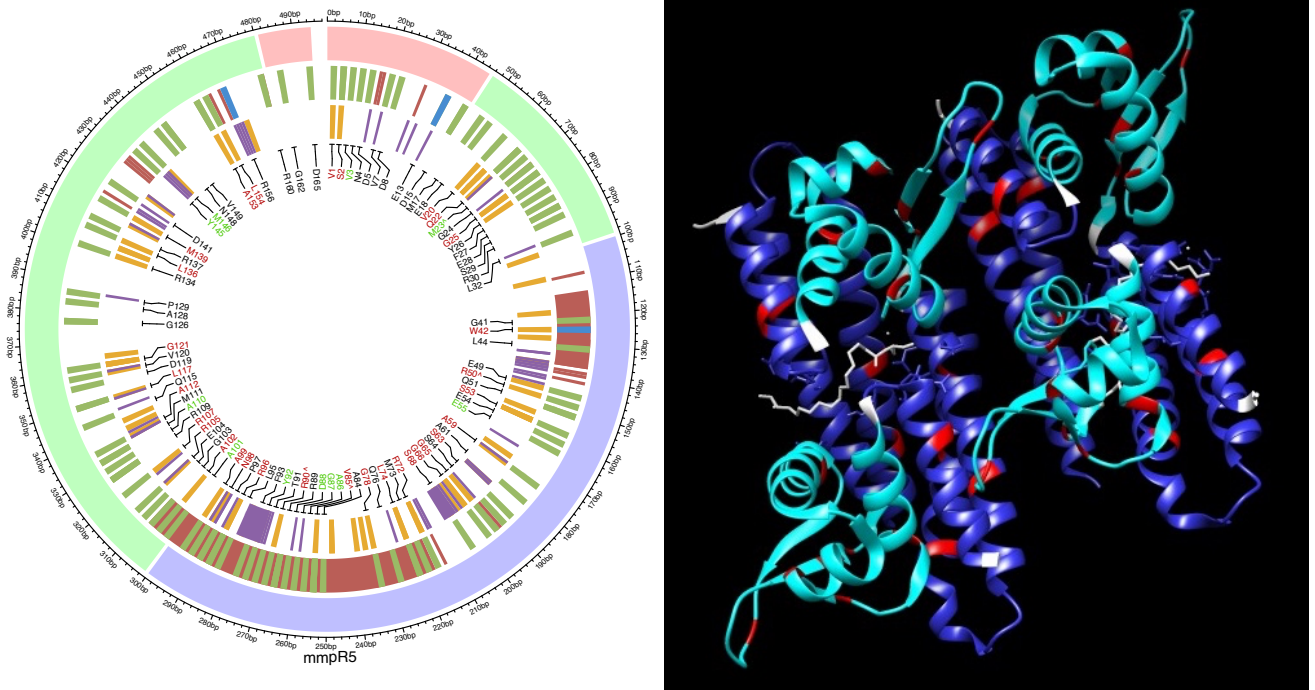


M.tuberculosis/1-81 - MDPT I AAGAL I GGGL I **M**AGGA I GAG **I**GDGVAGNAL I SGVARQP **S**AQGR**L****F****T**FF I TVGLVEAAYF **I**NLAFMALFVFA **I****P****V****K**-
M.phlei/1-83 - MADPT I VAGAL I GGGL I MAGGA I GAG I **G****S**IAGNAL I SGVARQPEAQSRLFT PFF I TV**G****L****V****E****A****A****T****F****I****N****S**AFMALFVFATPGAS

S4 Figure. Protein structure of *atpE* C9 ring and sequence. The c9 ring is composed by 3 subunits. Highlighted in blue are the residues known to interact with Bedaquiline (BDQ), in orange the residues predicted to give resistance, and in red the known and previously reported mutation associated with BDQ drug resistance.



S5 Figure. Non-synonymous SNPs and indels in the *pepQ* gene, a candidate for bedaquiline (BDQ) resistance. From outside to inside, first track represents indels (in red) and SNPs (in green) identified in the ~33k isolates. SNPs leading to premature stop codons in blue. The second track represents known resistant SNPs (yellow) and indels (purple). Labels show the residues where SNPs are identified in the ~33k isolates: in black residues with not known association to susceptibility/resistance; in green residues with known association to susceptibility; in red residues with known association to increased minimum inhibitory concentration (MIC) values.

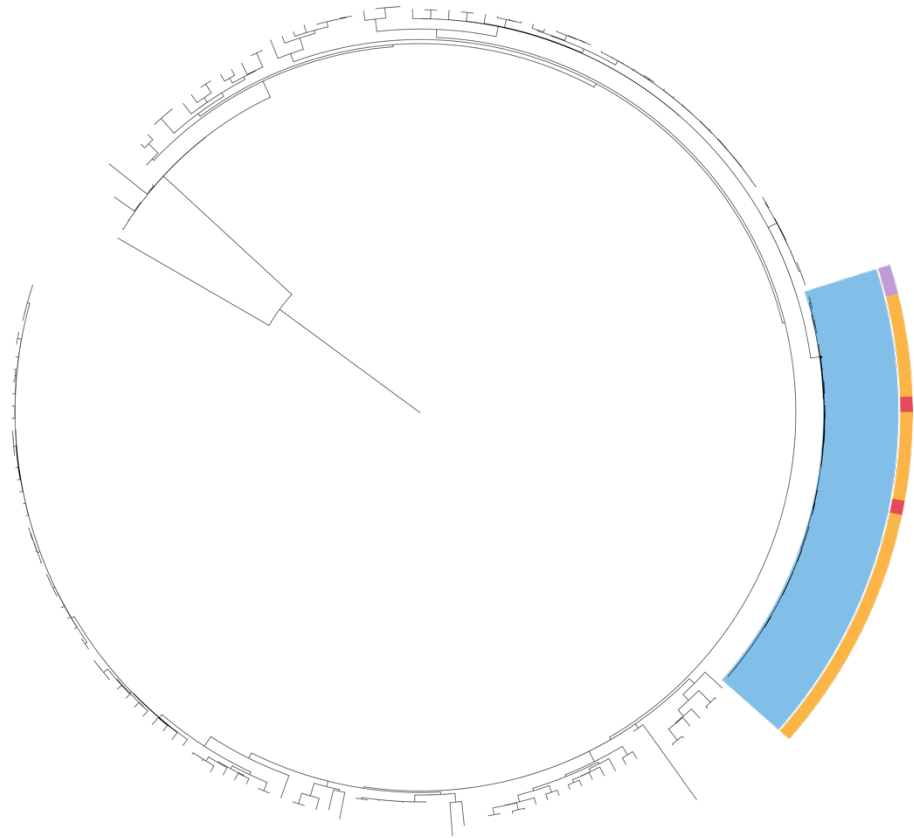


S6 Figure. The *mmpR5* gene variants position and protein structure. **(left)** Non-synonymous SNPs and indels along *mmpR5* gene. From outside to inside, first track represents the different domains of the protein: in red non-characterised; in green dimerization domain; in blue binding domain. The second track show indels (in red) and SNPs (in green) identified in the ~33k isolates. SNPs leading to premature stop codons in blue. The third track represents known resistant SNPs (yellow) and indels (purple). Labels show the residues where SNPs are identified in the ~33k isolates: in black residues with not known association to susceptibility/resistance; in green residues with known association to susceptibility; in red residues with known association to increased minimum inhibitory concentration; ^ = residues with association to resistance and susceptibility depending on alternate allele. Non-synonymous SNPs and indels position along the *mmpR5* gene. SNPs are coloured in green, indels in red. **(right)** Protein structure of *mmpR5* showing in red SNPs that have already seen reported as associated with bedaquiline resistance. Dark blue corresponds to the binding domain.

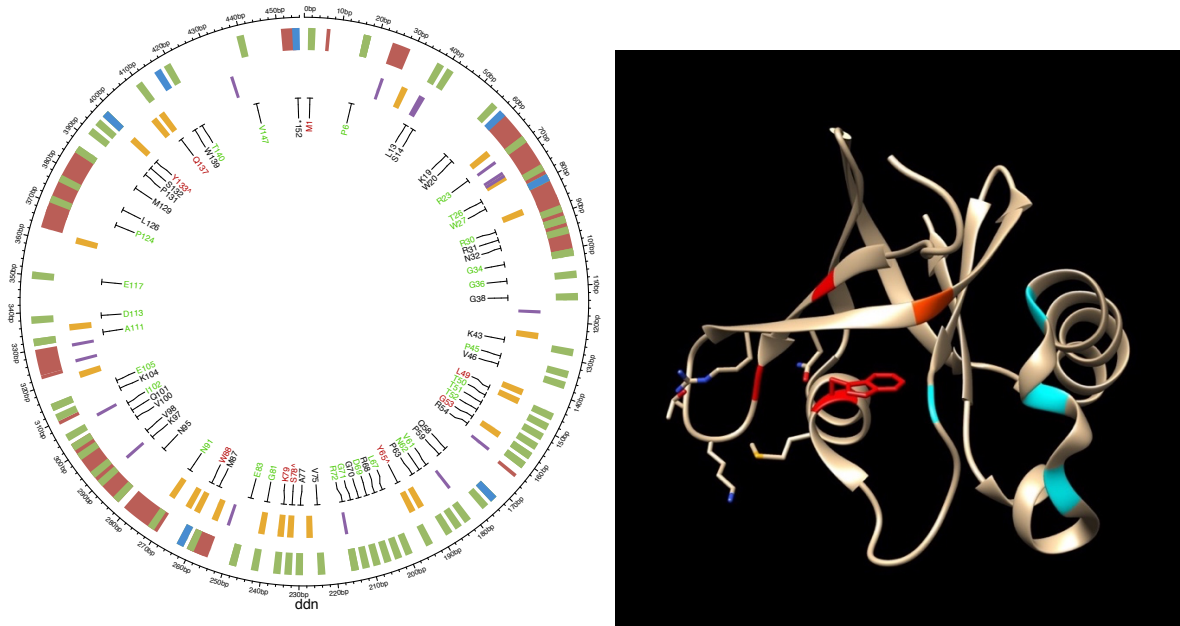
Tree scale: 0.1

mmpR5 192_193insG

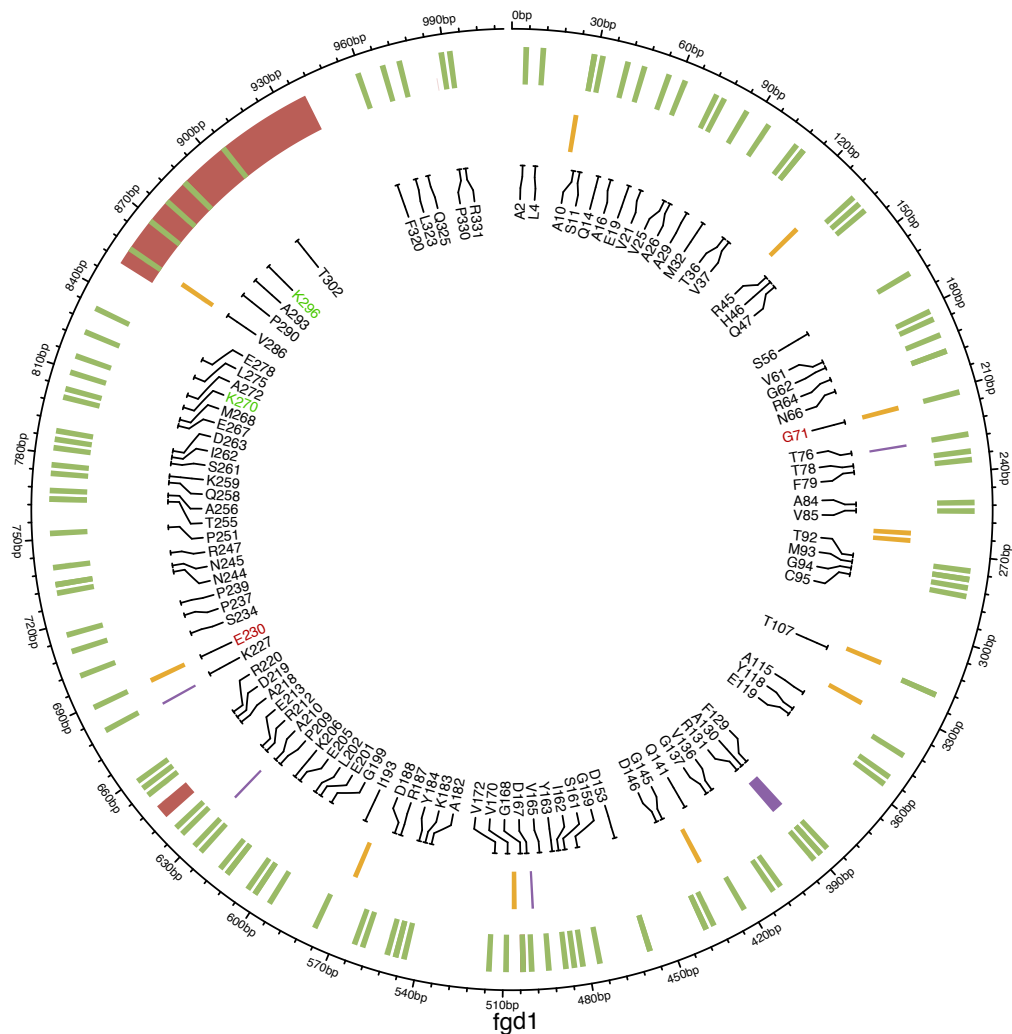
Resistance
Susceptible
MDR
XDR
Other resistance



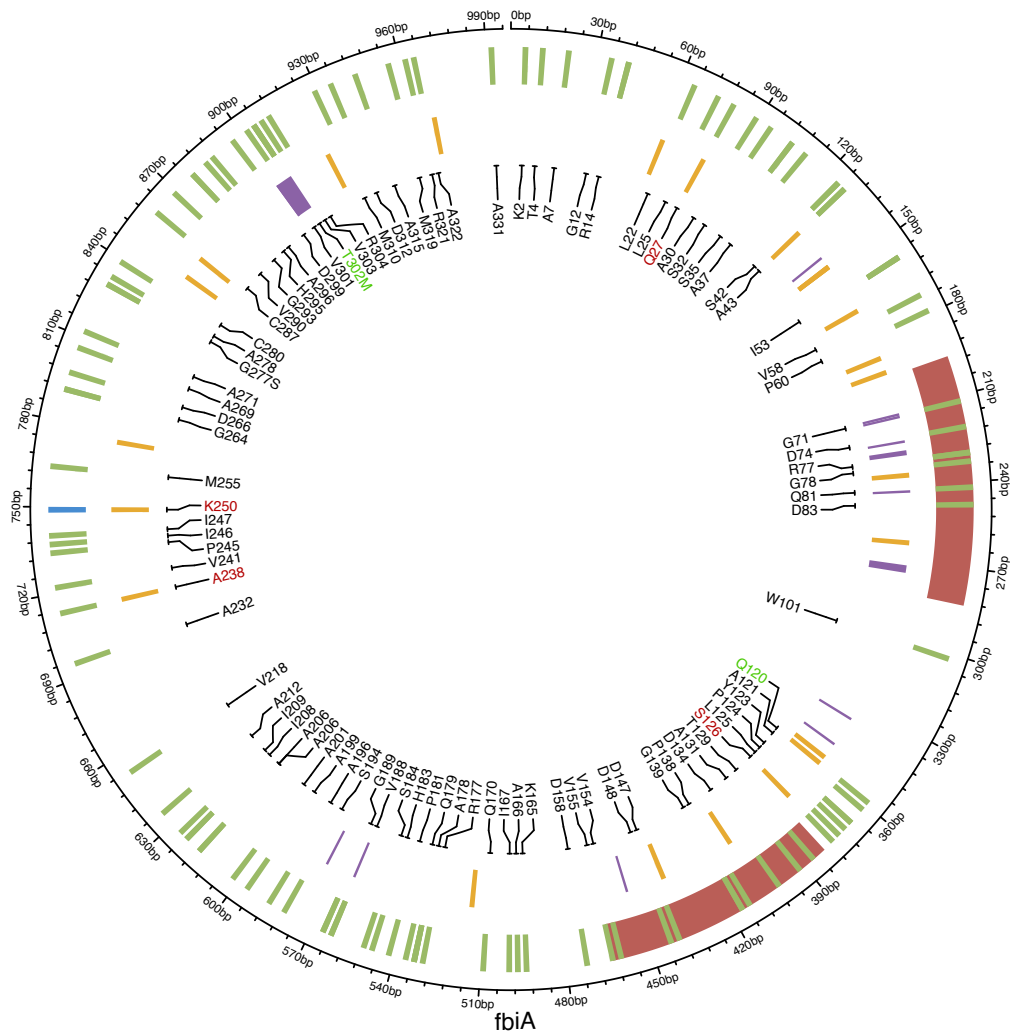
S7 Figure. Phylogenetic tree of lineage 4 strains. Coloured in blue are the samples that present the frameshift (192_193insG; I67F) in *mmpR5* for bedaquiline resistance. The outer track shows the resistance profile of the samples harbouring the frameshift.



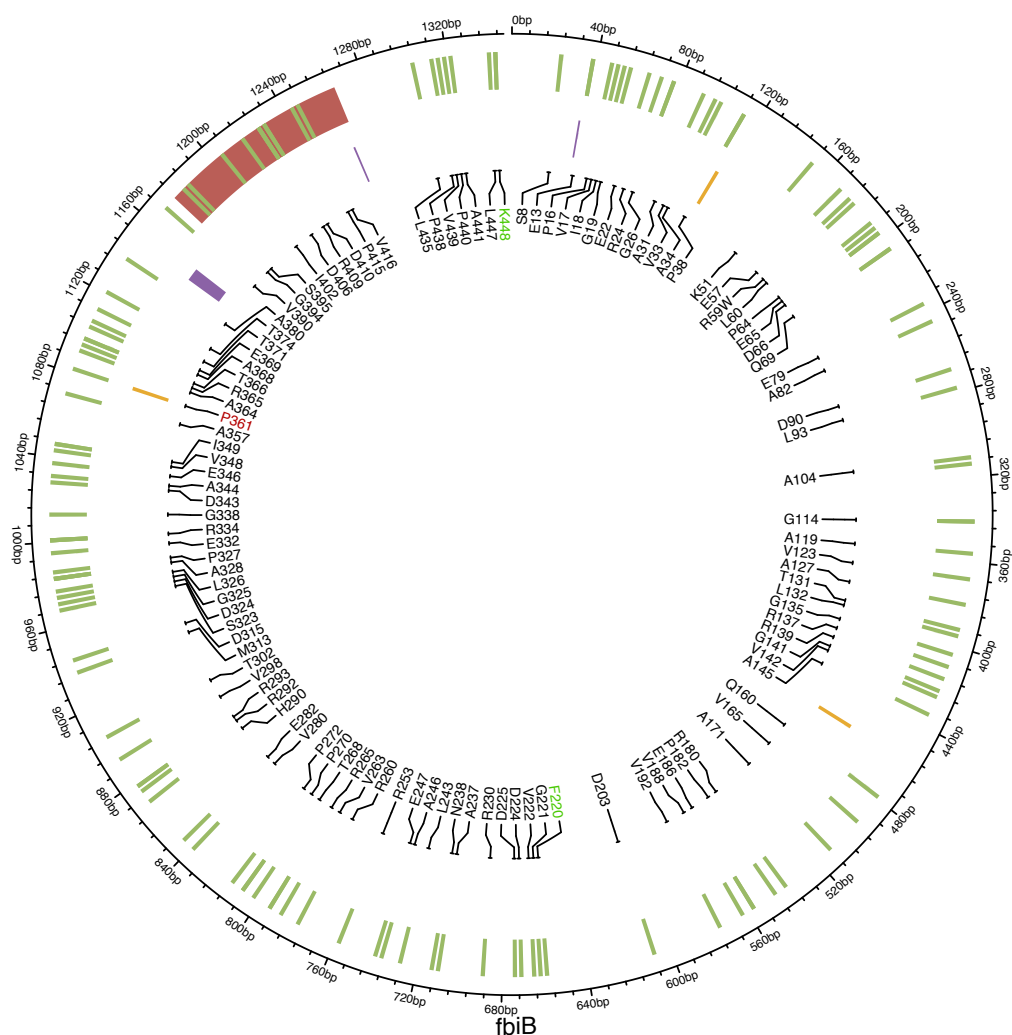
S8 Figure. A) Non-synonymous SNPs and indels along *ddn* gene. From outside to inside, first track represents indels (in red) and SNPs (in green) identified in the ~33k isolates. SNPs leading to premature stop codons in blue. The second track represents known resistant SNPs (yellow) and indels (purple). Labels show the residues where SNPs are identified in the ~33k isolates: in black residues with not known association to susceptibility/resistance; in green residues with known association to susceptibility; in red residues with known association to increased MIC; ^ = residues with association to resistance and susceptibility depending on alternate allele or drug (delamanid (DLM)/pretomanid (PTM)). **B)** Protein structure of *ddn* gene showing in red SNPs that have already seen reported as associated with DLM/PTM resistance, in blue residues known to be involved in PTM interaction, and in orange residues involved in PTM interaction that also confer resistance to DLM.



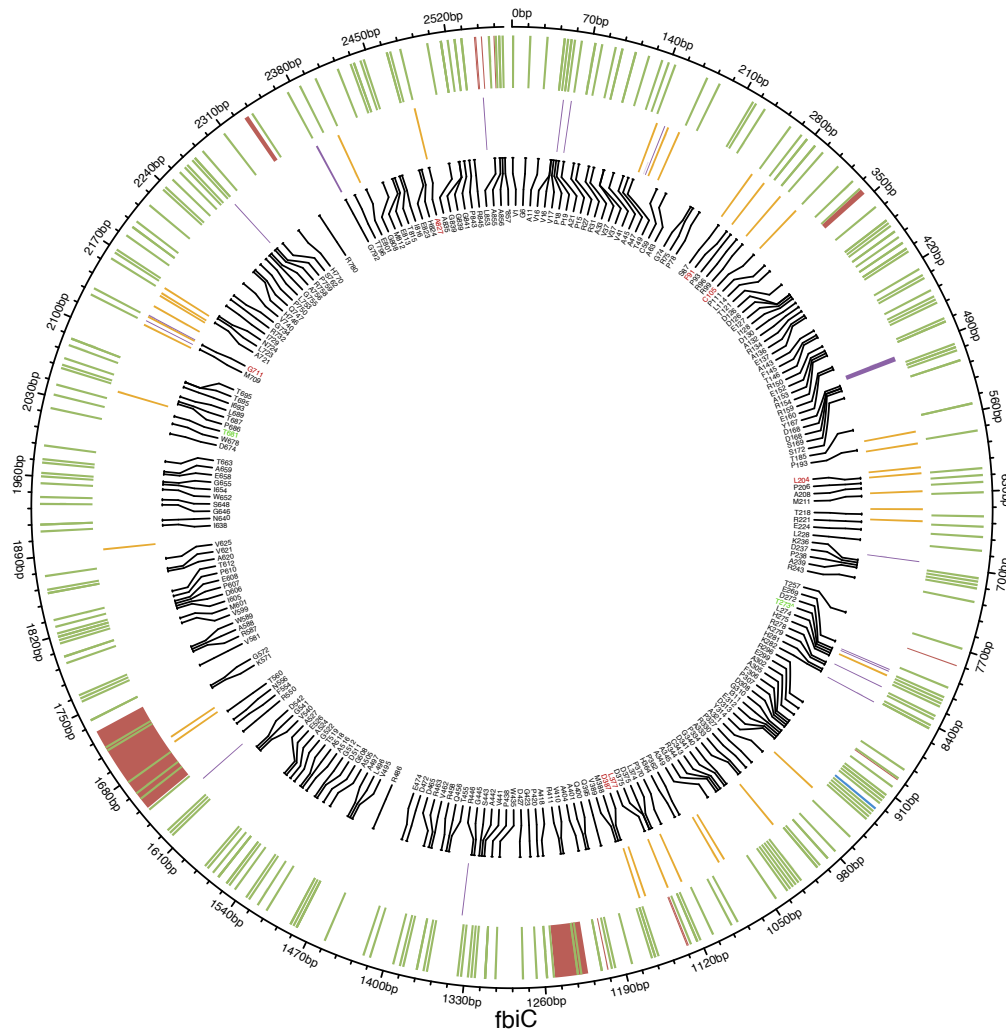
S9 Figure. Non-synonymous SNPs and indels in the *fgd1* gene, a candidate for delamanid (DLM)/pretomanid (PTM) resistance. From outside to inside, first track represents indels (in red) and SNPs (in green) identified in the ~33k isolates. SNPs leading to premature stop codons in blue. The second track represents known resistant SNPs (yellow) and indels (purple). Labels show the residues where SNPs are identified in the ~33k isolates: in black residues with not known association to susceptibility/resistance; in green residues with known association to susceptibility; in red residues with known association to increased MIC values.



S10 Figure. Non-synonymous SNPs and indels in the *fbiA* gene, a candidate for delamanid (DLM)/pretomanid (PTM) resistance. From outside to inside, first track represents indels (in red) and SNPs (in green) identified in the ~33k isolates. SNPs leading to premature stop codons in blue. The second track represents known resistant SNPs (yellow) and indels (purple). Labels show the residues where SNPs are identified in the ~33k isolates: in black residues with not known association to susceptibility/resistance; in green residues with known association to susceptibility; in red residues with known association to increased MIC.



S11 Figure. Non-synonymous SNPs and indels in the *fbiB* gene, a candidate for delamanid (DLM)/pretomanid (PTM) resistance. From outside to inside, first track represents indels (in red) and SNPs (in green) identified in the ~33k isolates. SNPs leading to premature stop codons in blue. The second track represents known resistant SNPs (yellow) and indels (purple). Labels show the residues where SNPs are identified in the ~33k isolates: in black residues with not known association to susceptibility/resistance; in green residues with known association to susceptibility; in red residues with known association to increased MIC.



S12 Figure. Non-synonymous SNPs and indels in the *fbiC* gene, a candidate for delamanid (DML)/pretomanid (PTM) resistance. From outside to inside, first track represents indels (in red) and SNPs (in green) identified in the ~33k isolates. SNPs leading to premature stop codons in blue. The second track represents known resistant SNPs (yellow) and indels (purple). Labels show the residues where SNPs are identified in the ~33k isolates: in black residues with no known association to susceptibility/resistance; in green residues with known association to susceptibility; in red residues with known association to increased MIC; ^ = residues with association to resistance and susceptibility depending on alternate allele or drug (DLM/PTM).

CHAPTER 5

Functional genetic variation in *pe/ppe*
genes contributes to diversity in
Mycobacterium tuberculosis lineages and
potential interaction with the human host

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	lsh1704009	Title	
First Name(s)	Paula Josefina		
Surname/Family Name	Gómez González		
Thesis Title	Analysis of Mycobacterium tuberculosis 'omics data to inform on loci linked to drug resistance, pathogenicity and virulence		
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?			
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.

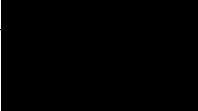
SECTION C – Prepared for publication, but not yet published

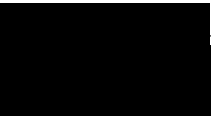
Where is the work intended to be published?	Genome Biology
Please list the paper's authors in the intended authorship order:	Gomez-Gonzalez, PJ; Grabowska, AD; Tientcheu, L; Hibberd, ML; Campino, S; Phelan, JE; Clark, TG
Stage of publication	Submitted

SECTION D – Multi-authored work

<p>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)</p>	<p>I cultured and extracted DNA from clinical isolates. I received the long-read sequence data and compiled a data set together with publicly available PacBio genomes. I performed the bioinformatic analysis, consisting in assembly, alignment, variant calling and phylogenetics. I designed a extraction pipeline for the pe/ppe genes, and custom scripts were used for the population genetics analysis. All statistical analysis and plotting was performed in R. I wrote the first draft of the manuscript and circulated to co-authors, and after receiving comments I edited the last version. I submitted the manuscript to the journal.</p>
---	--

SECTION E

Student Signature	
Date	28/01/2022

Supervisor Signature	
Date	28/01/2022

Functional genetic variation in *pe/ppa* genes contributes to diversity in *Mycobacterium tuberculosis* lineages and potential interactions with the human host

Paula Josefina Gómez-González ¹ paula-josefina.gomez-gonzalez@lshtm.ac.uk
Anna D. Grabowska ² dr.anna.grabowska@gmail.com
Leopold Tientcheu ³ leopold.tientcheu@lshtm.ac.uk
Martin L. Hibberd ¹ martin.hibberd@lshtm.ac.uk
Susana Campino ¹ susana.campino@lshtm.ac.uk
Jody E. Phelan ¹ jody.phelan@lshtm.ac.uk
Taane G. Clark ^{1,4,*} taane.clark@lshtm.ac.uk

1. Faculty of Infectious and Tropical Diseases, London School of Hygiene & Tropical Medicine, WC1E 7HT London, UK
2. Department of Biophysics, Physiology and Pathophysiology, Medical University of Warsaw, 02-004 Warsaw, Poland
3. MRC Unit The Gambia at the London School of Hygiene and Tropical Medicine, Vaccines and Immunity Theme, Atlantic Road, Fajara, The Gambia.
4. Faculty of Epidemiology and Population Health, London School of Hygiene & Tropical Medicine, WC1E 7HT London, UK

* Correspondence: taane.clark@lshtm.ac.uk, Department of Infection Biology, Faculty of Infectious and Tropical Diseases, London School of Hygiene & Tropical Medicine, Keppel Street, London, UK

Genome Biology

ABSTRACT

Background: Around 10% of the coding potential of *Mycobacterium tuberculosis* is constituted by two poorly understood gene families, the *pe* and *ppe* loci. Their repetitive nature and high GC content have hindered sequence analysis, leading to their exclusion from whole-genome studies. Although the functions of many of *pe/ppe* genes are still unknown, some are involved in host-pathogen interactions and thereby promising targets for vaccine development. Understanding the genetic diversity of *pe/ppe* families is essential to facilitate their potential translation into tools for tuberculosis prevention and treatment.

Results: We performed an *in silico* sequence analysis of the 169 *pe/ppe* genes across 72 long-read assemblies representing 6 different lineages of *M. tuberculosis* and *M. bovis* BCG. The characterised genes were classified into three groups based on the level of protein sequence conservation relative to the reference H37Rv, finding that indels in the *pe_pgrs* and *ppe_mptr* sub-families were the main drivers of structural variation. Overall, every isolate had >50% of its *pe/ppe* genes conserved. We observed gene rearrangements, such as duplications and changes in the open reading frames leading to gene fusions, notably between *pe* and *pe_pgrs* genes. Inter-strain diversity revealed lineage-specific SNPs and indels among the *pe/ppe* genes.

Conclusions: The high level of *pe/ppe* genes conservation, together with the lineage-specific findings, suggest their phylogenetic informativeness. However, structural variants and gene rearrangements differing from the reference were also identified, with potential implications for pathogenicity. Overall, improving our knowledge on these elusive complex gene families can inform the development of tools for tuberculosis control.

Word count: 250 words.

Keywords

Mycobacterium tuberculosis, *pe/ppe* genes, genomics

BACKGROUND

Tuberculosis disease (TB), caused by *Mycobacterium tuberculosis* bacteria, is a major global public health problem with drug resistance making its control difficult [1]. The available vaccine, Bacillus Calmette-Guérin (BCG), has limited efficacy and recent attempts to develop more protective vaccines have been unsuccessful, in part due to the insufficient understanding of host-pathogen interactions [2]. The *M. tuberculosis sensu stricto* genome has a low overall genetic diversity and a striking clonal population structure, with nine lineages (L1-L9), which are postulated to have different impacts on pathogenesis, disease outcome and vaccine efficacy [3]. For example, modern lineages, such as Beijing (L2) and Euro-American Haarlem (L4) strains exhibit more virulent phenotypes compared to ancient lineages, such as East African Indian (L1) [4]. Whilst some genetic differences between lineages have been identified [5], the molecular mechanisms responsible for differences in pathogenesis and virulence remain largely unknown.

The *Mycobacterium tuberculosis* genome (4.4 Mb) has unique *pe* (100 loci) and *ppe* (69 loci) genes, found in larger numbers in pathogenic mycobacteria than saprophytic or avirulent species [6–8], and therefore suggested to play a role in pathogenicity and virulence. Members of these two families constitute ~10% of the *M. tuberculosis* genome, and have a conserved domain with 110 and 180 amino acids respectively at the N-terminal, within which signature proline-glutamate (PE) and proline-proline-glutamate (PPE) motifs can be identified in most of the protein products [9]. In contrast, the C-terminal sequences are more variable and of various sizes, ranging from zero to more than a thousand residues in length. The *pe* and *ppe* genes are closely associated with the ESX secretion systems, and their evolution and expansion has been proposed to be linked to a series of duplication events of the ESAT-6 gene

clusters [6, 10], together with insertions, deletions and homologous recombination [11]. Through the phylogenetic reconstruction of their protein sequences, *pe* and *ppe* genes have been grouped into five sub-families each, with *pe_pgrs* and *ppe_mptr* being the most recent and polymorphic [6]. These two groups represent some of the most variable of all *M. tuberculosis* genomic regions, whilst other members are conserved across strains, therefore implying different functional roles [12].

Despite the function of PE and PPE proteins being poorly understood, some have been shown to be involved in host-pathogen interactions and immune evasion (*e.g.*, PPE34 or PE_PGRS11) [13, 14], while others have enzymatic activity, such as hydrolase (*e.g.*, LipY) [15]. The role of PE_PGRS in host-pathogen interaction varies, from triggering autophagy (PE_PGRS29) [16] to preventing phagosome maturation enhancing survival (PE_PGRS30) [17]. PE and PPE proteins have been demonstrated to be highly immunogenic, and therefore promising targets for vaccine and diagnostic development [18]. The apparent polymorphic and repetitive nature of *pe* and *ppe* genes could be a source of antigenic variation and consequent immune evasion [7, 19, 20]. In contrast, T-cell epitopes in *pe_pgrs* genes binding to HLA-I and -II molecules have been found to be highly conserved and mainly located in PE domains [21]. Despite a significant degree of conservation in some *pe/ppe* genes [21], a number of hot spots of polymorphisms and recombination have been observed [12, 19, 22–24], and overall, diversity seems to be higher in *pe/ppe* loci than in the rest of the genome. Although, the *pe_pgrs* sub-family is considered largely polymorphic, their PE domains are thought to share a higher homology than those of other *pe* genes, which implies important functional consequences [25]. Genetic diversity varies across different *pe_pgrs* genes, suggestive of non-redundant functionality [21]. Some *pe/ppe* genes have been suggested to be under positive selection

[12, 26], whilst other studies have found adaptive or diversifying selection only on the unique C-terminal domains of *pe_pgrs* genes, with the remaining gene under purifying selection [21, 24], supporting the lack of T-cell mediated immune selection of these proteins.

The subcellular localization of PE and PPE proteins require them to be secreted by the ESX system [27]. The type VII secretion system ESX-5 is involved, and PPE38 appears essential for the secretion of PE_PGRS and PPE_MPTR proteins and therefore linked to virulence [28]. The *ppe38* locus is duplicated in ancient strains (named *ppe71*), with this duplication typically lost in modern strains such as H37Rv (RvD7 deletion) [29], and resultantly linked to increased virulence [28]. Notably, Beijing strains harbour a unique disrupted copy of *ppe38* associated with a hypervirulent phenotype, and restoration of its PPE38-dependent secretion partially reverts this phenotype [28]. These insights show how strain specific structural variants that can cause gene rearrangements may affect the pathogenesis and virulence of different strains types. More generally, there is a need to fully characterise the genetic diversity across different strain types, to provide a better understanding of their role in pathogenesis, but also immune evasion and complement immunogenic assays and evaluations of vaccine candidates. However, *pe/ppe* gene families have been systematically excluded from analyses due to the difficulties in reliable aligning sequences to the high GC repetitive regions [12, 27].

Although the availability of high throughput short sequencing technologies has revolutionised the study of *M. tuberculosis* genetic diversity, a high number of coverage blind spots in short read sequencing occurs in *pe* and *ppe* genes [30], due to difficulties in successfully mapping high GC regions. This limitation can be overcome by long read sequencing technologies, such as the PacBio and Oxford Nanopore platforms [31]. In an attempt to characterise these elusive

genes and genetic variants, we have performed an *in silico* analysis of the 169 *pe/ppe* gene sequences across 72 *M. tuberculosis* strains with either (near-)complete assembled genomes, representing six different lineages. We have identified lineage specific markers among the conserved genes, as well as lineage patterns that are responsible for disrupted protein sequences, likely to have functional consequences. These include gene rearrangements, such as duplications or gene fusions that have not been previously reported. Overall, using long sequence data, we provide the first comprehensive analysis of the genetic diversity among the *pe/ppe* families, to assist the development of infection control tools for high burden TB.

RESULTS

The samples

A total of 72 strains with complete genomes (n=37) or with PacBio long-read sequencing data (n=35) [32, 33] were included in the analysis (see **Table S1** for ENA accession numbers, **Additional File 1**). These strains represented 6 different lineages of *M. tuberculosis*, including ancient (n: L1 11, L5 2, L6 7), modern (n: L2 20, L3 5, L4 27 including H37Rv and H37Ra) and one from *M. bovis* BCG (see **Table S1**, **Additional File 1**). They consisted of ten newly sequenced clinical isolates sourced from TB patients in Karonga (Malawi) between 2001 and 2009 (n: L2 2, L3 3, L4 5). *De novo* assembly was performed on the 35 strains with PacBio long reads, base correcting with Illumina data and leading to high quality assemblies (all with number of contigs \leq 8). The maximum SNP distance differences by lineage were $>$ 350 SNPs, ensuring there was genetic diversity amongst strains.

Genome-wide SNP nucleotide and indel diversity

All genomes were aligned to the reference H37Rv, and a total of 19,125 biallelic and polyallelic sites and 6,594 insertions or deletions (indels) were identified genome-wide across the 72 strains. Differences in per SNP or indel nucleotide diversity (π) and absolute divergence (d_{xy}) between the ancient and modern strains were observed in genomic regions containing *pe/ppe* genes (**Figure 1A**). There were four broad regions with high SNP or indel diversity ($\pi > 0.0009$), including loci: (i) *ppe1*; (ii) *ppe3*, *pe_pgrs3*, *pe_pgrs4*; (iii) *pe_pgrs9*, *pe_pgrs10*; (iv) *pe_pgrs50*, *pe_pgrs53-pe_pgrs57*, *ppe55*, *ppe57-ppe59*. There were overlapping SNP and indel diversity peaks, which is consistent with the idea of hot spots of polymorphisms being correlated with deletions [34]. The *ppe1* locus had a greater diversity in modern strains, with potential influence in the differentiation between the ancient and modern lineages ($d_{xy} > 0.001$). The region consisting of *ppe3*, *pe_pgrs3* and *pe_pgrs4* had a high SNP and indel nucleotide diversity, with the *pe_pgrs* genes being highly homologous and a potential recombination hotspot [12]. The region containing *pe_pgrs9* and *pe_pgrs10* had high SNP nucleotide diversity, especially in ancient strains, which also contributed to a high divergence between lineages ($d_{xy} > 0.002$). Finally, the largest region (H37Rv: 3.5 – 4.0 Mbp) had two peaks in SNP nucleotide diversity (*ppe55* and *pe_pgrs50*, and *ppe57* to *ppe59*) and one in indel diversity (*pe_pgrs53* to *pe_pgrs 57*). The *ppe57* and *ppe59* loci have been previously described as highly diverse [12]. The five *pe_pgrs* genes (*pe_pgrs53* to *pe_pgrs57*) harbour indels that differentiate between the ancient and the modern strains, suggesting that lineage-specific structural patterns might be found in these *loci*.

Overall, a higher mean diversity across the whole-genome was obtained among ancient strains (SNP $\pi = 0.000396$; indel $\pi = 0.00009$) than within modern strains (SNP $\pi = 0.000231$;

indel $\pi = 0.00006$; $P < 0.01$), in spite of L4 having a high value of indel π (see **Figure S1, Additional File 1**). Using SNPs only, there was significantly higher diversity in *pe/ppc* genes compared to other functional gene groups such as “cell wall and cell processes”, “lipid metabolism” or “information pathways” (adjusted $P < 0.01$) (**Table 1**). Likewise, *pe/ppc* genes showed significantly higher indel diversity than other gene groups except “insertion sequences” and the “unknown” categories. Similarly, across both SNPs and indels, sequence divergence (*dxy*) between the ancient and the modern strains was significantly higher in the *pe/ppc* gene family compared to other functional groups (adjusted $P < 0.01$) (**Table 1**), suggesting its genetic diversity contributes to lineage differentiation. Maximum-likelihood phylogenetic trees constructed using the genome-wide SNPs and indels resulted in the expected clustering of lineages (**Figure 1B and 1C**).

The *pe* and *ppc* gene family conservation and disruption

For each strain, the level of disruption caused by variants, relative to each H37Rv reference annotated gene, was assigned (0 = no variants or synonymous SNPs; 1 = non-synonymous SNPs; 2 = in-frame indels; and 3 = frameshifts/premature or delayed stop codons). To avoid bias due to potential high recombination, all gene sequences were aligned including flanking regions with non-*pe/ppc* sequences (see **Methods**). The number of truncated or absent *pe/ppc* genes per strain (level 3) varied from 4 in L4.9 to ≥ 30 in some L5, L6 or *M. bovis* BCG isolates. The number of *pe/ppc* genes with complete conserved protein sequences (level 0) per strain was on average 109 for L4, decreasing to 60 for most distant strains on the phylogenetic tree (**Figure 2**). Overall, strains had $> 55\%$ of their *pe/ppc* genes relatively conserved, only harbouring non-synonymous SNPs at most (level 1; median 118, range 93-163).

The 169 *pe/ppe* genes were classified into 3 different classes based on the presence or absence of structural variants, namely those that are: (i) conserved (C) (79/169; 27 *pe*, 20 *pe_pgrs* and 32 *ppe*), (ii) structurally non-conserved (S) (85/169; 9 *pe*, 40 *pe_pgrs* and 36 *ppe*), and (iii) with a unique *k-mer* profile (K) (5/169; 4 *pe_pgrs* and 1 *ppe*) (see **Methods** and **Figure S2** for pipeline; **Table S2** and **S3, Additional File 1**). The conserved genes (class C) did not have major structural variants, and included *ppe7* and *ppe9* where all sequences differed from the H37Rv annotated reference (including H37Rv and H37Ra PacBio genomes). The unique *k-mer* profile genes (Class K) had a high density of SNPs or in-frame indels (sizes: 100 bp to 1000 bp).

To support the classification of the genes into the three classes (C, S, K), we analysed short read sequencing data from ~30k isolates [5]. After alignment to the H37Rv reference, coverage per gene for each sample was obtained and normalised using four housekeeping genes (*gyrA*, *gyrB*, *rpoB*, *rpoC*). Mean normalised coverage of the *pe/ppe* genes (0.74) was found to be lower than the rest of the genome (0.93; adjusted $P < 0.01$). There was the expected depletion in coverage in repetitive regions, but not all *pe/ppe* genes fell in coverage blind spots. Both class K and S *pe/ppe* gene groups, here considered together, had lower mean coverage (0.67), because of their repetitive regions, compared to class C (0.82) or the rest of the genome (0.93; adjusted $P < 0.001$) (see **Figure S3A, Additional File 1**). The mean coverage of class C genes was also lower than non-*pe/ppe* genes (adjusted $P < 0.001$). Seventy of the hundred genes with the lowest mean coverage belonged to the *pe/ppe* families. Mapping these values per gene genome-wide revealed peaks of low coverage (see **Figure S3B, Additional File 1**), which coincided with regions of high SNP and indel diversity found earlier. Moreover, the 20 genes with lowest coverage had been classified into the two non-conserved

categories (class K, S), highlighting difficulties in robustly characterising their variants using a short-read alignment approach.

Diversity in *pe/ppc* genes

SNP nucleotide and indel diversity were calculated for each of the 169 *pe/ppc* gene sequence alignments previously obtained (see **Tables S2 and S3, Additional File 1**). As expected, indel diversity in genes with structural variants (class S; indel $\pi = 0.000585$) was significantly higher than in conserved genes (Class C; indel $\pi = 0.000083$; $P < 0.001$) (**Figure 3A**). However, there were no significant differences between classes in terms of SNP nucleotide diversity (T-test $P > 0.1$). SNP π was heterogeneous among the conserved and the structurally non-conserved genes (range: 0 to > 0.002). The genes with unique *k-mer* profiles (class K; $n=5$) had higher SNP diversity (mean SNP $\pi = 0.002938$) compared to other classes (mean SNP $\pi < 0.0007$) and higher indel diversity (mean indel $\pi = 0.000569$) than class C (mean indel $\pi = 0.000083$), but slightly lower than class S (mean indel $\pi = 0.000585$). A weak correlation between SNP and indel diversity at a gene level was found (Spearman's $\rho = 0.0416$; see **Figure S4, Additional File 1**).

Overall, the *pe_pgrs* subfamily accounted for the majority of the indel diversity compared to *pe* or *ppc* genes (**Figure 3B**), but diversity in the individual genes varies significantly (range π : from < 0.00002 to > 0.002). Interestingly, among *ppc* gene subfamilies, *ppc-svp* (subfamily IV) genes showed higher values of SNP and indel diversity than *ppc_mptr* (subfamily V) (see **Figure S4, Additional File 1**). In accordance with the rest of the genome, *pe/ppc* genes in ancient strains had a higher SNP diversity than modern strains (ancient $\pi = 0.00067$; modern $\pi = 0.00042$). Intra-lineage diversity was calculated for lineages L1, L2, L3, L4 and L6. A total

of 34 and 32 genes had zero SNP or indel diversity respectively in at least four of the five lineages studied, suggesting π values for these genes were driven by inter-lineage diversity. This was the case of some highly conserved genes like *pe10*, *pe23* or *pe_pgrs40*, with low numbers of SNPs or indels that occur in the whole lineage or various lineages. The dN/dS ratios were investigated in individual genes, finding 19, 16 and 19 genes under diversifying selection ($dN/dS > 1.5$; genome-wide average 0.71) in *pe*, *pe_pgrs* and *ppe* genes respectively (see **Tables S2 and S3, Additional File 1**). Despite showing selection pressure, thirty of these genes belonged to the conserved category, as they did not harbour any structural variants. Genome-wide, only the “insertion sequences” functional group showed a dN/dS ratio > 1 suggesting positive selection.

To assess whether the 169 *pe/ppe* genes were evenly diverse along the whole coding region or not, diversity in the different domains was investigated. In PE and PPE domains, which are found at the beginning of the gene, there was low indel diversity (**Figure 3C**), suggesting a certain structural conservation. In addition, these PE and PPE domains showed a higher SNP nucleotide diversity than indel diversity (adjusted $P < 0.01$) except in the *pe_pgrs* subfamily. Within the *pe* family there was a significant higher indel diversity after the PE domain (adjusted $P < 0.01$), which was driven by *pe_pgrs* genes. No significant differences in nucleotide diversity were found between the PE domain and the rest of the gene. In summary, *pe_pgrs* carried the majority of indels after the conserved PE domain, whilst diversity in *ppe* genes and the rest of the *pe* family was resulting predominantly from SNPs.

Large insertions

Structurally non-conserved genes (class S) had a high abundance of large insertions. The *Mycobacterium tuberculosis* complex-specific insertion sequence IS6110 is known to have been integrated into some members of the *pe* and *ppe* gene families, especially among the *ppe_mptr* genes [24, 29, 35]. Through whole genome analysis, we observed integration of IS6110 in regions around *pe/ppe* genes, which were similar across the different lineages (see **Figure S5, Additional File 1**). Thirteen genes (1 *pe* and 12 *ppe*, including 9 *ppe_mptr*, 0 *pe_pgrrs*) were found to harbour IS6110 in at least one isolate (see **Table S4, Additional File 1**). The IS6110 sequence was in most cases responsible for a shift in the reading frame, however, in some samples it was found in-frame. Nevertheless, in both cases, IS6110 was identified as causing premature stop codons and the consequent disruption of the protein sequence. Lineage L4.5 (n=3) and L2.2.1 (n=14/19) isolates harboured IS6110 in *ppe55* and *ppe16* respectively, leading to a truncated protein with a reduced number of MPTR repeats. The *ppe16* and *ppe34* loci are known to have IS6110 insertions [36]. Thirty-four of analysed samples (all L2/3 included) had IS6110 inserted in *ppe34*, disrupting the gene. The inserted IS6110 led to two shorter open reading frames of *ppe34* that were also annotated with PGAP, truncated at the N-terminal and at the C-terminal respectively when compared to H37Rv-PPE34 (see **Figure S6, Additional File 1**). Isolates from L1.1.3 and L3.1.1 were missing the SVP domain in *ppe49* due to the premature stop codon caused by IS6110, which, in summary, showed that some structural variation could be attributed to IS6110.

The *ppe38* genomic region as annotated in the H37Rv reference is rarely found in clinical isolates [29], but often encounters a duplication of *ppe38* (called *ppe71*) which together with *ppe38* flank two *esx* genes (*esxX* (*mt2419*) and *esxY* (*mt2420*)) [28]. We observed the two *esx*

genes and *ppe71* in a high proportion of isolates (n=38/72; 52.8%), including the laboratory strains H37Rv and H37Ra. However, no clear lineage patterns could be identified (see **Figure S7, Additional File 1**). For instance, in every lineage except L5 we found the presence of *ppe71* in at least one sample. Moreover, single isolates from L2 and L4 harboured a second duplication of *esxX/Y/ppe71*. Nevertheless, all Beijing (L2.1.1) isolates had only a single copy, which furthermore, was truncated by the insertion of *IS6110*, which is demonstrated to suppress the secretion of PE_PGRS and PPE_MPTR proteins [28]. We observed that downstream the *IS6110*, which is inserted at the N-terminal of *ppe38*, there is an open reading frame which translates into a homologue of PPE38, however, missing the PPE domain. The lack of a PPE domain in *ppe38* was also found in sporadic samples of other lineages. The contiguous gene, *ppe39*, was found in a different configuration in all isolates except the laboratory strains and lineages L4.6 to L4.9. Most isolates had an extra ~268 residues at the N-terminal which included a PPE domain that is not found in the reference H37Rv, but previously described in Beijing isolates [37]. The longer version of the PPE39 protein shares a high similarity with PPE40 (77% identity), including identical N-terminal sequences. In H37Rv and closely related isolates, PPE39 was truncated by *IS6110* integration leading to the short, annotated version of PPE39 without the PPE domain. Overall, the region between *ppe38* and *ppe40* is a hot spot for the insertion of *IS6110*, more frequently integrated among modern strains (modern: n=29/52; ancient: n=2/20). This locus also corresponds to the RD5, deleted in *M. bovis* BCG [38].

To understand the genetic context of every assembled insertion identified in *pe/pppe* genes across the strains, their sequences were mapped against the H37Rv reference genome. In total, half of the unique insertions > 25 bp identified (264 in thirty genes) mapped with > 70%

identity to a *pe* or *ppe* gene. Twenty-seven genes had 218 insertions that mapped elsewhere in the same gene; while 60 insertions (12 genes) matched in 11 different *pe/pppe* loci. Most of these multi-matched insertions were found in *pe_pgrs* and *ppe_mptr* genes, which contain repetitive regions. The insertions were identified mainly in the MPTR or PGRS domains, in several cases as in-frame insertions of the repetitive regions, and followed a lineage or strain specific pattern. Other insertions were in similar regions adjacent or close to *pe/pppe* genes, (e.g., *ppe54*, *ppe55* and *ppe56*), which could result from homologous recombination. In a few cases, these insertions were inversions of small regions of the gene itself, leading sometimes to stop codons and disrupted protein sequences. Finally, some insertions were gene duplications, like *ppe38/pppe71* presented above, which are identical genes (~100% sequence identity). For *ppe53*, all isolates except lineage L4.3 to L4.9 had an extra copy with the same N-terminal but different C-terminal domain (see **Figure S8, Additional File 1**). This extra copy of *ppe53* shared a 77% protein sequence identity with the H37Rv annotated *ppe53*.

Complex gene reorganisation

Genes classified as structurally non-conserved (class S) harboured different variants that disrupted the protein sequence, among them, changes in the open reading frames caused by big deletions or frameshifts. We found 10 pairs of *pe/pppe* genes that showed potential gene fusions compared to the H37Rv reference, including the fusion of the PE and PGRS domains of adjacent genes. The *pe_pgrs4/3* (L2) and *pe_pgrs20/19* (L1) loci are two examples of fusion of domains in single lineages due to a deletion. A large deletion covering the end of *pe_pgrs4* and beginning of *pe_pgrs3* was identified in all L2 and one L3 isolates. The merging of the remaining sequences of these two adjacent genes for those samples revealed a *pe_pgrs* gene with the PE domain from *pe_pgrs4* and the PGRS domain from *pe_pgrs3*, suggesting a

potential event of gene fusion in these strains (**Figure 4A**). Using AlphaFold prediction analysis, the protein structure of the PE_PGRS4/3 rearrangement in L2 revealed a *pe_pgrs* gene highly similar to *pe_pgrs3* and *pe_pgrs4* (**Figure 4B**). Likewise, the deletion in L1 isolates leads to the formation of *pe_pgrs20/19* fusion, which translates into a protein with the PE and PGRS domains of the constituent proteins.

In other situations, the open reading frame continued until the end of the adjacent gene because of a frameshift caused by a small indel. This situation was found in every lineage in *ppe6/5* except laboratory strains and L1.1.3 (due to a frameshift), in ancient lineages in *ppe8/7* and *pe_pgrs12/13*, in most lineages (except those closest to L4.9) in *pe_pgrs50/49* and *pe_pgrs55/56*, and in L1 and L5 in *ppe67/66* (**Figure 4A**). For example, *pe_pgrs55* and *pe_pgrs56* loci are found in a region of high SNP nucleotide and indel diversity, and most isolates had a *pe_pgrs55* gene that lacked the stop codon caused by a 1 bp deletion, continuing the reading frame until the end of *pe_pgrs56*, hence creating a unique protein sequence. Interestingly, both *pe_pgrs12* and *pe_pgrs55* have a PE domain, whilst in the downstream genes *pe_pgrs13* and *pe_pgrs56* this domain is absent, only showing PGRS motifs, and therefore the combination of them leads to a normal PE_PGRS-like structure inferred by AlphaFold software (**Figure 4C**). For *ppe8/7*, the *ppe7* locus does not have any PPE domain, thereby the gene fusion leads to a *ppe_mptr*-like structure. Similarly, for *ppe6/5*, where *ppe5* lacks the PPE domain, it adds MPTR motifs that form an enlarged *ppe6/5* gene. Finally, there are 4 *pe/ppe* genes in *M. tuberculosis* annotated as pseudogenes, located in 2 operons, where also small indels causing frameshifts led to a change in the open reading frame and the consequent formation of a single gene (**Figure 4A**). The *pe21* locus contains a PE domain, whilst *pe_pgrs36* harbours only the characteristic PGRS repetitive sequences. The

lack of stop codon in *pe21* brings its 3'-end into *pe_pgrs36*. When translated, the H37Rv joint *pe21/pe_pgrs36* sequence seems to create a truncated protein. Notwithstanding, a 1 bp insertion at the beginning of *pe_pgrs36*, present in all samples except L4 excluding L4.4, changes the reading frame relative to the reference genome. This change produces a PE_PGRS-like protein sequence, as determined by AlphaFold. A similar situation is found in *ppe48/ppe47*, where a 1 bp frameshift at the beginning of *ppe47* in all isolates generates a different structure than the one annotated for H37Rv. Overall, 3 out of 4 *pe_pgrs* and 2 out of 3 *ppe_mptr* genes that are annotated without a PE or PPE domain were found to be the continuation of the gene upstream in at least in one lineage. All these gene fusions or rearrangements were confirmed by PGAP annotation.

Duplication of *pe_pgrs3* gene

The *pe_pgrs3* locus is a potential recombination hotspot and several large indels have been identified when aligned to the H37Rv reference, including insertions linked to duplication of repetitive regions. Using the alignments, premature stop codons were identified, and the non-conserved nature of the gene confirmed. Surprisingly, the protein sequences obtained from the aligned region showed a duplication of *pe_pgrs3* in almost every sample analysed (**Figure 4A**). This gene duplication was confirmed by the annotation of the assemblies obtained by PGAP. The two *pe_pgrs3* genes identified are highly similar to the annotated *pe_pgrs3*, with the main differences being the presence/absence of the C-terminal domain from H37Rv-*pe_pgrs3*. There were some differences between lineages, including the absence of the C-terminal domain in the two *pe_pgrs3* genes in L5, L6 and *M. bovis* BCG, or the gene fusion between *pe_pgrs4* and *pe_pgrs3* in L2. This *pe_pgrs4/3* gene was followed in three L2 isolates by a truncated copy of *pe_pgrs4*, before a second copy of a *pe_pgrs3*-like gene.

In summary, only the two lab strains analysed (H37Rv and H37Ra) together with one L4.6 isolate showed the same arrangement than the reference. Despite this lack of concordance with the reference, we observed a significant degree of conservation within lineage. The *pe_pgrs3* gene is duplicated in *M. bovis* and *M. canetti* and until now it was believed not to be duplicated in *M. tuberculosis* [25]. However, we have seen how across different lineages of clinical isolates this gene is duplicated. The two copies of *pe_pgrs3* are slightly different, and also differ from the H37Rv-*pe_pgrs3*.

Conservation across the *pe* and *ppe* sub-families

Both *pe* and *ppe* families have been classified into 5 different subfamilies (named from I to V) based on the phylogenetic analysis of their protein sequences [6]. The *pe35* (*Rv3872*) and *ppe68* (*Rv3873*) loci are found in an operon in the region of difference RD1 (deleted in *M. bovis* BCG), and considered to be the most ancestral *pe* and *ppe* genes respectively (subfamily I for each family), located in the ESAT-6 gene cluster region 1, also present in *M. smegmatis* [6]. The *pe35* gene was found structurally non-conserved as the L5 isolates harboured a 1 bp deletion at the beginning of the gene leading to a premature stop codon, truncating the protein sequence which then lacks the PE domain. Analysis of the ~30k isolates database found that 98% of L5 and 4% of the other lineages harboured that deletion. All isolates except L4.9 are 1 amino acid shorter in sequence due to a SNP leading to a stop codon, however, unlikely to have functional effects.

The pe family

Subfamilies II and III of *pe* genes are formed by two and three genes respectively, all conserved across the different lineages (see **Figure S9A, Additional File 1**). Subfamily IV was

also mostly conserved across the different samples, however, *pe18* and *pe31* were in class S due to deletions or premature stop codons in sporadic samples. Moreover, *pe32* was also classified as non-conserved as it belongs to the RD8, deleted in *M. bovis* BCG and L6 [38, 39]; however, across the other lineages it remained with a 100% identical protein sequence (**Figure 2**). Subfamily V is formed mainly by all *pe_pgrs* and 19 other *pe* genes, 41% of them being structurally conserved, including two genes (*pe9* and *pe_pgrs40*) with a 100% protein sequence identity across the 72 samples. Overall, most of the structural diversity in *pe* family was found in *pe_pgrs* genes. The differences in protein lengths were mainly driven by deletions and were more common among subfamily V (see **Figure S9B, Additional File 1**). The *pe9* and *pe10* genes belong to subfamily V, and have been demonstrated to form a heterodimer which induces macrophage apoptosis through Toll-like receptor TLR4 interaction [40]. PE10 includes a carbohydrate-binding domain (CBM2, PF00553.21) at the C-terminal; however, using *Pfam* this CMB-2 domain was only identified in L2 and L3 isolates, which have a 1 bp deletion (fixed allele frequency in L2/L3) close to the C-terminal creating a frameshift that leads to a change in the last residues and an additional 27 amino acids.

The pe_pgrs genes

The *pe_pgrs* genes have been traditionally considered highly polymorphic. However, it has been observed that the hydrophilic/hydrophobic profile of the PE domain within the *pe_pgrs* subfamily is more conserved than within the other *pe* genes [25]. Across our samples we also found a higher identity between the protein sequences of the PE domains belonging to PE_PGRS proteins (60%) compared to those from other PE proteins (41%). Moreover, the *dN/dS* ratio in *pe_pgrs* was 0.57 compared to 1.20 in the rest of *pe* genes, suggestive of negative selection. Twenty *pe_pgrs* genes were classified as conserved (Class C) across the

isolates using our pipeline. The *pe_pgrs40* locus was the only gene that had the protein sequence completely conserved across all samples, and *pe_pgrs39* had only non-synonymous SNPs in a small number of samples. Other conserved *pe_pgrs* had at least in-frame indels in one sequence, however the protein sequence was conserved overall.

Among the class S *pe_pgrs* genes, some of the deleted loci correspond to known regions of difference, such as RD701 in specific *M. africanum* isolates, which involves the deletion of *pe_pgrs2* [41] in our L6 samples. As shown, *pe_pgrs3* appears duplicated in most lineages. Ancestral *M. canetti* shows the same structure [25], suggesting that laboratory strains like H37Rv and H37Ra have lost one of the copies of *pe_pgrs3* retaining a unique gene which combines N-terminal and C-terminal from the ancestral 2 copies. Thus, the structure found in most lineages is similar to that in *M. bovis pe_pgrs3* and *pe_pgrs3a*. However, L1 to L4 differ from *M. bovis* as they harbour the H37Rv-*pe_pgrs3* C-terminal domain in one of the *pe_pgrs3* copies (see **Figure S10, Additional File 1**). The *pe_pgrs28* gene also showed a pattern of differences between the clinical isolates and the laboratory strains (except for the L4.6 isolate, which matched with the laboratory strains). On the other hand, *pe_pgrs35* and *pe_pgrs47* had conserved sequences, although they were classified as class S due to missing samples. Genes within the unique *k-mer* class (K), despite showing diverse sequences at a nucleotide level, kept conserved protein sequences (> 97% protein sequence identity). Both *pe_pgrs17* and *pe_pgrs18* were included in this category. These two genes are highly similar and likely to be the result of a duplication event. The *pe_pgrs17* locus harbours a polymorphism termed 12/40 [42], which consists of an insertion of 12 bp followed by 40 SNPs. In our analysis, we considered the 12/40 polymorphism as a single indel event, and in line with previous works [42], it was found in all isolates except laboratory strains in *pe_pgrs17*, and in L4.1 sub-

lineages in *pe_pgrs18*. In summary, lineage or sub-lineage patterns of disruption at a protein level could be identified (**Figure 2**). However, in many cases the classification of a gene as non-conserved was due to sporadic mutations in single isolates, with others having gene rearrangements. However, *pe_pgrs* were more conserved at a protein sequence level compared to nucleotide level.

The ppe family

In contrast to the *pe* family, genes of *ppe* sub-families II and III harboured disruptive variants (see **Figure S9C, Additional File 1**). The *ppe* sub-family II is characterised by genes with the PPW domain. It is formed by 12 genes from which 7 were classified as conserved. Among the non-conserved genes, we found loci disrupted by frameshifts (*ppe37*), by IS6110 insertions (*ppe46*), deletions (*ppe66*) and gene fusions as previously seen (*ppe67/66* and pseudogenes *ppe48/47*). Interestingly, *ppe67* and *ppe48* do not have a PPW domain, however, in lineages where its open reading frame continues into the downstream gene they formed a PPE-PPW protein sequence. Sub-family III is formed by 6 genes with a variable C-terminal domain, and only 2 of them were conserved. The others were mainly disrupted due to deletions. Twenty-six genes formed sub-family IV, characterised by the C-terminal SVP domain, which was identified in all genes, including *ppe9*. In H37Rv, *ppe9* is annotated as a truncated gene without the SVP domain, however, all our samples including H37Rv and H37Ra showed a longer sequence with SVP domain. Similarly, *ppe50* in H37Rv, L3 and L4 do not have an SVP domain either. Nevertheless, L2, L5, L6 and *M. bovis* BCG carried an insertion with this domain. Thirteen of these genes were in class S, showing lineage specific patterns in some cases (**Figure 2**). For instance, *ppe65* belongs to the RD8 deleted in L6 and *M. bovis*, or *ppe43* and *ppe45* were truncated by a frameshift and a non-synonymous change in L5 and L6

respectively. PPE38 has been suggested to be involved in secretion of other PPE_MPTR and PE_PGRS proteins [28]. The disruption of *ppe38* would therefore cancel secretion of these immunological proteins, thereby enhancing immune evasion and persistence [11]. This gene is truncated in some strains (*e.g.*, Beijing strains) and it is located in regions of difference, such as RD5, deleted in *M. bovis* BCG [38]. Whilst *ppe38* was > 50% deleted in L2, it was reasonably well conserved across the other lineages.

The ppe_mptr genes

The *ppe* sub-family V has 24 loci, which are the *ppe_mptr* genes. They are the most polymorphic of *ppe* genes and thereby they account for the highest level of disruption in the protein sequence, with 16 of them non-conserved. These *ppe_mptr* genes generally have a PPE domain followed by different numbers of a pentapeptide repeat. This MPTR domain is found between 2- and 48-times (*e.g.*, *ppe8*) in *ppe_mptr* genes. Generally, the largest variation in gene length was found among members of *ppe_mptr* sub-family (see **Figure S9D**, **Additional File 1**). The *ppe55* and *ppe56* loci are in a peak of high nucleotide diversity in the genome. They both belong to the RD^{Rio} [43], deleted in L4.3.4 isolates, and harbour various variants creating truncated protein sequences in different lineages, including IS6110 insertions. As shown, *ppe_mptr* genes were the most common locations of insertion of IS6110, responsible for disrupting these proteins. In total, nine *ppe_mptr* genes had IS6110 insertions in at least one sample, including two of the conserved genes.

Lineage specific SNPs and indels in *pe/ppe* genes

A total of 3,571 SNPs and 1,247 indels were identified among the *pe* and *ppe* genes, from which 459 SNPs and 122 indels were found in the structurally conserved genes. Moreover,

seven of the conserved genes (*ppe7*, *pe9*, *pe13*, *pe19*, *pe22*, *pe25* and *pe_pgrrs40*) did not harbour any non-synonymous SNP or indel, having a completely conserved protein sequence across all the isolates. Nevertheless, one of them was *ppe7*, which as mentioned previously, even though is conserved across the samples analysed, the sequence differed from the annotated H37Rv gene in 1 bp insertion. The existence of inter- without intra-lineage diversity in some genes suggested a potential lineage specific pattern in *pe/pppe* genes. We performed a principal component analysis with SNP and indel matrices for each sample. Clustering by lineage was clear for indels, and with sub-groups for some lineages being observed using SNPs (see **Figure S11, Additional File 1**). Following the hypothesis of *pe* and *pppe* genes also showing a lineage and sub-lineage specific pattern, we built three maximum likelihood phylogenetic trees with only these SNPs, indels, and both (see **Figure S12, Additional File 1**). Sixteen genes where > 1.5% of its coding region were polymorphic sites or with a unique *k-mer* profile due to SNPs were removed for the reconstruction of the SNPs tree (1,946 SNPs discarded). The topologies of the trees with SNPs and indels were different; however, both showed a clear clustering by lineage, suggesting lineage specific patterns.

With the purpose of identifying these lineage specific polymorphisms, the fixation index (F_{ST}) was calculated comparing one lineage against the others for each of the variants found in *pe* and *pppe* genes across the 72 available genomes. Overall, 83 SNPs and 8 indels were identified with a F_{ST} of 1 (perfect differentiation) in one lineage within our dataset and with an allele frequency > 0.95 in the corresponding lineage within the ~30k isolate dataset (**Table 2**). Variants present in the ancient clade were also tested. Nine SNPs and four indels with F_{ST} of 1 in ancient strains and an allele frequency > 0.75 in ancient samples from the ~30k isolate database were identified (**Table 2**). In addition, two indels in L4.1 and L2/3 respectively were

identified only in those lineages. Among lineage specific variants, we found 2 SNPs leading to premature stop codons (*ppe10* W8* and *ppe45* W75*); eight frameshifts leading to disrupted proteins (*pe_pgrs6* 1557_1558insT, *pe_pgrs16* 1968_1969insG, *ppe16* 1279_1283del, *ppe43* 449_454del, *ppe56* 6586_6586del, *pe_pgrs55* 1411_1411del, *pe_pgrs56* 991_1086del and *ppe64* 63_64del) and two other frameshifts leading to longer protein sequences (*ppe8* 9889_9890insATA and *pe10* 337_337del).

DISCUSSION

The *pe* and *ppe* genes are important *M. tuberculosis* loci, but are routinely excluded from WGS studies, especially those using short sequence data, due to the difficulty in accurately mapping their repetitive and polymorphic regions [44]. To overcome this problem, we used PacBio assemblies to provide the most comprehensive picture to date of genetic diversity in all 169 *pe* and *ppe* genes. The sequence analysis revealed a large amount of both conservation and diversity in members of these two families. As expected, we observed greater nucleotide diversity in *pe/ppe* genes compared to the rest of the genome, especially in clusters of *pe/ppe* loci (e.g., *pe_pgrs53* to *pe_pgrs75*, *ppe57* to *ppe59*), with some predicted to be pathogenicity islands [45]. The diversity was driven not only by SNPs but also indels. One of the known drivers of diversity in these regions is the integration of *IS6110*, for which several transposition sites have been identified among these genes, especially within members of the *ppe* subfamily V (*ppe_mptr*) [24, 29, 35, 36, 46]. Consistent with previous findings [35], we observed a tendency of occurrence of *IS6110* insertions in genomic regions with *pe/ppe* genes. We identified one *pe* and twelve *ppe* genes disrupted by *IS6110*, with some of these genes exhibiting lineage-specific patterns. For example, *ppe38* represents a hot spot for

IS6110 integration, being truncated by IS6110 (RvD7) in all our Beijing and other sporadic isolates, known to lead to hypervirulence [28, 29]. However, *ppe38* also belongs to RD5, which is deleted in attenuated strains such as *M. bovis* BCG [38]. The contiguous gene, *ppe39*, has also been characterised in Beijing strains with a different sequence to that annotated in the H37Rv reference [37]. We found the complete version of PPE39 in most of our isolates except L4.7 to L4.9, which contained the short, annotated version, truncated by IS6110 in the N-terminal. This shorter H37Rv-*ppe39* does not present a PPE domain.

Evidence of homologous recombination, especially in repetitive regions of *pe/ppe* genes [12, 23], and events of gene conversion [42] have been described. The *ppe38* locus is a hotspot for recombination and indel events, which is highly variable between isolates, not only due to the insertion of IS6110 but also the presence or absence of a second copy of the gene (*ppe71*) [29]. Similar patterns are also observed for other genes such as the *pe_pgrs3/4* locus, which has a different configuration to that found in H37Rv. Homologous recombination due to the repetitive nature of the PGRS domain has been previously suggested to occur in this region [12]. Most of the samples, except laboratory strains, had a second copy of *pe_pgrs3*, leading to a similar arrangement as found in *M. bovis* and *M. canetti*, where 2 copies can be found [25]. Due to the similarity to the ancestral configuration, we suggest recombination events have resulted in the loss of one copy in H37Rv and related strains. Other gene arrangements identified include gene fusions. Some of these were found in single lineages (e.g., *pe_pgrs20/19*), while others were in all samples (e.g., *ppe48/47*). Interestingly, the four *pe/ppe* genes annotated as pseudogenes, organised in two operons in H37Rv, were found to form a single open reading frame in most isolates, leading to a potentially functional protein. This lack of consistency between the H37Rv annotated sequences and the predicted protein

sequences in the clinical isolates could potentially mislead and hinder the capture of variants when using mapping methods.

As expected, overall SNP and indel nucleotide diversity of *pe/pppe* genes was greater than the rest of the genome, but there was high heterogeneity across the genes. The class S genes displayed greater indel diversity, but a similar SNP diversity to class C. This finding is consistent with the lack of correlation between SNP and indel diversity found across the *pe/pppe* genes. Previous analysis has found a heterogeneous diversity profile across 27 *pe_pgrs* genes [21], and interestingly, the PE domains of these genes, where the T-cell epitopes are mostly found, were relatively more conserved than those of other *pe* genes [25]. In fact, the main source of diversity in *pe_pgrs* genes was identified after the PE domain, being mainly driven by indels. In contrast, diversity was more often the result of SNPs in *pe* and *pppe* genes. Despite some *pppe* and *pe_pgrs* genes having been reported to be under selective pressure [12], we found them to be overall under more purifying selection than *pe* genes. Nevertheless, in line with previous work, the *dN/dS* ratios obtained broadly varied across individual genes in both families [21]. The inter-lineage diversity found in some *pe/pppe* genes, together with its substantial impact to the phylogenetic differences between the ancient and the modern strains, suggested the presence of lineage-specific variants in these regions. We identified numerous lineage and clade specific SNPs and indels across the *pe/pppe* genes, which were validated in ~30k *M. tuberculosis* with whole genome sequencing data. Protein disruption was a frequent outcome of the lineage specific indels, which considering the role in host-pathogen interaction of these proteins, could provide insights into different behaviour between strains. One limitation of the use of short read sequencing data for the validation work was the lack of accuracy on detecting big indels, especially among repetitive regions.

All *pe/ppe* genes were classified based on the conservation observed across the 72 isolates. Structural variants, such as frameshifts, changes in start and stop codons and large deletions were responsible of the classification of numerous genes as non-conserved, which often, were identified across one or multiple sub-lineages. Sub-families V, which are the result of the most recent duplication and recombination events [6], were found in higher numbers amongst the non-conserved genes. Importantly, this classification was based on the alignment to the H37Rv sequence, which, as shown, does not always represent the functional locus, as some genes are truncated in the reference (*e.g.*, *ppe39* or *ppe48*). However, on average, more than half of the *pe/ppe* gene members per sample were found to be conserved, suggesting an important role. The various levels of diversity and conservation that different genes display have been proposed to imply non-redundant functions [21]. Nevertheless, the complex gene layout that is found in the different strains, with some genes highly conserved in some lineages whilst disrupted in others, requires more investigation in order to understand the functional consequences of the variation observed. One difficulty is the lack of structural data for PE/PPE proteins that restricts the prediction of functional consequences. However, the use of novel *in silico* tools, such as AlphaFold [47], can be of assistance.

The expansion of *pe/ppe* families in slow growing pathogenic mycobacteria [6, 7], the attenuated phenotype of *M. bovis* BCG associated to RD1 [38] or *ppe25-pe19* knockout mutants [48], all demonstrate the association of *pe/ppe* with virulence. Moreover, one of the most intriguing aspects of PE and PPE proteins is their role in host-pathogen interactions. The study of individual members of these families has revealed different functions. For instance,

PE_PGRS33 is known to induce pro-inflammatory cytokines through TL2 interaction [49], whilst PE31 increases expression of anti-inflammatory cytokines like IL-10 [50]. Thus, they can act as modulators of the immune response driving dormant or multiplying stages [11]. Consequently, multiple epitopes have been characterised on these proteins being investigated as targets for vaccine development [11]. T-cell epitopes are found in the conserved PE domains of *pe_pgrs* genes rather than the variable sequences, supporting the hypothesis that this conservation favours infection [21, 51]. Furthermore, PE and PPE domains are important for the cellular localisation of the proteins [52, 53]. It is plausible then that gene fusions where genes with absent PE/PPE domains are transcribed together with the upstream gene, lead to a likely functional protein. On this premise, understanding of the structural diversity of the *pe/ppe* genes and consequent effect on these proteins is crucial for its potential use in vaccine development, which ideally would target conserved sequences across the different lineages. Additionally, the role of these proteins in cell wall localisation and small molecule transportation means they should be explored as drug targets [27].

CONCLUSIONS

In conclusion, the *pe/ppe* genes represent various levels of diversity and conservation, which moreover show lineage specific profiles and can therefore be phylogenetically informative. Although there is a significant amount of variation in these genes, some are relatively conserved and could be included in whole genome sequencing analysis rather than removed. Moreover, the use of lineage specific reference genomes could assist with more successful alignments for those duplicated genes absent in the H37Rv reference. PE/PPE proteins play important roles in virulence and host-pathogen interaction, and therefore it is important to

elucidate their function to gain a better understanding of the complexity of these two families. Here, we provide the first analysis of genetic diversity across all 169 genes. Future studies in a larger number of isolates should further explore the diversity and conservation across these gene families, and combined with functional characterisation, will lead to insights that can assist with the control of tuberculosis disease.

MATERIALS AND METHODS

Selection of samples, culture and sequencing

A total of 72 PacBio assemblies were used for the analysis. Ten samples were cultured at LSHTM CL3 laboratories and sequenced for this study, being sourced from TB patients in Karonga district (Malawi). Briefly, *M. tuberculosis* clinical isolates derived from patient's sputum were cultured to mid-log phase (optical density = 0.6 - 0.8) in Middlebrook 7H9 supplemented with 0.05% Tween 80 and 10% albumin-dextrose-catalase (ADC) at 37°C in roller bottles. DNA was extracted from passage 2 by heat-inactivation followed by the CTAB-chloroform-isoamyl alcohol method [54]. DNA samples were sequenced with single-molecule real time (SMRT) sequencing technology from Pacific Biosciences (PacBio) RSII through The Applied Genomics Centre at LSHTM. Raw sequencing data from the 10 isolates together with other 27 samples previously sequenced [32, 33] were processed to generate the assemblies using Flye software [55]. These assembled genomes were base corrected using Illumina short reads where possible by using Pilon software [56]. To ensure good quality of the assemblies, only those with a maximum of 8 contigs were included in the analysis. The remaining 35 assembled genomes studied were publicly available and sourced from the ENA (for accession

numbers see **Table S1, Additional File 1**). Lineage and sub-lineage profiling were performed with TB-Profiler [57].

Population genetics analysis

For all the population genetics analysis the H37Rv reference genome (ASM19595v2) was used. Snippy software [58] was used to simulate reads from assemblies and to call variants (SNPs and indels) at a whole genome level. The R packages PopGenome [59] and SeqinR [60] were used for the population genetics analysis. In brief, Nei's π nucleotide diversity per site (SNP π), indel diversity per site (indel π) and absolute divergence (d_{xy}) were calculated in sliding windows throughout the genome for the different populations (ancient and modern strains or by lineage). The average of the three parameters was calculated for the comparison between populations. The dN/dS pairwise ratios were calculated by concatenating the coding regions relative to the reference H37Rv. Statistical differences in diversity and divergence parameters between gene functional groups were calculated using ANOVA, where p-values were corrected by multiple comparisons using Tukey's Honest Significant Differences (HSD) test. FastTree [61] was used for the phylogenetic reconstruction of the samples using SNPs and indels. The NCBI prokaryotic genome annotation pipeline, PGAP [62], was used to annotate the genomes and validate gene rearrangements.

pe/ppe gene extraction, alignment and classification

The *pe* and *ppe* gene alignments were generated using a customized pipeline. In brief, non-*pe/ppe* flanking genes were chosen and mapped against the H37Rv reference genome as anchors for the extracted sequence that were subsequently aligned with MAFFT [63]. Genomes where flanking genes were in different contigs or could not be mapped to the

reference were considered as missing samples. Single *pe/ppa* genes alignments were obtained relative to the H37Rv sequences and curated manually if necessary. SNPs and indels for each individual gene were obtained using the H37Rv reference. Levels of disruption that these variants caused on the protein sequence were assigned (0 = no variants or synonymous SNPs; 1 = non-synonymous SNPs; 2 = in-frame indels; 3 = SNPs or frameshifts leading to changes in start or stop codons, deletions of > 50% of the coding region, missing samples or insertions > 1,000 bp). To investigate whether individual genes were conserved across the different lineages, each *pe/ppa* gene was classified into one of the three categories: conserved (C), structurally non-conserved (S) and unique *k-mer* profile (K) (see **Figure S2, Additional File 1**). Briefly, for each gene alignment, if two or more samples were assigned a value of 3 as described above, the gene was considered as structurally non-conserved. In some genes it was observed that some samples had a high density of SNPs in some regions whilst still maintaining the same sequence length as the reference. Other genes had samples which contained completely novel sequence insertions. In an attempt to characterise the presence of these, DSK software [64] was used to count *k-mers*. For each gene alignment the *k-mer* profile was obtained and those that did not show structural variants as previously described, but had enrichments of unique *k-mers* as a consequence of SNPs or indels, were considered as a different category.

Illumina short-reads coverage

A data set of ~30k short read Illumina samples representing every lineage (L1-L6 and *M. bovis* BCG) was used [5]. Short reads were aligned to the reference with BWA-MEM [65] and the coverage per gene per sample was calculated with BEDTools [66]. The coverage was normalised by four housekeeping genes (*gyrA*, *gyrB*, *rpoB* and *rpoC*) and compared between

pe/*ppe* genes and the rest of the genome. For the comparison between groups, *pe*/*ppe* genes were divided into the previously explained categories, in this case including the “unique *k*-mer” category in “structurally non-conserved” due to small numbers of genes. Statistical differences in the means between categories were assessed using T-tests.

The pe and ppe genes sequence analysis

Population genetics parameters (nucleotide and indel diversity and divergence) for individual genes were calculated using PopGenome R package [59]. The BUSTED method was used for the calculation of *dN/dS* ratios [67]. Identification of known domains was performed with *Pfam* software [68]. T-tests were applied to calculate the statistical differences for nucleotide and indel diversity between the different domains or gene groups. AlphaFold software [47] was used for the prediction of protein structure models. For all variants identified in PE/PPE genes, fixation index (F_{ST}) values to assess allele frequency differences for each lineage were calculated. As validation of variants with F_{ST} values of 1 (perfect differentiation), allele frequencies in a database of ~30k short read Illumina genomes were obtained [5]. For the consideration of lineage specific variants, an allele frequency of 0 in other lineages and > 0.95 in the corresponding lineage was required.

Availability of data and materials

The sequence data supporting the conclusions of this article have been deposited in the ENA (for accession numbers see **Table S1, Additional File 1**).

REFERENCES

1. World Health Organization (WHO). Global Tuberculosis Report 2021. 2021.
2. Sable SB, Posey JE, Scriba TJ. Tuberculosis Vaccine Development: Progress in Clinical Evaluation. *Clin Microbiol Rev.* 2019;33.
3. Coscolla M, Gagneux S. Consequences of genomic diversity in *Mycobacterium tuberculosis*. *Semin Immunol.* 2014;26:431–44.
4. Bottai D, Frigui W, Sayes F, Di Luca M, Spadoni D, Pawlik A, *et al.* TbD1 deletion as a driver of the evolutionary success of modern epidemic *Mycobacterium tuberculosis* lineages. *Nat Commun.* 2020;11.
5. 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 Med.* 2020;12:114.
6. Gey van Pittius NC, Sampson SL, Lee H, Kim Y, van Helden PD, Warren RM. Evolution and expansion of the *Mycobacterium tuberculosis* PE and PPE multigene families and their association with the duplication of the ESAT-6 (*esx*) gene cluster regions. *BMC Evol Biol.* 2006;6:95.
7. Akhter Y, Ehebauer MT, Mukhopadhyay S, Hasnain SE. The PE/PPE multigene family codes for virulence factors and is a possible source of mycobacterial antigenic variation: Perhaps more? *Biochimie.* 2012;94:110–6.
8. McGuire A, Weiner B, Park S, Wapinski I, Raman S, Dolganov G, *et al.* Comparative analysis of *Mycobacterium* and related actinomycetes yields insight into the evolution of *Mycobacterium tuberculosis* pathogenesis. *BMC Genomics.* 2012;13:120.
9. 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.

10. Abdallah AM, Gey van Pittius NC, DiGiuseppe Champion PA, Cox J, Luirink J, Vandenbroucke-Grauls CMJE, *et al.* Type VII secretion — mycobacteria show the way. *Nat Rev Microbiol.* 2007;5:883–91.

11. Medha, Sharma S, Sharma M. Proline-Glutamate/Proline-Proline-Glutamate (PE/PPE) proteins of *Mycobacterium tuberculosis*: The multifaceted immune-modulators. *Acta Trop.* 2021;222 June:106035.

12. 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.

13. Bansal K, Sinha AY, Ghorpade DS, Togarsimalemath SK, Patil SA, Kaveri S V., *et al.* Src Homology 3-interacting Domain of Rv1917c of *Mycobacterium tuberculosis* Induces Selective Maturation of Human Dendritic Cells by Regulating PI3K-MAPK-NF- κ B Signaling and Drives Th2 Immune Responses. *J Biol Chem.* 2010;285:36511–22.

14. Basu J, Shin DM, Jo EK. Mycobacterial signaling through toll-like receptors. *Front Cell Infect Microbiol.* 2012;2 November:145.

15. Mishra KC, De Chastellier C, Narayana Y, Bifani P, Brown AK, Besra GS, *et al.* Functional role of the PE domain and immunogenicity of the *Mycobacterium tuberculosis* triacylglycerol hydrolase LipY. *Infect Immun.* 2008;76:127–40.

16. Chai Q, Wang X, Qiang L, Zhang Y, Ge P, Lu Z, *et al.* A *Mycobacterium tuberculosis* surface protein recruits ubiquitin to trigger host xenophagy. *Nat Commun.* 2019;10:1973.

17. Iantomasi R, Sali M, Cascioferro A, Palucci I, Zumbo A, Soldini S, *et al.* PE_PGRS30 is required for the full virulence of *Mycobacterium tuberculosis*. *Cell Microbiol.* 2012;14:356–

67.

18. Qian J, Chen R, Wang H, Zhang X. Role of the PE/PPE Family in Host–Pathogen Interactions and Prospects for Anti-Tuberculosis Vaccine and Diagnostic Tool Design. *Front Cell Infect Microbiol.* 2020;10:1–8.
19. Talarico S, Zhang L, Marrs CF, Foxman B, Cave MD, Brennan MJ, *et al.* *Mycobacterium tuberculosis* PE_PGRS16 and PE_PGRS26 genetic polymorphism among clinical isolates. *Tuberculosis.* 2008;88:283–94.
20. Tundup S, Pathak N, Ramanadham M, Mukhopadhyay S, Murthy KJR, Ehtesham NZ, *et al.* The Co-Operonic PE25/PPE41 Protein Complex of *Mycobacterium tuberculosis* Elicits Increased Humoral and Cell Mediated Immune Response. *PLoS One.* 2008;3:e3586.
21. Copin R, Coscollá M, Seiffert SN, Bothamley G, Sutherland J, Mbayo G, *et al.* Sequence Diversity in the *pe_pgrs* Genes of *Mycobacterium tuberculosis* Is Independent of Human T Cell Recognition. *MBio.* 2014;5:1–11.
22. Talarico S, Cave MD, Marrs CF, Foxman B, Zhang L, Yang Z. Variation of the *Mycobacterium tuberculosis* PE_PGRS33 Gene among Clinical Isolates. *J Clin Microbiol.* 2005;43:4954–60.
23. Karboul A, Mazza A, Gey van Pittius NC, Ho JL, Brousseau R, Mardassi H. Frequent Homologous Recombination Events in *Mycobacterium tuberculosis* PE/PPE Multigene Families: Potential Role in Antigenic Variability. *J Bacteriol.* 2008;190:7838–46.
24. McEvoy CRE, Cloete R, Müller B, Schürch AC, van Helden PD, Gagneux S, *et al.* Comparative Analysis of *Mycobacterium tuberculosis* *pe* and *ppe* Genes Reveals High Sequence Variation and an Apparent Absence of Selective Constraints. *PLoS One.* 2012;7:e30593.
25. De Maio F, Berisio R, Manganeli R, Delogu G. PE_PGRS proteins of *Mycobacterium tuberculosis*: A specialized molecular task force at the forefront of host–pathogen interaction. *Virulence.* 2020;11:898–915.
26. Zhang Y, Zhang H, Zhou T, Zhong Y, Jin Q. Genes under positive selection in *Mycobacterium*

tuberculosis. Comput Biol Chem. 2011;35:319–22.

27. Ates LS. New insights into the mycobacterial PE and PPE proteins provide a framework for future research. Mol Microbiol. 2020;113:4–21.

28. Ates LS, Dippenaar A, Ummels R, Piersma SR, van der Woude AD, van der Kuij K, *et al*. Mutations in *ppe38* block PE_PGRS secretion and increase virulence of *Mycobacterium tuberculosis*. Nat Microbiol. 2018;3:181–8.

29. McEvoy CR, Van Helden PD, Warren RM, Van Pittius NCG. Evidence for a rapid rate of molecular evolution at the hypervariable and immunogenic *Mycobacterium tuberculosis* PPE38 gene region. BMC Evol Biol. 2009;9:1–21.

30. Modlin SJ, Robinhold C, Morrissey C, Mitchell SN, Ramirez-Busby SM, Shmaya T, *et al*. Exact mapping of Illumina blind spots in the *Mycobacterium tuberculosis* genome reveals platform-wide and workflow-specific biases. Microb Genomics. 2021;7.

31. Elghraoui A, Modlin SJ, Valafar F. SMRT genome assembly corrects reference errors, resolving the genetic basis of virulence in *Mycobacterium tuberculosis*. BMC Genomics. 2017;18:302.

32. Gomez-Gonzalez PJ, Andreu N, Phelan JE, de Sessions PF, Glynn JR, Crampin AC, *et al*. An integrated whole genome analysis of *Mycobacterium tuberculosis* reveals insights into relationship between its genome, transcriptome and methylome. Sci Rep. 2019;9:1–11.

33. Phelan J, De Sessions PF, Tientcheu L, Perdigao J, Machado D, Hasan R, *et al*. Methylation in *Mycobacterium tuberculosis* is lineage specific with associated mutations present globally. Sci Rep. 2018;8:1–7.

34. Tsolaki AG, Hirsh AE, DeRiemer K, Enciso JA, Wong MZ, Hannan M, *et al*. Functional and evolutionary genomics of *Mycobacterium tuberculosis*: Insights from genomic deletions in 100 strains. Proc Natl Acad Sci. 2004;101:4865–70.

35. Reyes A, Sandoval A, Cubillos-Ruiz A, Varley KE, Hernández-Neuta I, Samper S, *et al.* IS-seq: a novel high throughput survey of in vivo IS6110 transposition in multiple *Mycobacterium tuberculosis* genomes. BMC Genomics. 2012;13.
36. Yesilkaya H, Dale JW, Strachan NJC, Forbes KJ. Natural transposon mutagenesis of clinical isolates of *Mycobacterium tuberculosis*: How many genes does a pathogen need? J Bacteriol. 2005;187:6726–32.
37. Han SJ, Song T, Cho YJ, Kim JS, Choi SY, Bang HE, *et al.* Complete genome sequence of *Mycobacterium tuberculosis* K from a Korean high school outbreak, belonging to the Beijing family. Stand Genomic Sci. 2015;10:1–8.
38. Gordon S V., Brosch R, Billault A, Garnier T, Eiglmeier K, Cole ST. Identification of variable regions in the genomes of tubercle bacilli using bacterial artificial chromosome arrays. Mol Microbiol. 1999;32:643–55.
39. Groenheit R, Ghebremichael S, Svensson J, Rabna P, Colombatti R, Riccardi F, *et al.* The Guinea-Bissau Family of *Mycobacterium tuberculosis* Complex Revisited. PLoS One. 2011;6:1–8.
40. Tiwari B, Ramakrishnan UM, Raghunand TR. The *Mycobacterium tuberculosis* protein pair PE9 (*Rv1088*)-PE10 (*Rv1089*) forms heterodimers and induces macrophage apoptosis through Toll-like receptor 4. Cell Microbiol. 2015;17:1653–69.
41. Mostowy S, Onipede A, Gagneux S, Niemann S, Kremer K, Desmond EP, *et al.* Genomic Analysis Distinguishes *Mycobacterium africanum*. J Clin Microbiol. 2004;42:3594–9.
42. Karboul A, Van Pittius NCG, Namouchi A, Vincent V, Sola C, Rastogi N, *et al.* Insights into the evolutionary history of tubercle bacilli as disclosed by genetic rearrangements within a PE_PGRS duplicated gene pair. BMC Evol Biol. 2006;6:1–18.
43. Lazzarini LCO, Huard RC, Boechat NL, Gomes HM, Oelemann MC, Kurepina N, *et al.*

Discovery of a novel *Mycobacterium tuberculosis* lineage that is a major cause of tuberculosis in Rio de Janeiro, Brazil. *J Clin Microbiol.* 2007;45:3891–902.

44. 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. *Nat Rev Microbiol.* 2019;17:533–45.

45. Xie J, Zhou F, Xu G, Mai G, Hu J, Wang G, *et al.* Genome-wide screening of pathogenicity islands in *Mycobacterium tuberculosis* based on the genomic barcode visualization. *Mol Biol Rep.* 2014;41:5883–9.

46. Namouchi A, Mardassi H. A genomic library-based amplification approach (GL-PCR) for the mapping of multiple IS6110 insertion sites and strain differentiation of *Mycobacterium tuberculosis*. *J Microbiol Methods.* 2006;67:202–11.

47. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature.* 2021;596:583–9.

48. Bottai D, Di Luca M, Majlessi L, Frigui W, Simeone R, Sayes F, *et al.* Disruption of the ESX-5 system of *Mycobacterium tuberculosis* causes loss of PPE protein secretion, reduction of cell wall integrity and strong attenuation. *Mol Microbiol.* 2012;83:1195–209.

49. Basu S, Pathak SK, Banerjee A, Pathak S, Bhattacharyya A, Yang Z, *et al.* Execution of Macrophage Apoptosis by PE_PGRS33 of *Mycobacterium tuberculosis* Is Mediated by Toll-like Receptor 2-dependent Release of Tumor Necrosis Factor- α . *J Biol Chem.* 2007;282:1039–50.

50. Ali MK, Zhen G, Nzungize L, Stojkoska A, Duan X, Li C, *et al.* *Mycobacterium tuberculosis* PE31 (*Rv3477*) Attenuates Host Cell Apoptosis and Promotes Recombinant *M. smegmatis* Intracellular Survival via Up-regulating GTPase Guanylate Binding Protein-1. *Front Cell Infect Microbiol.* 2020;10 February:1–12.

51. Comas I, Chakravarti J, Small PM, Galagan J, Niemann S, Kremer K, *et al.* Human T cell

epitopes of *Mycobacterium tuberculosis* are evolutionarily hyperconserved. *Nat Genet.* 2010;42:498–503.

52. Donà V, Ventura M, Sali M, Cascioferro A, Provvedi R, Palù G, *et al.* The PPE Domain of PPE17 Is Responsible for Its Surface Localization and Can Be Used to Express Heterologous Proteins on the Mycobacterial Surface. *PLoS One.* 2013;8:1–8.

53. Fishbein S, van Wyk N, Warren RM, Sampson SL. Phylogeny to function: PE/PPE protein evolution and impact on *Mycobacterium tuberculosis* pathogenicity. *Mol Microbiol.* 2015;96:901–16.

54. Somerville W, Thibert L, Schwartzman K, Behr MA. Extraction of *Mycobacterium tuberculosis* DNA: A question of containment. *J Clin Microbiol.* 2005;43:2996–7.

55. Kolmogorov M, Yuan J, Lin Y, Pevzner PA. Assembly of long, error-prone reads using repeat graphs. *Nat Biotechnol.* 2019;37:540–6.

56. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, *et al.* Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One.* 2014;9.

57. 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.

58. Seemann T. Snippy: fast bacterial variant calling from NGS reads. 2015.

59. Pfeifer B, Wittelsbürger U, Ramos-Onsins SE, Lercher MJ. PopGenome: An efficient swiss army knife for population genomic analyses in R. *Mol Biol Evol.* 2014;31:1929–36.

60. Charif D, Lobry JR. SeqinR 1.0-2: A Contributed Package to the R Project for Statistical Computing Devoted to Biological Sequences Retrieval and Analysis. In: *Structural Approaches to Sequence Evolution. Biological and Medical Physics, Biomedical Engineering.* 2007. p. 207–

32.

61. Price MN, Dehal PS, Arkin AP. Fasttree: Computing large minimum evolution trees with profiles instead of a distance matrix. *Mol Biol Evol.* 2009;26:1641–50.

62. Tatusova T, Dicuccio M, Badretdin A, Chetvernin V, Nawrocki EP, Zaslavsky L, *et al.* NCBI prokaryotic genome annotation pipeline. *Nucleic Acids Res.* 2016;44:6614–24.

63. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol Biol Evol.* 2013;30:772–80.

64. Rizk G, Lavenier D, Chikhi R. DSK: K-mer counting with very low memory usage. *Bioinformatics.* 2013;29:652–3.

65. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics.* 2009;25:1754–60.

66. Quinlan AR, Hall IM. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics.* 2010;26:841–2.

67. Murrell B, Weaver S, Smith MD, Wertheim JO, Murrell S, Aylward A, *et al.* Gene-wide identification of episodic selection. *Mol Biol Evol.* 2015;32:1365–71.

68. Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, *et al.* Pfam: The protein families database in 2021. *Nucleic Acids Res.* 2021;49:D412–9.

Acknowledgements

PJG-G is funded by an MRC-LID PhD studentship. JEP is funded by a Newton Institutional Links Grant (British Council, no. 261868591). TGC was funded by the Medical Research Council UK (Grant no. MR/M01360X/1, MR/N010469/1, MR/R025576/1, and MR/R020973/1) and BBSRC (Grant no. BB/R013063/1). SC was 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

SC, JEP and TGC conceived and directed the project. PJG-G and ADG undertook sample processing and DNA extraction. PJG-G performed bioinformatic and statistical analyses under the supervision of SC, JEP and TGC. LT provided data. SC led the generation of sequence data, with some assistance from MLH. PJG-G, SC, JEP and TGC interpreted results. PJG-G 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 manuscript. PJG-G, JEP, and TGC compiled the final manuscript.

Table 1. Statistical significance of differences in SNP and indel diversity between each gene functional category and *pe/ppe* genes.

Functional Group	π				<i>dxy</i>			
	SNPs		Indels		SNPs		Indels	
	diff*	p adj	diff*	p adj	diff*	p adj	diff*	p adj
Cell wall and cell processes	0.00013	0.00518	0.00030	0	2.98E-04	1.39E-05	0.00032	0
Regulatory proteins	9.98E-05	0.40263	0.00029	2.03E-07	0.00026	0.01723	0.00035	5.77E-07
Virulence detoxification adaptation	0.00012	0.10995	0.00027	7.83E-06	2.95E-04	0.00136	0.00032	2.13E-05
Conserved hypotheticals	0.00012	0.01087	0.00028	0	0.00031	2.56E-06	0.00030	1.40E-11
Information pathways	0.00017	0.00171	0.00034	1.43E-10	0.00039	1.01E-06	0.00040	2.52E-09
Insertion seqs and phages	8.43E-05	0.87033	5.93E-05	0.96167	8.78E-05	0.99593	5.10E-05	0.99689
Intermediary metabolism and respiration	0.00018	2.44E-06	0.00035	0	0.00039	8.10E-09	0.00040	0
Lipid metabolism	0.00019	0.00010	0.00034	0	0.00041	6.60E-08	0.00037	2.37E-11
Unknown	-9.57E-05	0.99921	0.00025	0.46870	3.18E-05	1	0.00023	0.83652

*diff = difference between mean *dxy* or π in *pe/ppe* and the other functional group of comparison; p adj = P-value adjusted for multiple comparisons using Tukey's Honest Significant Differences. In bold, statistically significant adjusted P values (p adj < 0.01).

Table 2. Lineage- or clade-specific variants.

Gene (locus)	Lineage	Variant	AF in lineage *	AF in rest **	Gene Class	Comment
<i>ppe1</i> (Rv0096)	L6	P298P	0.986	0	C	
<i>pe1</i> (Rv0151c)	<i>M. bovis</i>	G26R	0.972	0	S	
	L3	G369R	1	0	S	
	L6	P494L	1	0	S	
<i>pe3</i> (Rv0159c)	<i>M. bovis</i>	P255T	1	0	C	
	L3	S175P	0.999	0	C	
<i>pe4</i> (Rv0160c)	L1	K164N	1	0	C	
	L3	F197S	0.999	0	C	
<i>ppe2</i> (Rv0256c)	L1	T412T	0.984	0	C	
	L5	E140G	1	0	C	
	L5	D431N	1	0	C	
<i>ppe3</i> (Rv0280)	L5	E448D	1	0	C	
	L6	M450T	1	0	C	
<i>ppe4</i> (Rv0286)	L3	L52M	0.955	0	C	
	Ancient	A185A	1	0	C	
<i>pe_pgrs5</i> (Rv0297)	L1	G225D	0.952	0	C	
<i>ppe5</i> (Rv0304c)	L1	I1273V	0.999	0	S	
	L3	G960A	0.955	0	S	
	Ancient	S1765F	0.998	0	S	
<i>ppe8</i> (Rv0355c)	L1	139_139del	1	0	S	
	<i>M. bovis</i>	G2403G	1	0	S	
	L1	V118A	1	0	S	
	L3	D741N	0.983	0	S	
	L3	S1920F	0.954	0	S	
	L5	F414V	1	0	S	
	Ancient	I3250F	1	0	S	
	Ancient	9889_9890insATA	0.999	0	S	Change in ORF of PPE8 until the end of PPE7 (gene fusion)
<i>ppe10</i> (Rv0442c)	<i>M. bovis</i>	W8*	0.991	0	C	Truncated protein
	L3	W147S	1	0	C	
	L6	G288A	1	0	C	
<i>pe_pgrs6</i> (Rv0532)	L3	A124V	0.997	0	S	
	Ancient	1557_1558insT	0.778	0	S	Truncated protein
<i>pe_pgrs7</i> (Rv0578c)	L1	G951R	0.981	0	C	
	L3	G405G	0.978	0	C	
<i>pe_pgrs10</i> (Rv0747)	L3	G799G	0.953	0	S	

pe_pgrs11 (Rv0754)	L1	G280R	0.999	0	C	
ppe12 (Rv0755c)	L5	G378S	0.996	0	S	
	Ancient	R545K	0.999	0	S	
pe_pgrs14 (Rv0834c)	L1	G668D	0.977	0	S	
	Ancient	A246A	0.950	0	S	
pe_pgrs15 (Rv0872c)	<i>M. bovis</i>	L113L	0.991	0	S	
ppe13 (Rv0878c)	L1	G336G	1	0	C	
	L6	N244N	0.993	0	C	
ppe14 (Rv0915c)	L5	T293M	1	0	C	
pe_pgrs16 (Rv0977)	L4.1	1968_1969insG	1	0	S	Truncated protein
pe10 (Rv1089)	L2/L3	337_337del	0.999	0	S	Delayed STOP, 26 more residues
pe_pgrs22 (Rv1091)	L2	G730G	0.952	0	S	
	L5	G118G	1	0	S	
ppe16 (Rv1135c)	L5	G349R	0.984	0	S	
	L6	1279_1283del	1	0	S	Truncated protein
ppe17 (Rv1168c)	L2	P167L	0.982	0	C	
pe12 (Rv1172c)	L5	L217F	1	0	C	
ppe18 (Rv1196)	L5	H234R	1	0	K	
pe14 (Rv1214c)	L2	A106A	0.981	0	C	
pe_pgrs23 (Rv1243c)	L5	G280G	0.952	0	S	
pe_pgrs24 (Rv1325c)	L5	L101R	1	0	C	
ppe19 (Rv1361c)	L3	F4V	0.965	0	S	
ppe20 (Rv1387)	<i>M. bovis</i>	V94A	0.991	0	C	
pe_pgrs25 (Rv1396c)	Ancient	N336T	0.853	0	S	
pe16 (Rv1430)	L2	A96A	0.998	0	C	
pe_pgrs28 (Rv1452c)	Ancient	1505_1506insGG CCGGCGG	0.866	0	S	In-frame
pe17 (Rv1646)	L3	T285I	1	0	C	
pe_pgrs30 (Rv1651c)	<i>M. bovis</i>	A172V	0.978	0	C	
	L3	T600N	0.999	0	C	
	L5	R115L	1	0	C	
ppe23 (Rv1706c)	L6	S37P	1	0	C	
ppe24 (Rv1753c)	L5	S716R	1	0	S	

<i>pe_pgrs31</i> (Rv1768)	Ancient	1064_1065insCG GTAACGGTGGGG GCGG	0.851	0	C	In-frame
<i>ppe25</i> (Rv1787)	<i>M. bovis</i>	925_927del	1	0	S	In-frame
<i>ppe28</i> (Rv1800)	Ancient	C144W	0.994	0	C	
<i>ppe29</i> (Rv1801)	L5	A366P	0.996	0	C	
<i>pe_pgrs32</i> (Rv1803c)	L5	E76D	1	0	S	
	L5	A483T	1	0	S	
<i>ppe31</i> (Rv1807)	L5	H188Y	1	0	C	
<i>ppe33</i> (Rv1809)	L3	G22S	0.985	0	S	
<i>ppe36</i> (Rv2108)	L5	I25I	1	0	C	
<i>ppe37</i> (Rv2123)	L5	V124M	1	0	S	
<i>pe_pgrs39</i> (Rv2340c)	L5	A109T	1	0	C	
<i>pe_pgrs40</i> (Rv2371)	L5	D29D	1	0	C	
<i>pe_pgrs41</i> (Rv2396)	<i>M. bovis</i>	S26N	0.991	0	S	
<i>pe24</i> (Rv2408)	L2	G216V	0.982	0	C	
<i>pe_pgrs42</i> (Rv2487c)	L5	G125G	1	0	S	
<i>pe_pgrs43</i> (Rv2490c)	L6	W1503R	0.971	0	C	
<i>pe26</i> (Rv2519)	L3	S330L	0.955	0	C	
	L5	G160S	1	0	C	
<i>pe_pgrs44</i> (Rv2591)	L5	A439A	0.984	0	C	
	Ancient	G478G	0.994	0	C	
<i>pe_pgrs45</i> (Rv2615c)	L3	G437G	0.998	0	K	
<i>pe_pgrs47</i> (Rv2741)	L1	S20S	1	0	S	
	Ancient	G383G	0.969	0	S	
<i>ppe43</i> (Rv2768c)	L5	449_454del	0.988	0	S	Truncated protein
<i>ppe44</i> (Rv2770c)	L1	G59V	1	0	C	
<i>ppe45</i> (Rv2892c)	L6	W75*	1	0	S	Truncated protein
<i>ppe48</i> (Rv3022A)	L3	I64L	0.999	0	C	
<i>lipY</i> (Rv3097c)	L4	A58G	1	0	S	
	L5	F129S	1	0	S	
<i>ppe54</i> (Rv3343c)	L3	G2189S	0.982	0	S	

<i>ppe56</i> (Rv3350c)	L6	6586_6586del	1	0	S	Truncated protein
<i>pe_pgrs55</i> (Rv3511)	L5	1411_1411del	0.956	0	S	Truncated protein
<i>pe_pgrs56</i> (Rv3512)	L5	991_1086del	0.940	0	S	Truncated protein
<i>ppe61</i> (Rv3532)	L1	T257M	1	0	C	
<i>ppe63</i> (Rv3539)	L1	Y365N	1	0	C	
<i>ppe64</i> (Rv3558)	L1	G306S	0.998	0	S	
	L3	63_64del	0.955	0	S	Truncated protein
<i>pe_pgrs58</i> (Rv3590c)	L2	A314V	0.969	0	C	
<i>pe_pgrs59</i> (Rv3595c)	L5	G22D	1	0	C	

AF = Allele frequency

* AF in indicated lineage; ** AF in the group of samples from other lineages.

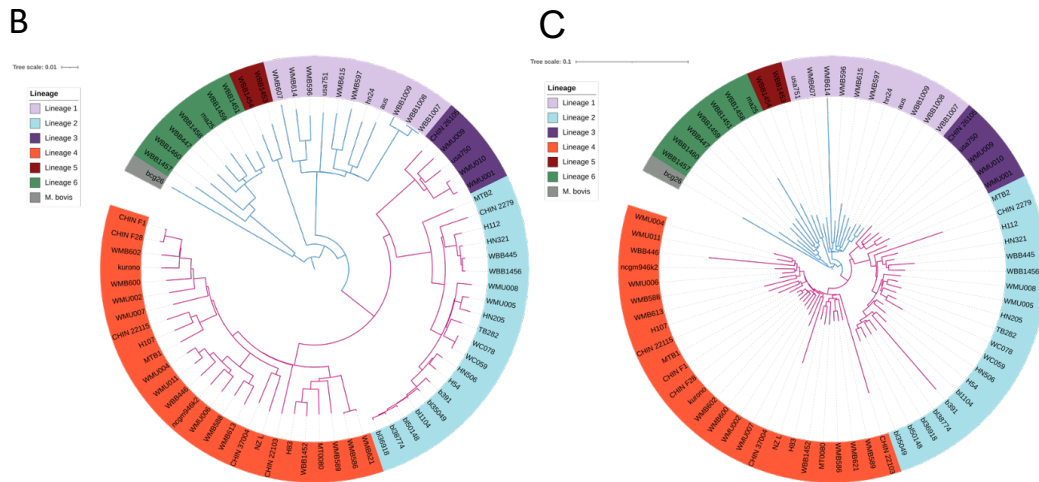
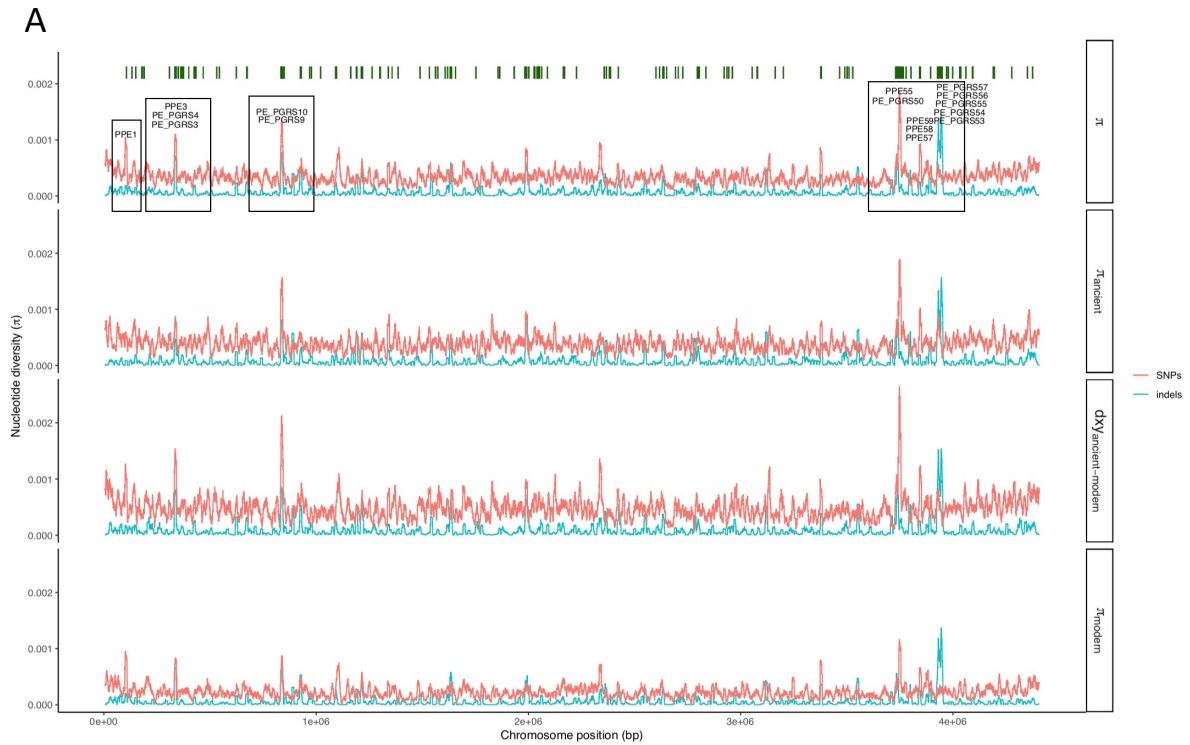


Figure 1. (A) Whole-genome SNPs nucleotide diversity and indel diversity. From top to bottom, the first track shows nucleotide diversity along the chromosome, with the peaks over 0.001 highlighted in a box. The *pe/ppe* genes in the peaks of nucleotide diversity are annotated. Green bars show where *pe/ppe* genes are located along the genome. Second track shows nucleotide diversity in ancient lineages. Third track shows absolute divergence between ancient and modern lineages. Fourth track shows nucleotide diversity in modern lineages. Line in read

represents SNPs diversity and in blue indel diversity. **(B)** Maximum likelihood phylogenetic tree reconstructed with whole genome SNPs (n=19,125). **(C)** Maximum likelihood phylogenetic tree reconstructed with whole genome indels (n=6,594). Ancient lineages are represented in blue, modern lineages in pink.

Tree scale: 0.1

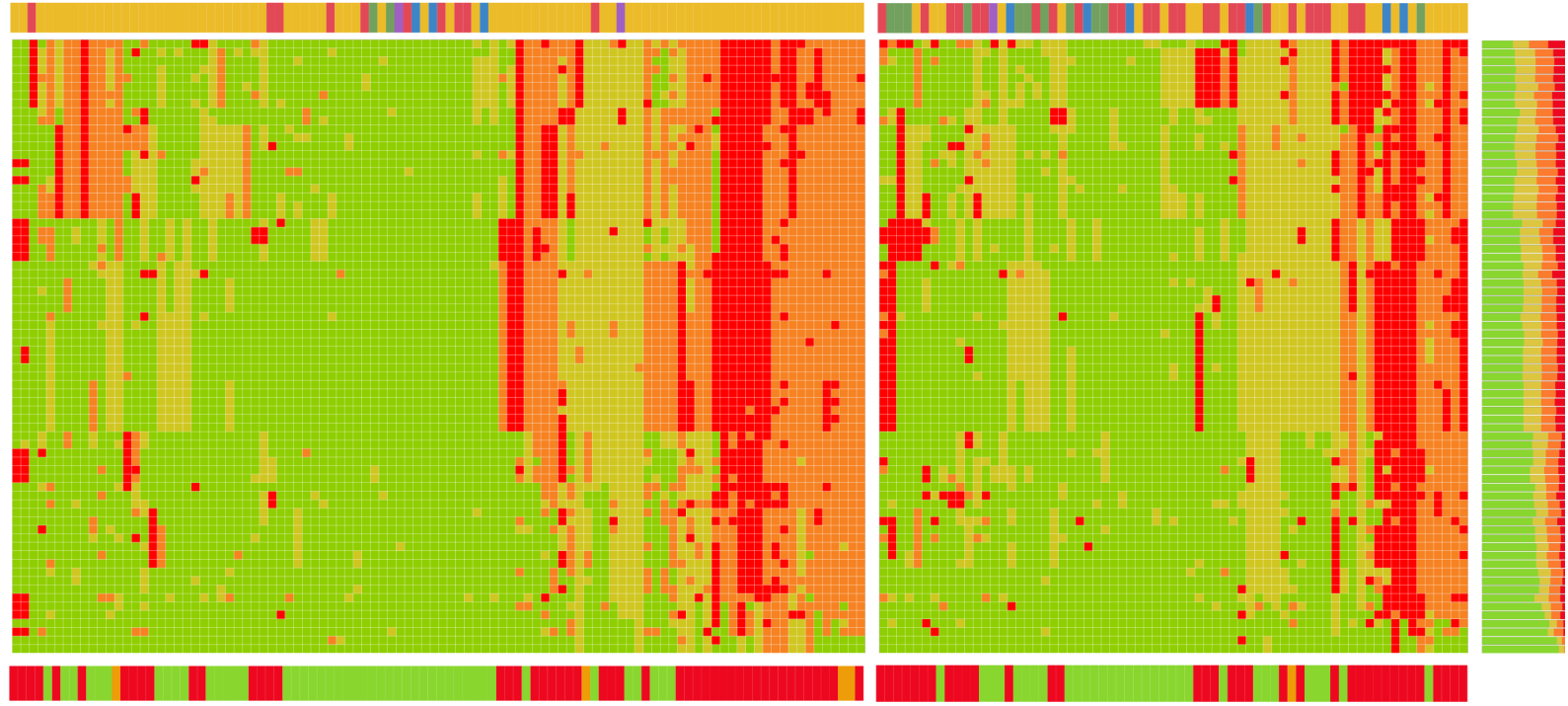
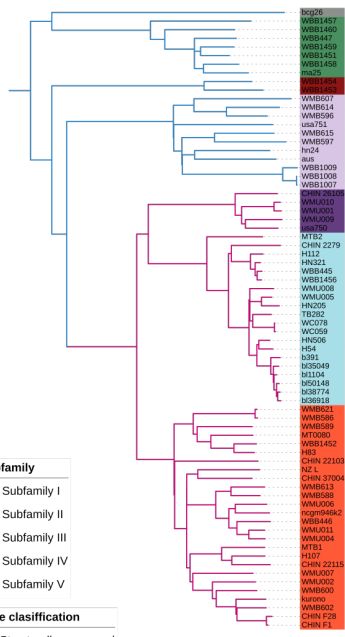
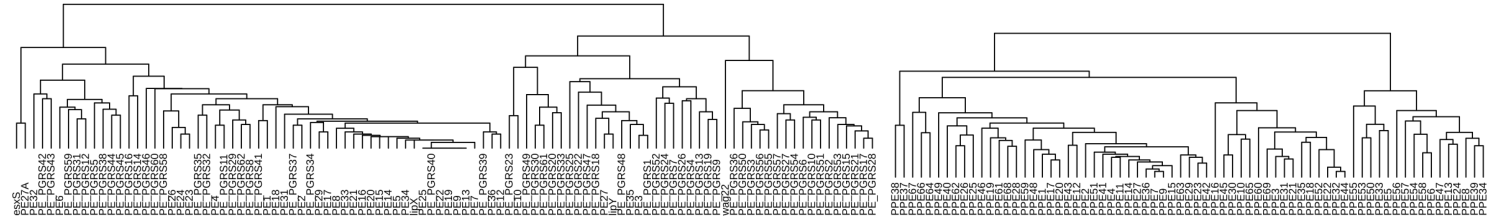
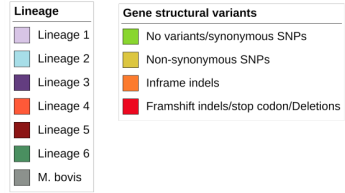


Figure 2. Heatmap showing the structural classification of each gene for each sample. Each row represents a separate sample, following the order based on the phylogenetic tree shown on the left. Genes on columns, *pe* family on the left, *ppe* family on the right. In green, genes without variants or synonymous SNPs; in yellow, genes with non-synonymous SNPs; in orange, genes with in-frame indels; in red, genes with frameshifts, changes in start/stop codons or large deletions. Top track shows the sub-family of each gene based on Gey Van Pittius *et al.* classification [6]. Bottom track summarises the structural classification of each gene across all samples in one of the following categories: structurally conserved (class C) in green, structural variants (class S) in red and unique *k-mer* profile (class K) in yellow. Barplot on the right shows the distribution of genes with each type of variant by sample.

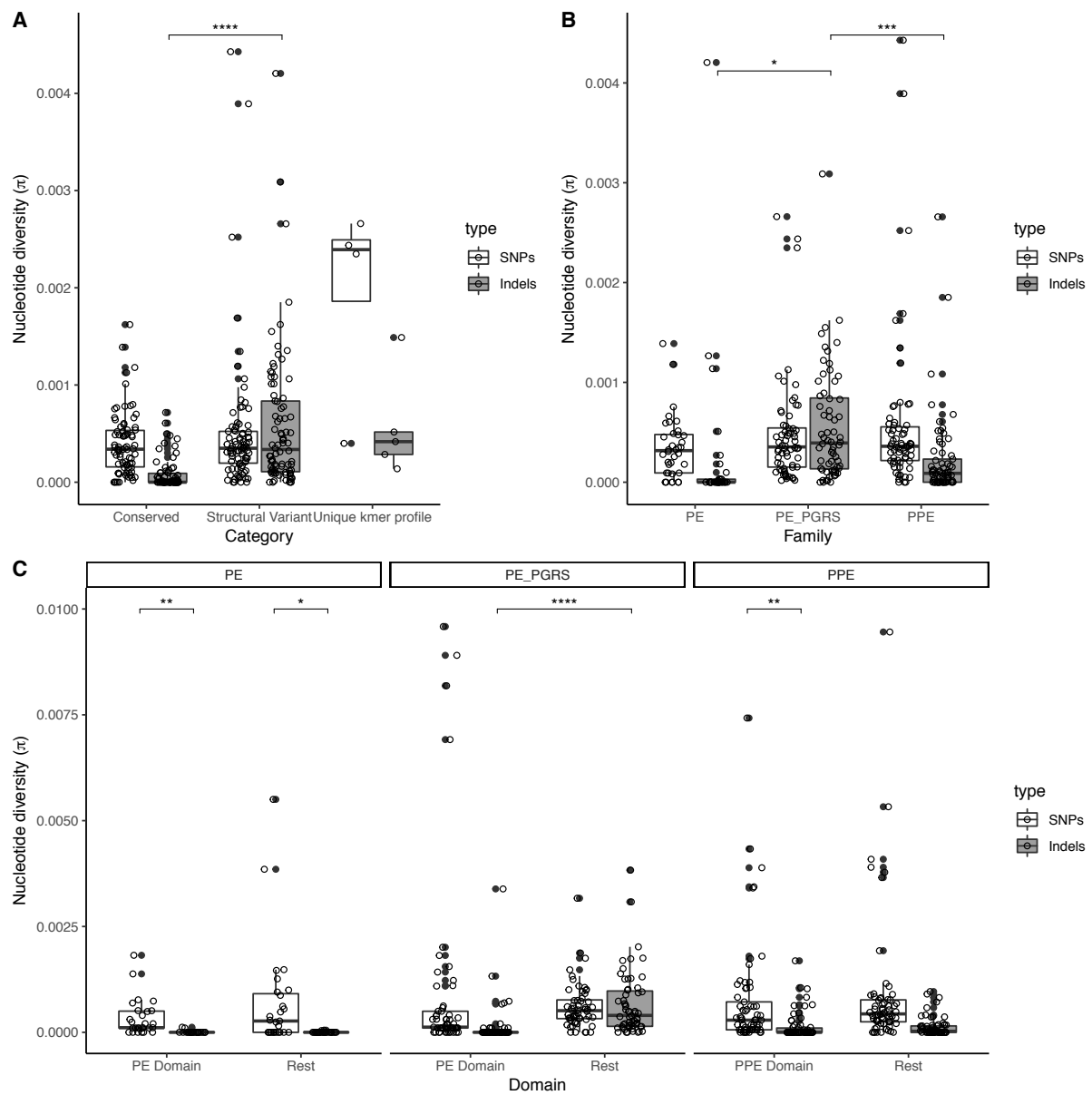


Figure 3. Boxplots of SNP and indel diversity in the 169 *pe/ppe* genes compared by **(A)** gene classification; **(B)** gene family and **(C)** domain within gene family. Outliers with $\pi > 0.005$ in **(A)** and **(B)** and $\pi > 0.01$ in **(C)** have been removed from figure. Adjusted P-value significant at (*) 5%, (**) 1%, (***) 0.1% or (****) 0.01%.

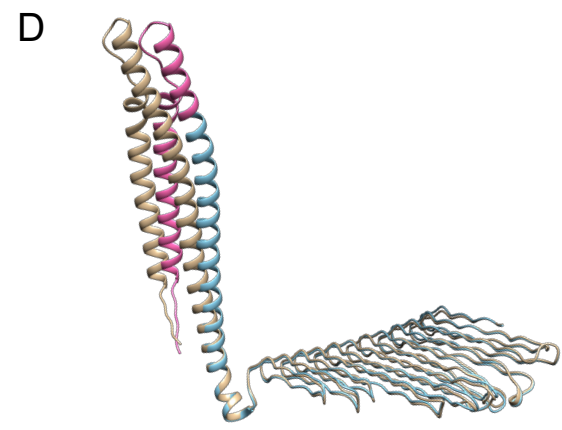
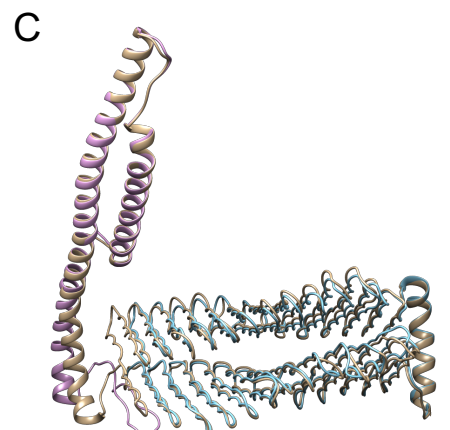
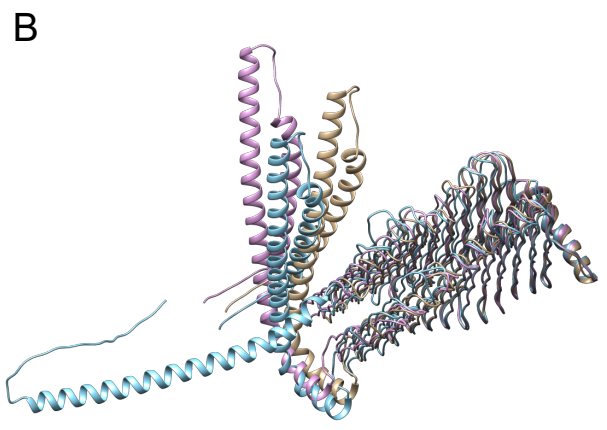
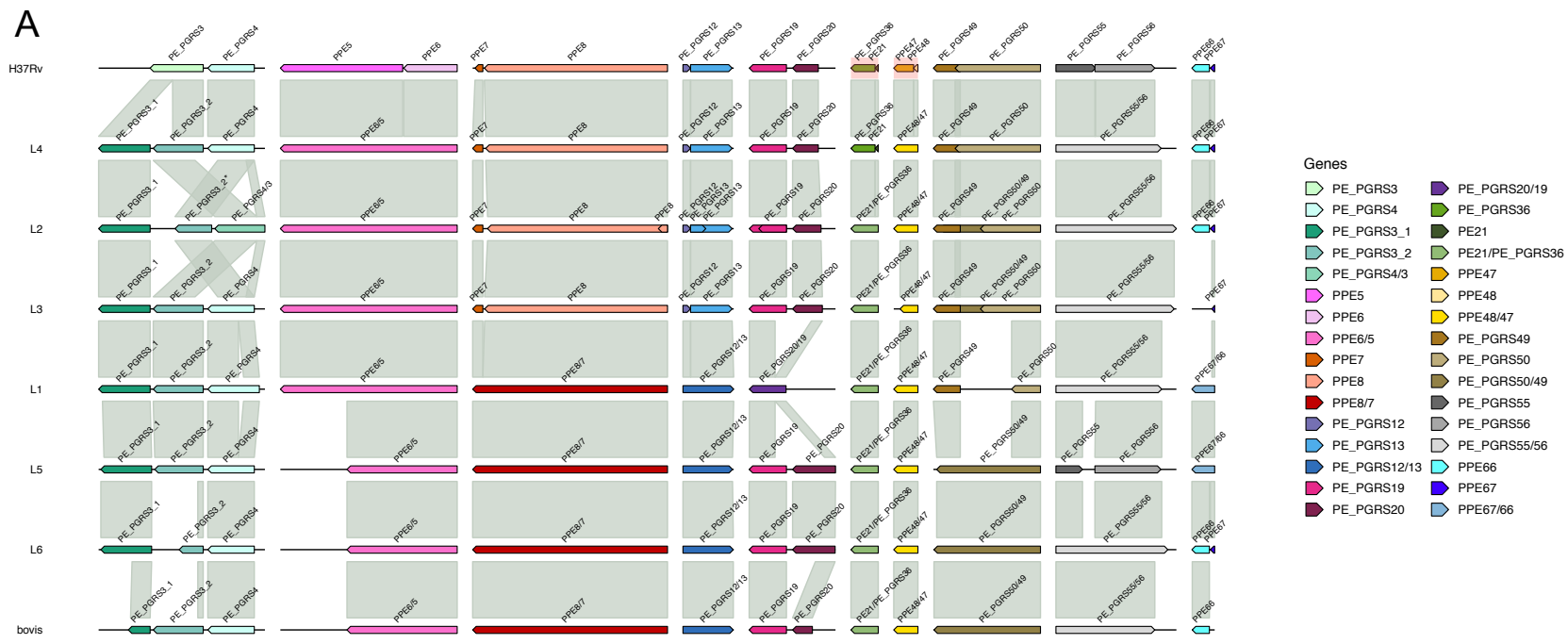


Figure 4. (A) Gene organisation of 10 pairs of consecutive genes where variants modify the open reading frame generating a gene fusion in at least one lineage. Gene organisation shown with representatives for each lineage; * only in some isolates from the lineage. **(B), (C)** and **(D)** Predicted protein structures by AlphaFold of **(B)** PE_PGRS4/3, **(C)** PE_PGRS12/13 and **(D)** PE21/PE_PGRS36. In beige, structure of the fused protein; in blue PE_PGRS4 **(B)**, PE_PGRS13 **(C)** and PE_PGRS36 **(D)**; in pink PE_PGRS3 **(B)**, PE_PGRS12 **(C)** and PE21 **(D)**.

Functional genetic variation in *pe/ppa* genes contributes to diversity in *Mycobacterium tuberculosis* lineages and potential interactions with the human host

Paula Josefina Gómez-González ¹ paula-josefina.gomez-gonzalez@lshtm.ac.uk

Anna D. Grabowska ² dr.anna.grabowska@gmail.com

Leopold Tientcheu ³ leopold.tientcheu@lshtm.ac.uk

Martin L. Hibberd ¹ martin.hibberd@lshtm.ac.uk

Susana Campino ¹ susana.campino@lshtm.ac.uk

Jody E. Phelan ¹ jody.phelan@lshtm.ac.uk

Taane G. Clark ^{1,4,*} taane.clark@lshtm.ac.uk

1. Faculty of Infectious and Tropical Diseases, London School of Hygiene & Tropical Medicine, WC1E 7HT London, UK
2. Department of Biophysics, Physiology and Pathophysiology, Medical University of Warsaw, 02-004 Warsaw, Poland
3. MRC Unit The Gambia at the London School of Hygiene and Tropical Medicine, Vaccines and Immunity Theme, Atlantic Road, Fajara, The Gambia.
4. Faculty of Epidemiology and Population Health, London School of Hygiene & Tropical Medicine, WC1E 7HT London, UK

* Correspondence: taane.clark@lshtm.ac.uk, Department of Infection Biology, Faculty of Infectious and Tropical Diseases, London School of Hygiene & Tropical Medicine, Keppel Street, London, UK

Additional File 1

Table S1. Metadata of samples analysed

Sample ID	Lineage	Sub-lineage	# Contigs	Length	Accession number	Assembly	Comments
kurono	4	4.9	1	4415078	AP014573	HGAP2	No Illumina data
CHIN_F1 (H37Rv)	4	4.9	1	4415075	CP010329	HGAP3	Illumina data (SRR3647351), corrected with pilon
CHIN_F28 (H37Ra)	4	4.9	1	4421998	CP010330	HGAP3	Illumina data (SRR3647352), corrected with pilon
WMB602	4	4.9	2	4432700		Flye	Illumina data (ERR221595), corrected with pilon
WMB600	4	4.8	1	4393237		Flye	Illumina data (ERR216945), corrected with pilon
WMU002	4	4.7	1	4398930		Flye	Illumina data (ERR163993), corrected with pilon
WMU007	4	4.6.1.2	1	4404359		Flye	Illumina data (ERR216919), corrected with pilon
CHIN_22115	4	4.5	1	4401829	CP010337	HGAP3	Illumina data (SRR3647361), corrected with pilon
MTB1	4	4.5	1	4433542	CP020381	HGAP2	No Illumina data
H107	4	4.5	1	4418796	CP019612	HGAP2	No Illumina data
CHIN_37004	4	4.4.2	1	4417090	CP010338	HGAP3	Illumina data (SRR3647362), corrected with pilon
NZ_L	4	4.4.1.1	1	4416671	CP044345	Canu	No Illumina data
WBB446	4	4.3.4.2	1	4369979		Flye	No Illumina data
WMU011	4	4.3.4.2.1	1	4363273		Flye	Illumina data (ERR181979), corrected with pilon
WMU004	4	4.3.4.2.1	1	4366577		Flye	No Illumina data
WMU006	4	4.3.4.1	1	4374435		Flye	Illumina data (ERR163992), corrected with pilon
ncgm946k2	4	4.3.4.1	1	4380602	AP017901	minimus2	No Illumina data
WMB613	4	4.3.3	2	4404702		Flye	No illumina data
WMB588	4	4.3.3	8	4401217		Flye	Illumina data (ERR181745), corrected with pilon
CHIN_22103	4	4.2.2	1	4399422	CP010339	HGAP3	Illumina data (SRR3647353), corrected with pilon
MT0080	4	4.1.2	1	4426525	CP041207	Canu	No Illumina data
WMB589	4	4.1.2	1	4424878		Flye	Illumina data (ERR181717), corrected with pilon
H83	4	4.1.2.1	1	4413214	CP019611	HGAP2	No Illumina data
WBB1452	4	4.1.2.1	1	4416367		Flye	No illumina data
WMB586	4	4.1.1.3	1	4400674		Flye	Illumina data (ERR181742), corrected with pilon

WMB621	4	4.1.1.3	3	4419731		Flye	Illumina data (ERR216982), corrected with pilon
CHIN_2279	2	2.2.1	1	4405033	CP010336	HGAP3	Illumina data (SRR3647360), corrected with pilon
bl35049	2	2.2.1	1	4427062	CP017593	Canu/Pilon	Illumina data used in assembly
bl36918	2	2.2.1	1	4441591	CP017594	Canu/Pilon	Illumina data used in assembly
bl38774	2	2.2.1	1	4431885	CP017595	Canu/Pilon	Illumina data used in assembly
b391	2	2.2.1	1	4406925	CP017596	Canu/Pilon	Illumina data used in assembly
bl50148	2	2.2.1	1	4444417	CP017597	Canu/Pilon	Illumina data used in assembly
bl1104	2	2.2.1	1	4380156	CP017598	Canu/Pilon	Illumina data used in assembly
H54	2	2.2.1	1	4416938	CP019610	HGAP2	No Illumina data
H112	2	2.2.1	1	4406346	CP019613	HGAP2	No Illumina data
WC078	2	2.2.1	1	4413712	CP022577	HGAP2	No Illumina data
WC059	2	2.2.1	1	4413669	CP022578	HGAP2	No Illumina data
HN205	2	2.2.1	1	4411033	AP018034	HGAP3	No Illumina data
HN321	2	2.2.1	1	4421540	AP018035	HGAP3	No Illumina data
HN506	2	2.2.1	1	4413362	AP018036	HGAP3	No Illumina data
WBB1456	2	2.2.1	1	4409920		Flye	No illumina data
WBB445	2	2.2.1	1	4410526		Flye	No illumina data
WMU008	2	2.2.1	1	4418906		Flye	Illumina data (ERR245831), corrected with pilon
WMU005	2	2.2.1	1	4421515		Flye	Illumina data (ERR181965), corrected with pilon
TB282	2	2.2.1.2	1	4425860	CP017920	HGAP2	No Illumina data
MTB2	2	2.2.2	1	4417716	CP022014	HGAP2	No Illumina data
CHIN_26105	3	3	1	4426920	CP010340	HGAP3	Illumina data (SRR3647354), corrected with pilon
usa750	3	3	1	4434666	CP046309	HGAP3	Illumina data used in assembly
WMU009	3	3	1	4441198		Flye	Illumina data (ERR190402), corrected with pilon
WMU001	3	3.1.1	1	4426849		Flye	Illumina data (ERR212147), corrected with pilon
WMU010	3	3.1.1	1	4423118		Flye	Illumina data (ERR212002), corrected with pilon
usa751	1	1	1	4441988	CP046308	Canu	Illumina data used in assembly
hn24	1	1.1.1	1	4399916	AP018033	HGAP3	Illumina data used in assembly

aus	1	1.1.1	1	4414769	CP045962	Canu	Illumina data (SRR10520175), corrected with pilon
WMB597	1	1.1.2	1	4427144		Flye	Illumina data (ERR181798), corrected with pilon
WMB615	1	1.1.2	1	4436831		Flye	Illumina data (ERR212157), corrected with pilon
WBB1007	1	1.1.3	1	4432578		Flye	No illumina data
WBB1008	1	1.1.3	1	4432521		Flye	No illumina data
WBB1009	1	1.1.3	1	4422821		Flye	No illumina data
WMB614	1	1.2.2	1	4427580		Flye	Illumina data (ERR212155), corrected with pilon
WMB607	1	1.2.2	3	4415876		Flye	Illumina data (ERR221596), corrected with pilon
WMB596	1	1.2.2	3	4456111		Flye	Illumina data (ERR181794), corrected with pilon
WBB1453	5	5	1	4424589		Flye	No illumina data
WBB1454	5	5	1	4419154		Flye	No illumina data
ma25	6	6	1	4386422	CP010334	HGAP3	Illumina data (SRR3647358), corrected with pilon
WBB1451	6	6	1	4373719		Flye	No illumina data
WBB1457	6	6	1	4389577		Flye	No illumina data
WBB1458	6	6	1	4358247		Flye	No illumina data
WBB447	6	6	1	4382892		Flye	No illumina data
WBB1459	6	6	2	4385170		Flye	No illumina data
WBB1460	6	6	2	4400942		Flye	No illumina data
bcg26	<i>bovis</i>	<i>bovis</i>	1	4351313	CP010331	HGAP3	Illumina data (SRR3647355), corrected with pilon

Table S2. Classification and diversity of *pe* genes

Gene (locus)	Sub-family *	Pfam Domains	Class **	Comments	# SNPs	SNPs π	# Indels	Indels π	<i>dN/dS</i>
<i>pe1 (Rv0151c)</i>	V	PE, PE-PPE	S	Truncated in L1.1.3 (852_853ins)	14	0.00061176	1	3.10E-05	4.5773
<i>pe2 (Rv0152c)</i>	V	PE, PE-PPE	C		8	0.0002869	1	1.76E-05	0.3512
<i>pe3 (Rv0149c)</i>	V	PE, PE-PPE	C		7	0.0003245	0	NA	28.8236
<i>pe4 (Rv0160c)</i>	V	PE, PE-PPE	C		7	0.0004672	0	NA	26.5716
<i>pe5 (Rv0285)</i>	II	PE	C		2	0.0001798	0	NA	25.418
<i>pe6 (Rv0335c)</i>	V	PE	S	Truncated in L1 (139_139del)	2	0.00039882	1	0.00050876	21.2085
<i>pe7 (Rv0916c)</i>	IV	PE	C		1	9.26E-05	0	NA	17.325
<i>pe8 (Rv1040c)</i>	IV	PE, PPE-SVP	C		5	0.000327	1	3.35E-05	0.4109
<i>pe9 (Rv1088)</i>	V	PE	C		0	NA	0	NA	0.9251
<i>pe10 (Rv1089)</i>	V	-	S	Delayed STOP in L2 and L3 (337_337del, 26 residues more)	3	0.0005141	1	0.0012664	16.9303
<i>lipX/pe11 (Rv1169c)</i>	IV	PE	C		1	9.17E-05	0	NA	22.3073
<i>pe12 (Rv1172c)</i>	V	PE	C		5	0.0003651	0	NA	1.2589
<i>pe13 (Rv1195)</i>	IV	PE	C		0	NA	0	NA	0.9251
<i>pe14 (Rv1214c)</i>	V	PE	C		3	0.0013887	0	NA	0.7575
<i>pe15 (Rv1386)</i>	II	PE	C		1	0.0001773	0	NA	14.0187
<i>pe16 (Rv1430)</i>	V	PE, PE-PPE	C		3	0.0003084	0	NA	0.6162
<i>pe17 (Rv1646)</i>	V	PE	C		3	0.0002	1	2.98E-05	23.4983
<i>pe18 (Rv1788)</i>	IV	PE	S	Deleted in some samples	2	0.0003469	1	0.00026995	0.2942
<i>pe19 (Rv1791)</i>	IV	PE	C		0	NA	0	NA	0.9251
<i>pe20 (Rv1806)</i>	IV	PE	C		3	0.0002778	0	NA	0.7106
<i>pe21 (Rv2099c)</i>	V	PE	C	Pseudogene (no stop codon), continues into PE_PGRS36	2	0.00047443	0	NA	22.5983

<i>pe22 (Rv2107)</i>	III	PE	C		0	NA	0	NA	0.9251
<i>pe23 (Rv2328)</i>	V	PE	C		2	0.0007542	0	NA	26.3814
<i>pe24 (Rv2408)</i>	V	PE	C		2	0.0006037	0	NA	12.4811
<i>pe25 (Rv2431c)</i>	III	PE	C		1	9.26E-05	0	NA	0
<i>pe26 (Rv2519)</i>	V	PE	C		8	0.0004875	0	NA	0.3857
<i>pe27 (Rv2769c)</i>	IV	PE	C		6	0.0011803	0	NA	28.4766
<i>pe27a (Rv3018A)</i>	V	-	S	Deleted in some samples	1	0.00025633	1	0.00420467	26.2812
<i>esxS/pe28 (Rv3020c)</i>	V	WXG100	S	Deleted in some samples	0	NA	1	0.00113778	1
<i>pe29 (Rv3022A)</i>	V	PE	C		3	0.0005924	0	NA	8.3594
<i>pe31 (Rv3477)</i>	IV	PE	S	Truncated in sporadic samples	3	0.00045315	0	NA	17.5541
<i>pe32 (Rv3622c)</i>	IV	PE	S	Deleted in L6 and <i>bovis</i> (RD8)	0	NA	0	NA	1.002
<i>pe33 (Rv3650)</i>	V	PE	C		2	0.0001949	1	9.75E-05	0.4656
<i>pe34 (Rv3746c)</i>	I	PE	C		1	8.27E-05	0	NA	10.8427
<i>pe35 (Rv3872)</i>	I	PE	S	Truncated in L5 (5_5del), deleted in <i>bovis</i> (RD1)	1	0.0003495	1	0.00018258	0.9246
<i>pe36 (Rv3893c)</i>	III	PE	C		1	0.0001947	1	1.86E-05	28.5633
<i>pe_pgrs1 (Rv0109)</i>	V	PE	C		2	0.0001947	1	1.86E-05	18.0182
<i>pe_pgrs2 (Rv0124)</i>	V	PE	S	Deleted in L6 (RD701), truncated in L4.3.3 (591_591insG)	19	0.00050294	16	0.00067077	0.8504
<i>pe_pgrs3 (Rv0278c)</i>	V	PE	S	Gene fusion with PE_PGRS4 in L2 due to deletion, duplication of PE_PGRS3 in other lineages (except H37Rv/Ra/4.6)	17	0.00013804	24	0.00082985	0.7585
<i>pe_pgrs4 (Rv0279c)</i>	V	PE	S	Gene fusion with PE_PGRS3 in L2 due to deletion, sporadic premature STOPS	56	0.00106364	18	0.00044835	0.4797
<i>pe_pgrs5 (Rv0297)</i>	V	PE	C		10	0.0003423	7	0.0004465	1.5899
<i>pe_pgrs6 (Rv0532)</i>	V	PE	S	Truncated in ancient lineages (1557_1558insT)	13	0.0004103	11	0.00106302	1.6463
<i>pe_pgrs7 (Rv0578c)</i>	V	PE	C		23	0.0004749	10	0.000344	0.8156
<i>pe_pgrs8 (Rv0742)</i>	V	PE	C		1	5.26E-05	1	0.0004972	19.5189
<i>pe_pgrs9 (Rv0746)</i>	V	PE	S	Truncated in sporadic samples	16	7.27E-05	29	0.00122112	0.4987
<i>pe_pgrs10 (Rv0747)</i>	V	PE	S	Truncated in L5 (1742_1824del), and sporadic samples	13	5.56E-05	21	0.00112428	0.4029
<i>pe_pgrs11 (Rv0754)</i>	V	PE, His_Phos_1	C		7	0.0002599	1	0.0001014	0.4889

<i>pe_pgrs12 (Rv0832)</i>	V	PE	S	Gene fusion with PE_PGRS13 in ancient lineages (392_393insG)	1	6.71E-05	1	0.00101211	0
<i>pe_pgrs13 (Rv0833)</i>	V	-	S	Gene fusion with PE_PGRS12 in ancient lineages, truncated in some L2 and sporadic samples	13	0.00029595	29	0.00108468	0.5459
<i>pe_pgrs14 (Rv0834c)</i>	V	PE	S	Truncated in L1.1.3 (472_472del) and sporadic samples	14	0.00045474	11	0.00021253	0.3931
<i>pe_pgrs15 (Rv0872c)</i>	V	PE	S	Truncated in some L2 samples (589_589del)	7	0.00034397	6	0.00016393	0.362
<i>pe_pgrs16 (Rv0977)</i>	V	PE	S	Truncated in L4.1 (1968_1969insG)	9	0.00011686	10	0.00018475	0.2054
<i>pe_pgrs17 (Rv0978c)</i>	V	PE, NHL	K	Differences in sequence in lab strains (H37Rv and H37Ra)	12	0.00243698	3	0.00013788	0.2417
<i>pe_pgrs18 (Rv0980c)</i>	V	PE, NHL	K	Differences in sequence in L4.1	17	0.00266035	4	0.0002856	0.2674
<i>pe_pgrs19 (Rv1067c)</i>	V	PE	S	Gene fusion with PE_PGRS20 in L1 due to deletion, in-frame insertions in L6 leading to extra PGRS motifs, truncated in sporadic samples	25	0.00036781	23	0.00101167	0.8365
<i>pe_pgrs20 (Rv1068c)</i>	V	PE	S	Gene fusion with PE_PGRS19 in L1 due to deletion, truncated in sporadic samples	17	0.00045452	15	0.0013994	0.2407
<i>pe_pgrs21 (Rv1087)</i>	V	PE	K	Differences in sequence in L3	15	0.00039939	31	0.00148836	1.5651
<i>pe_pgrs22 (Rv1091)</i>	V	PE	S	Truncated in L1.1.3 (Q68*) and L5 (409_409del)	27	0.00077056	21	0.00065511	0.4182
<i>pe_pgrs23 (Rv1243c)</i>	V	PE	S	Truncated in L3 (661_661del)	4	7.92E-05	6	0.00038429	0.9196
<i>pe_pgrs24 (Rv1325c)</i>	V	PE	C		14	0.0005368	6	0.0003012	0.587
<i>pe_pgrs25 (Rv1396c)</i>	V	PE	S	Truncated in some L2 and L4, different fs	15	0.0009773	8	0.00035055	0.7405
<i>pe_pgrs26 (Rv1441c)</i>	V	PE	C		11	0.0005635	14	0.0007167	1.9764
<i>pe_pgrs27 (Rv1450c)</i>	V	PE	S	Truncated in some samples, different sequences	53	0.00084699	33	0.00062245	0.2947
<i>pe_pgrs28 (Rv1452c)</i>	V	PE	S	Different sequences, truncated in L5	131	2.00E-05	16	0.00131344	0.4814
<i>pe_pgrs29 (Rv1468c)</i>	V	PE	C		5	0.0005206	2	9.77E-05	0.5473
<i>pe_pgrs30 (Rv1651c)</i>	V	PE	C		15	0.0003366	6	0.0002088	0.7663
<i>wag22 (Rv1759c)</i>	V	-	S	Deleted in several samples (RD152)	12	0.00017089	11	0.00040235	0.4067
<i>pe_pgrs31 (Rv1768c)</i>	V	PE	C		9	0.0001492	2	0.0002551	40.1193
<i>pe_pgrs32 (Rv1803c)</i>	V	PE	S	Truncated in sporadic samples	13	0.00041875	1	1.45E-05	2.5034
<i>pe_pgrs33 (Rv1818c)</i>	V	PE	S	Truncated in L1 (1009_1009del)	9	0.00049081	8	0.00075895	0.8811
<i>pe_pgrs34 (Rv1840c)</i>	V	PE	C		1	1.79E-05	2	5.33E-05	49.3495
<i>pe_pgrs35 (Rv1983)</i>	V	PE	S	Missing in sporadic samples	9	0.00024729	0	NA	0.7638

<i>pe_pgrs36 (Rv2098c)</i>	V	PE	S	Pseudogene (no start codon), continuation of PE21 ORF in all non-L4 and L4.4 (4_5insC) leading to gene fusion PE21/PE_PGRS36	5	0.00012711	5	0.00050366	1.0747
<i>pe_pgrs37 (Rv2126c)</i>	V	-	C		2	0.0001071	2	7.21E-05	0.3612
<i>pe_pgrs38 (Rv2162c)</i>	V	PE	C		11	0.0004903	8	0.0004018	0.6228
<i>pe_pgrs39 (Rv2340c)</i>	V	PE	C		10	0.000493	0	NA	0.6159
<i>pe_pgrs40 (Rv2371)</i>	V	PE	C		2	0.0011274	0	NA	0
<i>pe_pgrs41 (Rv2396)</i>	V	PE	S	Truncated in L3.1.1 (397_397del)	10	0.00077491	3	0.00010159	0.3329
<i>pe_pgrs42 (Rv2487c)</i>	V	PE	S	Truncated in 2 L4 samples	11	0.00030942	4	0.00014017	0.4927
<i>pe_pgrs43 (Rv2490c)</i>	V	PE	C		24	0.0002429	11	0.0001237	1.2423
<i>pe_pgrs44 (Rv2591)</i>	V	PE	C		11	0.0010117	7	0.0003721	0.4302
<i>pe_pgrs45 (Rv2615c)</i>	V	PE	K	Differences in sequence	20	0.00234798	4	0.00041721	0.4864
<i>pe_pgrs46 (Rv2634c)</i>	V	PE	S	Truncated in L5 (1490_1491insG) and sporadic samples	17	0.00057154	6	0.00011685	2.0905
<i>pe_pgrs47 (Rv2741)</i>	V	PE	S	Truncated in L6 and <i>bovis</i> (28_28del)	11	0.00071603	5	0.00021446	0.2756
<i>pe_pgrs48 (Rv2853)</i>	V	PE	S	Sequences missing/deleted	17	0.0005352	4	8.93E-05	1.5105
<i>lipY (Rv3097c)</i>	V	PE, Abhydrolase_3	S	Truncated in sporadic samples	10	0.00066367	3	6.34E-05	1.5359
<i>pe_pgrs49 (Rv3344c)</i>	V	-	S	Change in ORF in all except L4 (20_20del) making it continuation of PE_PGRS50 (gene fusion)	9	0.00029255	9	0.00086475	2.3838
<i>pe_pgrs50 (Rv3345c)</i>	V	PE	S	Truncated in L1 and some L2 (811_811del); rest of L2/3/5/6/ <i>bovis</i> ORF continues into PE_PGRS49 (4356_4356del = PE_PGRS49 20_20del) leading to gene fusion	35	0.00010067	47	0.0011893	1.1135
<i>pe_pgrs51 (Rv3367)</i>	V	PE	S	Truncated in L5 (309_391del) and sporadic samples	15	0.00032836	4	0.00012266	0.8626
<i>pe_pgrs52 (Rv3388)</i>	V	PE	S	Truncated in sporadic samples	9	0.00015268	16	0.00083645	0.6388
<i>pe_pgrs53 (Rv3507)</i>	V	PE	S	Truncated in L5 and some L2samples (1111_1111del)	29	0.00046484	36	0.0008926	0.8128
<i>pe_pgrs54 (Rv3508)</i>	V	PE	S	Truncated in L6 (461_462insC), some L3 (3718_1718del) and sporadic samples	422	4.57E-05	95	0.00162233	0.7237
<i>pe_pgrs55 (Rv3511)</i>	V	PE	S	Truncated in L5 (1213_1213del), rest except 4.7-9 ORF continues into PE_PGRS56 (2108_2108del) leading to gene fusion	12	0.00032557	37	0.00135428	0.9282

<i>pe_pgrs56 (Rv3512)</i>	V	-	S	Truncated in L5, continuation of PE_PGRS55 in the rest except L4.7-9 (1_1del = PE_PGRS55 2108_2108del) leading to gene fusion	48	3.60E-05	52	0.00155033	0.9011
<i>pe_pgrs57 (Rv3514)</i>	V	PE	S	Truncated in L6 (461_462insC), truncated in most of L2 (796_850del) and in sporadic samples	674	3.75E-05	208	0.00308849	0.6838
<i>pe_pgrs58 (Rv3590c)</i>	V	PE	C		12	0.0006998	4	0.0001525	1.7062
<i>pe_pgrs59 (Rv3595c)</i>	V	PE	C		7	0.0003628	4	0.000481	0.6077
<i>pe_pgrs60 (Rv3652)</i>	V	PE	S	Change in ORF in L4.3 (249_249del) leading to longer protein sequence	1	0.00081973	2	0.0006533	28.6189
<i>pe_pgrs61 (Rv3653)</i>	V	PE	S	Truncated in most L3 (115_115del)	5	0.00018364	4	0.0003227	19.6831
<i>pe_pgrs62 (Rv3812)</i>	V	PE	C		6	0.0003352	1	1.83E-05	0.6114

* Sub-family classification based on Gey Van Pittius *et al.* (2006) [7].

** Class: C = conserved; S = structural variant; K = unique *k-mer* profile

Table S3. Classification and diversity of *ppe* genes

Gene (locus)	Sub-family *	Pfam Domains	Class **	Comments	# SNPS	SNPs π	# Indels	Indels π	<i>dN/dS</i>
<i>ppe1</i> (Rv0096)	II (PPW)	PPE, PPE-PPW	C		10	0.0007993	0	NA	1.1121
<i>ppe2</i> (Rv0256c)	II (PPW)	PPE, PPE-PPW	C		10	0.0005202	0	NA	2.7497
<i>ppe3</i> (Rv0280)	II (PPW)	PPE, PPE-PPW	C		7	0.0007669	0	NA	1.1022
<i>ppe4</i> (Rv0286)	II (PPW)	PPE, PPE-PPW	C		9	0.0004828	0	NA	0.1912
<i>ppe5</i> (Rv0304c)	V (MPTR)	MPTR	S	Truncated in L5/6/ <i>bovis</i> (2997_2997del) and in sporadic samples	34	0.00037107	10	0.00015809	0.4433
<i>ppe6</i> (Rv0305c)	V (MPTR)	PPE, MPTR	S	All samples except L1.1.3 (truncated 2678_2678del) and lab strains H37Rv/Ra change in ORF (2429_2429del) which continues until the end of PPE5	7	0.00032062	7	0.0001227	0.42
<i>ppe7</i> (Rv0354c)	V (MPTR)	-	C	Different from H37Rv, 42 aa longer (372_373insG)	0	NA	1	0	0.9251
<i>ppe8</i> (Rv0355c)	V (MPTR)	PPE, MPTR	S	Truncated in some L2 (453_453del); ancient lineages change in ORF (9889_9890insTA) leading to 211 residues more (until the end of PPE7 ORF)	41	0.00041735	15	0.0001845	0.4087
<i>ppe9</i> (Rv0388c)	IV (SVP)	PPE, PPE-SVP	C	Different from H37Rv, 263 aa longer and SVP domain (492_493insC, 501_502insC)	1	5.12E-05	2	0	1.3175
<i>ppe10</i> (Rv0442c)	V (MPTR)	PPE, MPTR	C		9	0.0003439	3	0.0001438	2.0863
<i>ppe11</i> (Rv0453)	II (PPW)	PPE, PPE-PPW	C		4	0.0001872	0	NA	0.3283
<i>ppe12</i> (Rv0755c)	V (MPTR)	PPE, MPTR	S	Truncated in L5 (87_87del)	12	0.00038861	4	7.13E-05	0.4687
<i>ppe13</i> (Rv0878c)	V (MPTR)	PPE, MPTR	C	polyC/polyA region masked in analysis, as there might be errors due to sequencing	10	0.0005381	5	0.00060741	0.5377
<i>ppe14</i> (Rv0915c)	IV (SVP)	PPE, PPE-SVP	C		4	0.0001504	0	NA	0.93
<i>ppe15</i> (Rv1039c)	IV (SVP)	PPE, PPE-SVP	C		2	7.02E-05	0	NA	23.8303
<i>ppe16</i> (Rv1135c)	V (MPTR)	PPE, MPTR	S	Truncated in most L2 (IS6110) and L6 (1279_1283del)	8	0.00021974	2	0.00026693	14.9883
<i>ppe17</i> (Rv1168c)	IV (SVP)	PPE, PPE-SVP	C		5	0.0006419	0	NA	0.4819

<i>ppe18</i> (Rv1196)	IV (SVP)	PPE, PPE-SVP	K	Different sequences	98	0.0068438	8	0.00051633	0.3642
<i>ppe19</i> (Rv1361c)	IV (SVP)	PPE, PPE-SVP	S	Truncated in L1.1.3 (Q145*)	104	0.00389265	2	6.93E-05	0.4275
<i>ppe20</i> (Rv1387)	II (PPW)	PPE, PPE-PPW	C		9	0.0003804	0	NA	1.3113
<i>ppe21</i> (Rv1548c)	V (MPTR)	PPE, MPTR	C		17	0.0002516	3	5.24E-05	0.4219
<i>ppe22</i> (Rv1705c)	IV (SVP)	PPE, PPE-SVP	C		6	0.0005706	0	NA	27.207
<i>ppe23</i> (Rv1706c)	IV (SVP)	PPE, PPE-SVP	C		4	0.0002882	0	NA	1.0831
<i>ppe24</i> (Rv1753c)	V (MPTR)	PPE, MPTR	S	Truncated in sporadic samples	29	0.00052289	10	0.00031217	0.1083
<i>ppe25</i> (Rv1787)	IV (SVP)	PPE, PPE-SVP	S	Deleted in some samples	17	0.00119295	3	0.00021842	0.7069
<i>ppe26</i> (Rv1789)	IV (SVP)	PPE, PPE-SVP	S	Deleted in some samples	9	0.00026744	2	9.20E-05	0.3614
<i>ppe27</i> (Rv1790)	IV (SVP)	PPE, PPE-SVP	S	Deleted in some samples	2	0.00024448	1	7.69E-05	0.2488
<i>ppe28</i> (Rv1800)	V (MPTR)	PPE, PE-PPE	C		9	0.0004781	1	1.41E-05	2.5387
<i>ppe29</i> (Rv1801)	IV (SVP)	PPE, PPE-SVP	C		14	0.0004869	0	NA	0.9722
<i>ppe30</i> (Rv1802)	IV (SVP)	PPE, PPE-SVP	S	Truncated in L6 (Q162*), truncated in some L2	7	0.00024031	1	3.93E-05	1.6392
<i>ppe31</i> (Rv1807)	IV (SVP)	PPE, PPE-SVP	C		9	0.0003593	0	NA	0.7717
<i>ppe32</i> (Rv1808)	IV (SVP)	PPE, PPE-SVP	C		3	0.0007869	0	NA	0.2067
<i>ppe33</i> (Rv1809)	IV (SVP)	PPE, PPE-SVP	S	L1/5/6 with 1 residue more (*469S), truncated in <i>bovis</i>	10	0.00058393	1	1.97E-05	0.2663
<i>ppe34</i> (Rv1917c)	V (MPTR)	PPE, MPTR	S	Truncated in most lineages due to IS6110	43	0.00013756	25	0.00077604	0.5102
<i>ppe35</i> (Rv1918c)	V (MPTR)	PPE, MPTR	S	Truncated in sporadic samples	14	0.00035995	2	2.79E-05	1.7086
<i>ppe36</i> (Rv2108)	III	PPE	C		3	0.0001507	0	NA	0.1015
<i>ppe37</i> (Rv2123)	II (PPW)	PPE, PPE-PPW	S	Truncated in L2 (503_503del) and L3 (1219_1219del), delayed STOP in some L4 (1016_1017del) adding 23 residues)	12	0.00035602	4	0.00049001	0.74

<i>ppe38</i> (Rv2352c)	IV (SVP)	PPE, PPE-SVP	S	Deletion of beginning of gene in L2 (RD185), samples missing	2	4.32489E-05	7	0.0006647025	12.0194
<i>ppe39</i> (Rv2353c)	V (MPTR)	MPTR	S	Deletion of beginning of the gene in most isolates, missing samples	4	0.000222252	6	0.0007001844	0.2822
<i>ppe40</i> (Rv2356c)	V (MPTR)	PPE, MPTR	S	Missing samples, truncated in sporadic samples (IS6110)	2	2.88E-05	4	8.89E-05	18.7541
<i>ppe41</i> (Rv2430c)	III	PPE	C		1	4.75E-05	1	4.75E-05	0
<i>ppe42</i> (Rv2608)	V (MPTR)	PPE, PE-PPE	C		6	0.0002101	0	NA	0.2935
<i>ppe43</i> (Rv2768c)	IV (SVP)	PPE, PPE-SVP	S	Truncated in L5 (449_454del)	5	0.00038959	1	4.62E-05	26.114
<i>ppe44</i> (Rv2770c)	IV (SVP)	PPE, PPE-SVP	C		8	0.0007808	0	NA	0.6649
<i>ppe45</i> (Rv2892c)	IV (SVP)	PPE, PPE-SVP	S	Truncated in L6 (W75*)	5	0.00036318	0	NA	19.2502
<i>ppe46</i> (Rv3018c)	II (PPW)	PPE, PPE-PPW	S	Truncated in 4.1.1.3 (IS6110) and in other sporadic samples	28	0.00252083	4	0.00018677	0.4056
<i>ppe47</i> (Rv3021c)	II (PPW)	PPE, PPE-PPW	S	Pseudogene, all different to reference (12_13insG) making the ORF to continue until the end of PPE47; deleted in some samples	3	0.0003273	8	0.00054272	0.3328
<i>ppe48</i> (Rv3022c)	II (PPW)	PPE	C	Pseudogene, no stop codon until end of PPE47 except in ref, where fs in PPE47 (12_13insG) creates premature stop	1	0.0016213	1	0.0001143	14.7954
<i>ppe49</i> (Rv3125c)	IV (SVP)	PPE, PPE-SVP	S	Truncated in L1.1.3 and L3.1.1 (IS6110), truncated in sporadic samples	10	0.00029808	5	0.00023155	0.5397
<i>ppe50</i> (Rv3135)	IV (SVP)	PPE, PPE-SVP	S	L1 deleted; insertion in L2/5/6/ <i>bovis</i> adding SVP domain (331_332ins)	5	0.00075796	4	0.00265825	0.4003
<i>ppe51</i> (Rv3136)	IV (SVP)	PPE, PPE-SVP	C		3	7.19E-05	0	NA	0.1923
<i>ppe52</i> (Rv3144c)	V (MPTR)	PPE	S	Truncated in 3.1.1 (970_970del)	8	0.0005261	2	8.91E-05	0.4428
<i>ppe53</i> (Rv3159c)	V (MPTR)	PPE, MPTR	S	L1/2/3/4.1/4.2/5/6 truncated (88_89ins or IS6110)	12	5.72E-04	3	0.000387926	2.3535
<i>ppe54</i> (Rv3343c)	V (MPTR)	PPE, MPTR	S	Truncated in sporadic samples (IS6110/big insertions)	127	0	27	0.00038333	0.3772
<i>ppe55</i> (Rv3347c)	V (MPTR)	PPE, MPTR	S	Truncated in L4.5 (IS6110), L5/6/ <i>bovis</i> and sporadic samples, missing samples	151	0.00134505	20	0.00012244	0.4674
<i>ppe56</i> (Rv3350c)	V (MPTR)	PPE, MPTR	S	Truncated in L2 (6081_6081del) and L6 (6586_6586del), missing samples	223	0.00030085	19	0.00010512	0.3403
<i>ppe57</i> (Rv3425)	III	PPE	S	Deleted in all L1, half of L4 and some other sporadic samples; truncated in L2 (226_226del)	11	0.00038461	5	0.00185156	0.6238

<i>ppe58</i> (Rv3426)	III	PPE	S	Deleted in some samples; truncated in all except L4.9 (373_373del)	6	0.00032183	2	0.0010836	2.2304
<i>ppe59</i> (Rv3429)	III	PPE	S	Deleted (>50%) in sporadic samples	45	0.00948948	1	0.000102	2.4137
<i>ppe60</i> (Rv3478)	IV (SVP)	PPE, PPE-SVP	S	Truncated in sporadic samples	85	0.00442826	5	0.0005147	0.6678
<i>ppe61</i> (Rv3532)	IV (SVP)	PPE, PPE-SVP	C		6	0.0003509	3	0.0001528	24.5627
<i>ppe62</i> (Rv3533c)	V (MPTR)	PPE, MPTR	C		5	9.35E-05	3	4.76E-05	0.183
<i>ppe63</i> (Rv3539)	V (MPTR)	PPE, PE-PPE	C		6	0.0003831	0	NA	1.6317
<i>ppe64</i> (Rv3558)	V (MPTR)	PPE, MPTR	S	Truncated in L3 (63_64del)	4	0.00020847	3	0.0002936	0.2259
<i>ppe65</i> (Rv3621c)	IV (SVP)	PPE, PPE-SVP	S	Deleted in L6/ <i>bovis</i> (RD8);	2	0.00019593	0	NA	0.3229
<i>ppe66</i> (Rv3738c)	II (PPW)	PPE, PPE-PPW	S	Deleted in L3	6	0.00043581	1	0.00016343	0.6127
<i>ppe67</i> (Rv3739c)	II (PPW)	PPE	S	Truncated in L3 (152_234del); L1 and L5 delayed STOP (*78W) leading ORF to continue until the end of PPE66	5	0.00168867	2	0.00067881	1.1496
<i>pep68</i> (Rv3873)	I	PPE	C	Deleted in <i>bovis</i> (RD1)	4	0.0003315	0	NA	27.142
<i>ppe69</i> (Rv3892c)	III	PPE	S	Truncated in some L2 due to deletion	9	0.00059859	2	9.06E-05	0.6972

* Sub-family classification based on Gey Van Pittius *et al.* (2006) [7]/

** Class: C = conserved; S = Structural variant; K = unique *k-mer* profile

Table S4. Genes with IS6110 integrated within the coding region

Locus	Gene	# Samples with IS6110*	Consequence
<i>Rv1040c</i>	<i>pe8</i>	1	frameshift
<i>Rv1135c</i>	<i>ppe16</i>	14 (L2.2.1)	frameshift
<i>Rv1753c</i>	<i>ppe24</i>	1	frameshift
<i>Rv1800</i>	<i>ppe28</i>	1	frameshift
<i>Rv1917c</i>	<i>ppe34</i>	34 (n=20 L2, n=5 L3)	frameshift/stop codon
<i>Rv2352c</i>	<i>ppe38</i>	17 (L2)	Frameshift/stop codon
<i>Rv2356c</i>	<i>ppe40</i>	1	frameshift
<i>Rv3018c</i>	<i>ppe46</i>	3 (n=2 L4.1.1.3)	frameshift/stop codon
<i>Rv3021c</i>	<i>ppe47</i>	1	stop codon
<i>Rv3125c</i>	<i>ppe49</i>	7 (n=3 L3.1.1, n=2 L1.1.3)	frameshift/stop codon
<i>Rv3159c</i>	<i>ppe53</i>	2	frameshift/stop codon
<i>Rv3343c</i>	<i>ppe54</i>	1	stop codon
<i>Rv3347c</i>	<i>ppe55</i>	3 (L4.5)	frameshift

* In brackets, if there is lineage patterns, number and lineage where samples belonged to.

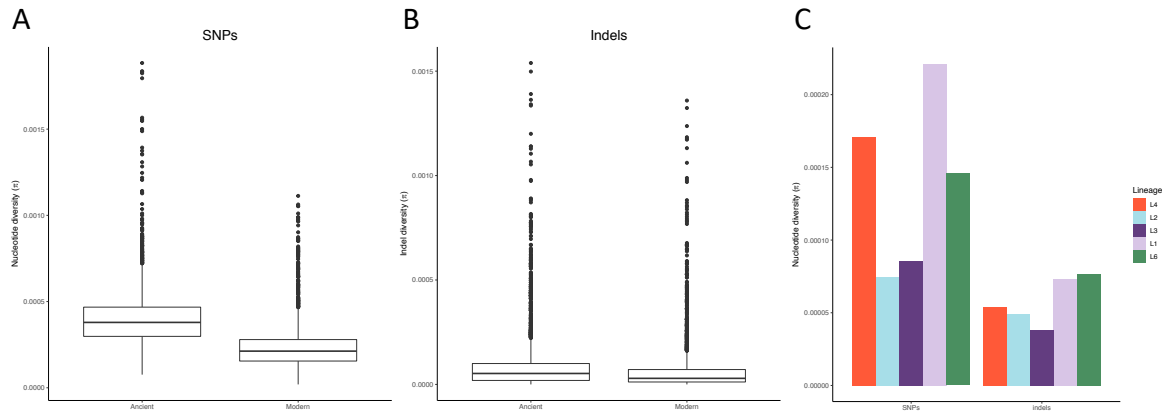


Figure S1. Boxplots of whole-genome nucleotide diversity (π) for **(A)** SNPs and **(B)** indels between ancient and modern lineages. **(C)** SNP and indel π by lineage (L5 and *bovis* excluded due to low number of isolates).

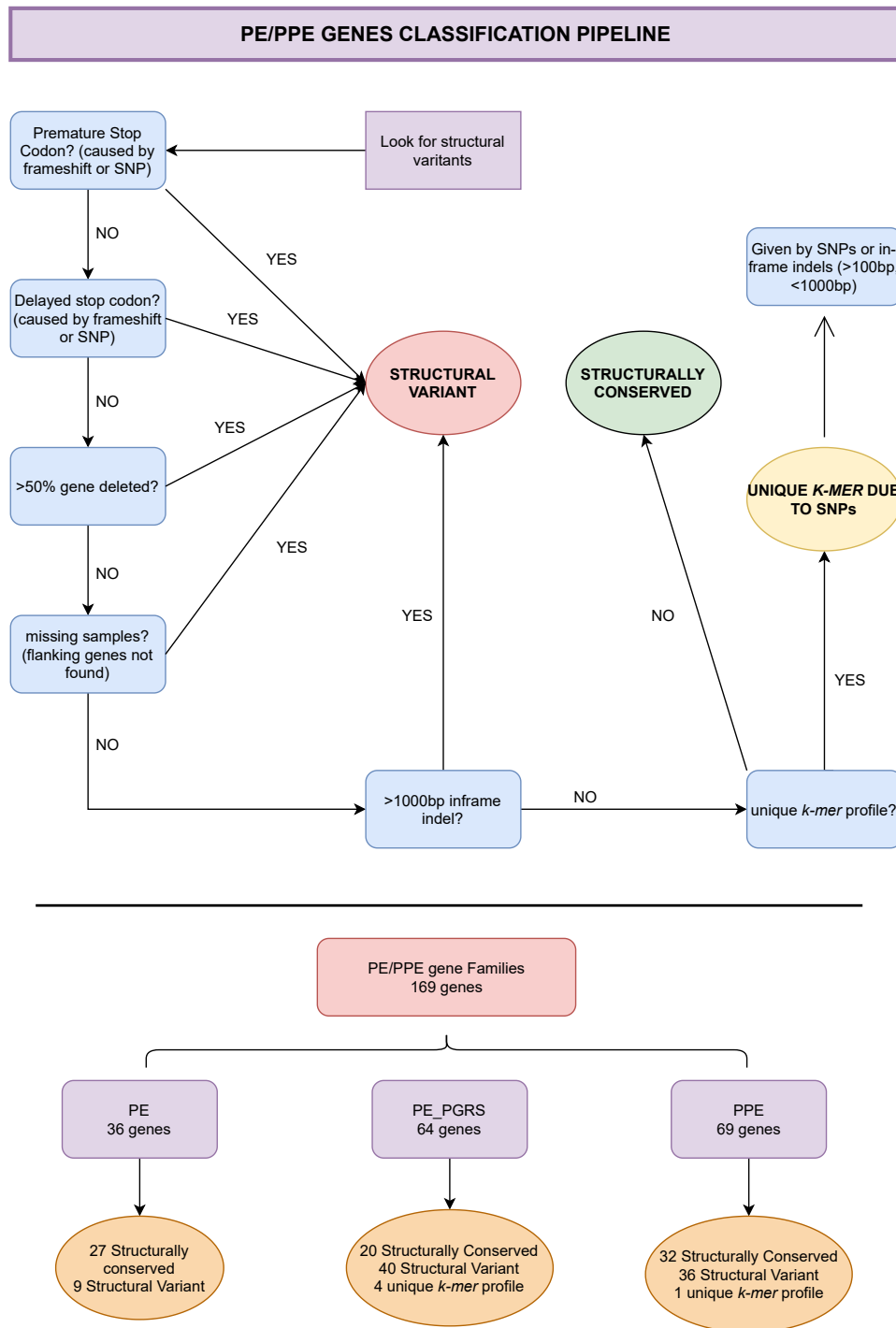
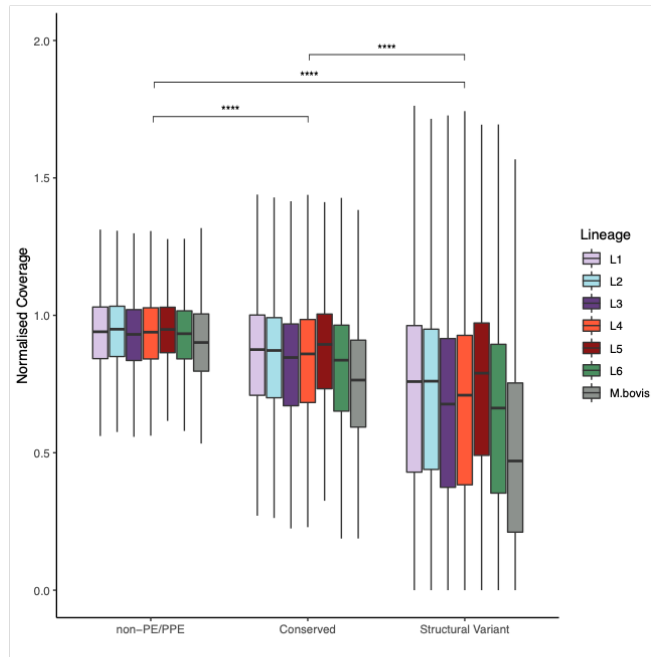


Figure S2. Flowchart showing the pipeline followed for the classification of *pe* and *ppe* genes.

A



B

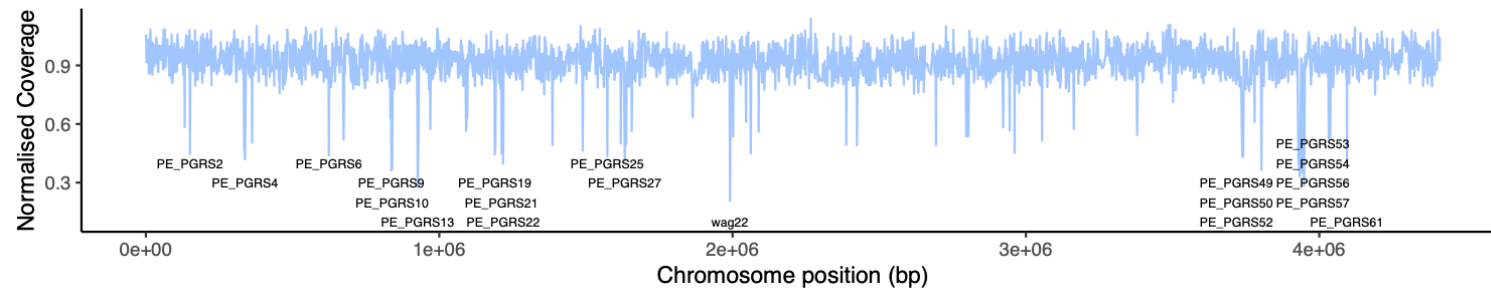


Figure S3. (A) Normalised coverage by gene category. The *pe/ppc* genes are divided in "Conserved" and "Structural Variant" based on the classification pipeline in **Figure S2**. Genes belonging to the "Unique k-mer" category are included in "Structural Variant". Every other gene in the genome is under "non-PE/PPE". Normalised coverage is shown by lineage for each category. Statistical differences were calculated between the means for each category.

*** = P-value adjusted < 0.001.

(B) Mean normalised coverage per gene along the genome. The 20 genes with the lowest mean normalised coverage are annotated.

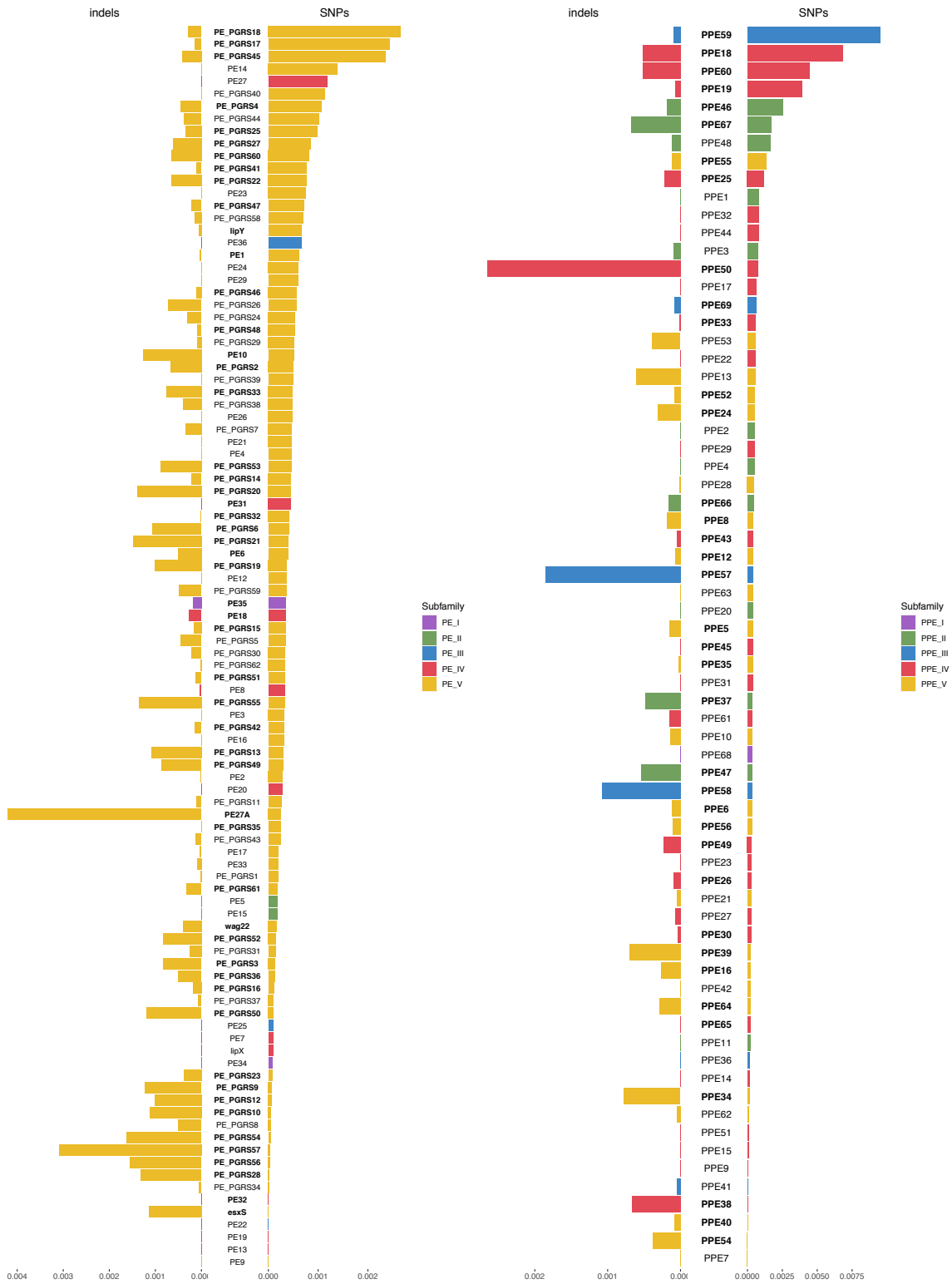


Figure S4. SNP and indel nucleotide diversity in both *pe* and *ppe* gene families. Colours respond to subfamilies. Genes in bold belong to the class S or K.

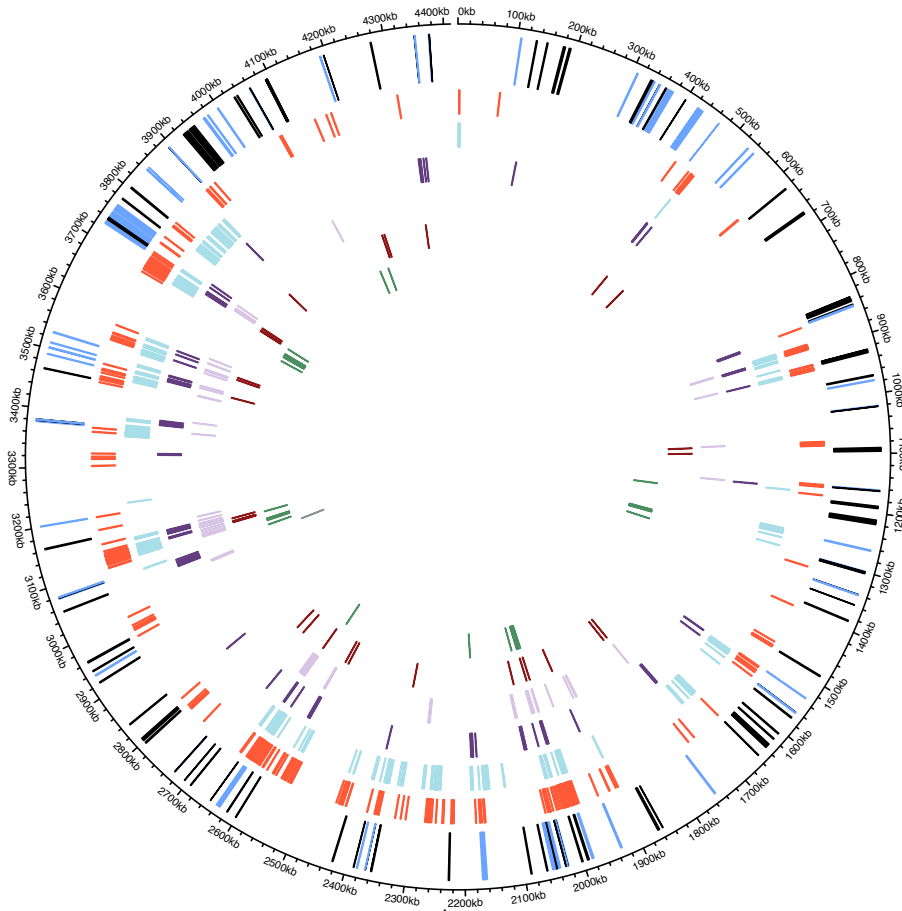


Figure S5. Circos plot showing the location *IS6110* along the genome for the different lineages. First track (outside to inside) refers to the location of *pe* (in black) and *ppe* (in blue) genes. Second to eighth track represent each of the position where *IS6110* is integrated in the samples belonging to each lineage as follows (in order): red for L4, light blue for L2, purple for L3, lilac for L1, brown for L5, green for L6 and grey for *M. bovis* BCG.

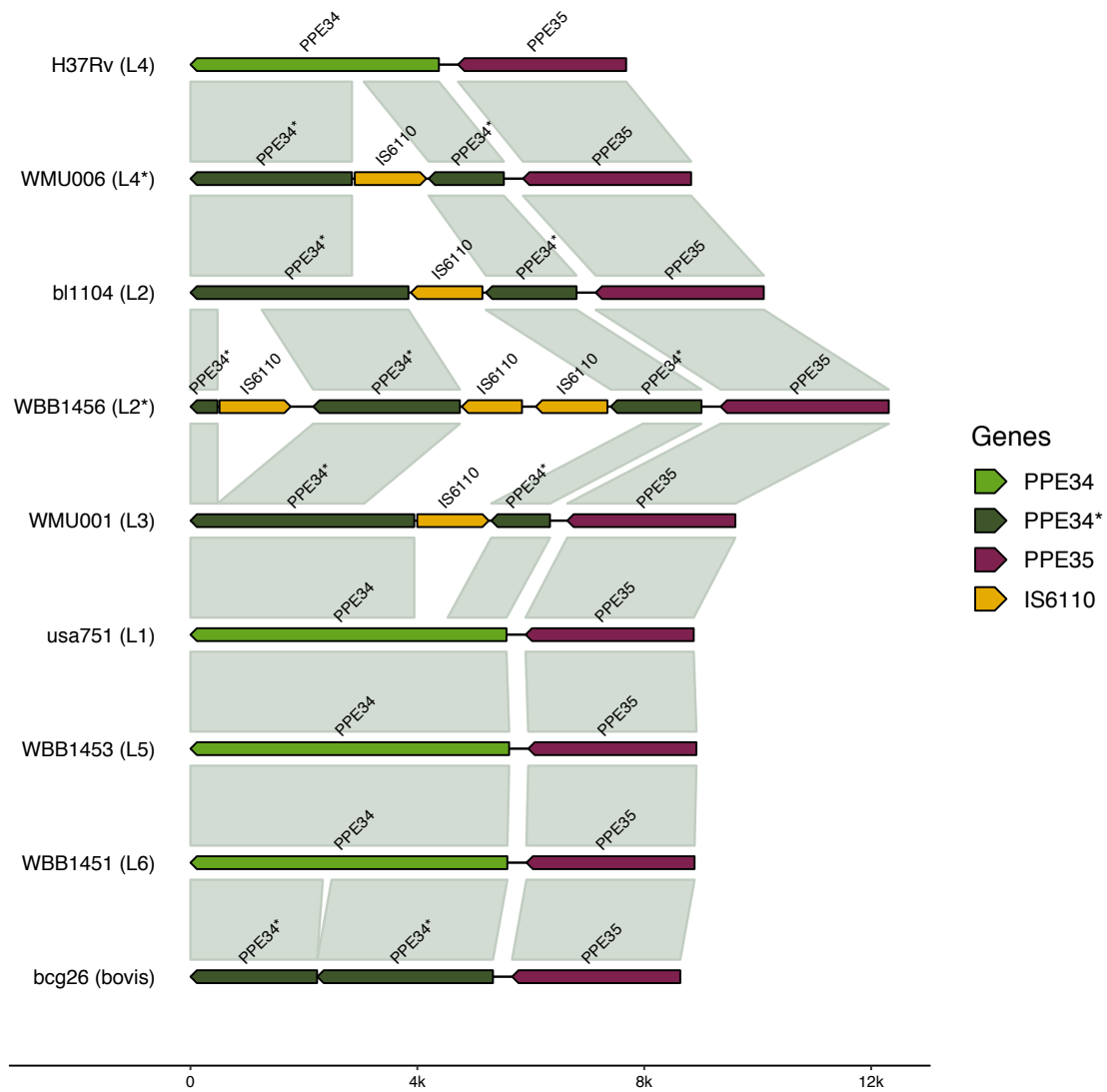


Figure S6. The *ppe34/35* loci organisation in representative strains for each lineage. PPE34* =

truncated *ppe34* gene.

* Sporadic isolates.

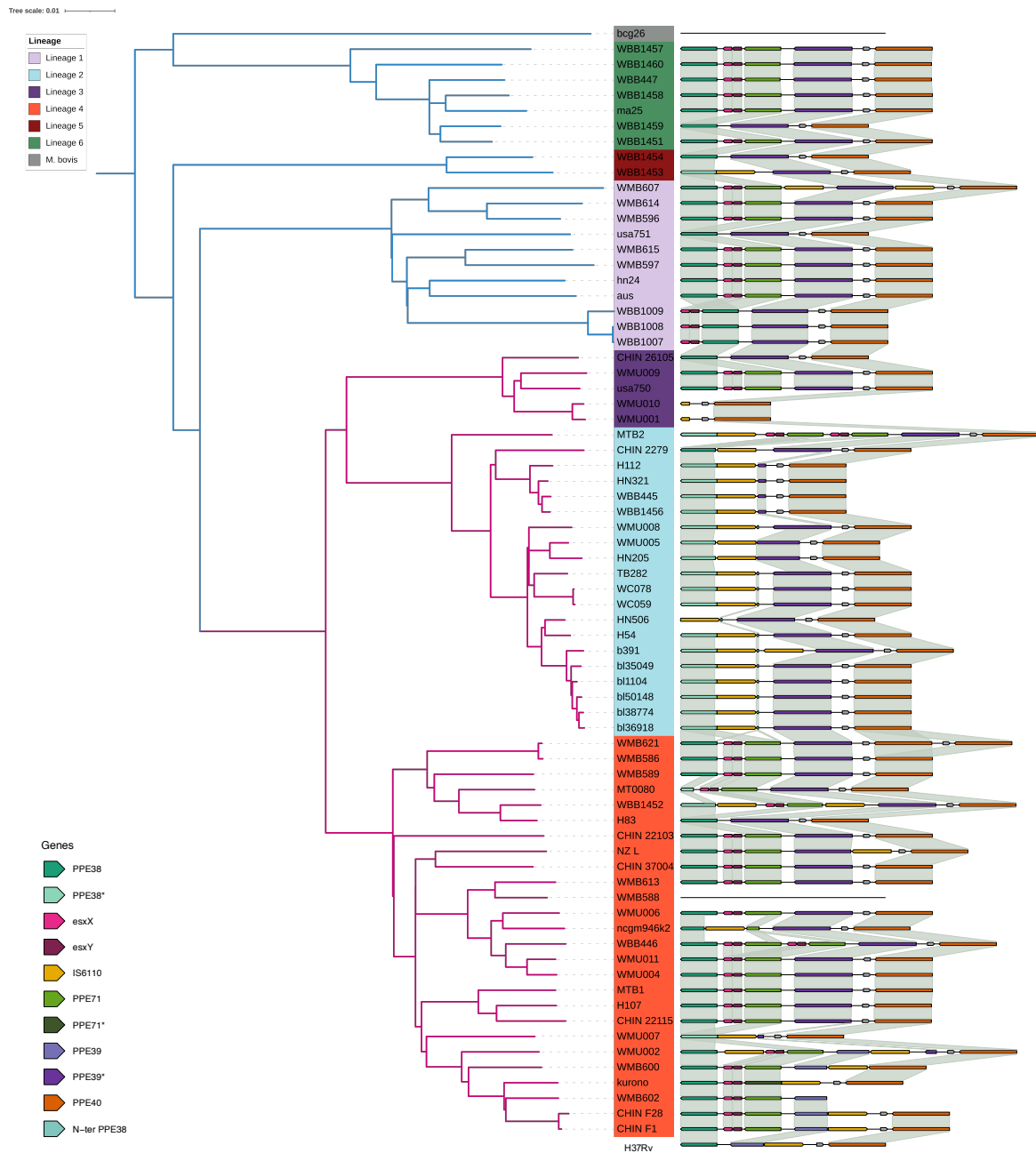


Figure S7. Configuration of *ppe38-ppe40* loci across the 72 samples analysed. H37Rv annotation shown at the bottom track. Samples ordered based on the phylogenetic tree shown on the left. Branches in pink represent modern lineages, in blue ancient lineages.

* Genes with different N-/C-terminal.

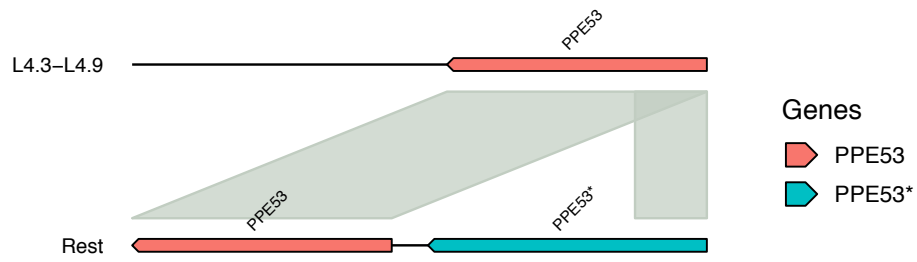


Figure S8. The *ppe53* locus representation in H37Rv and 4.3-4.9 lineage (first track) and the rest of lineages (second track). PPE53* indicates the 77% similar duplicated gene.

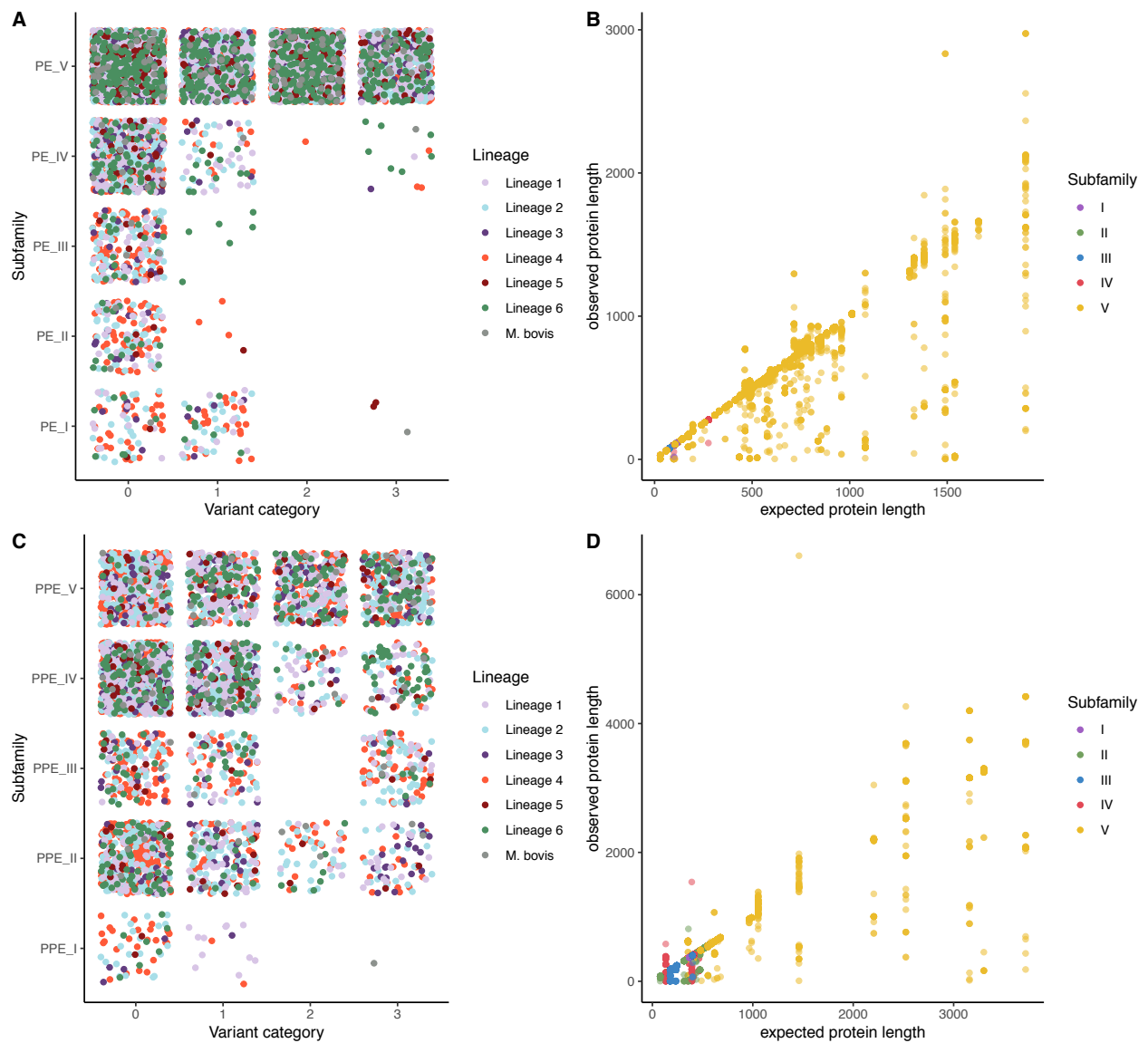


Figure S9. (A) and (C) Distribution of genes relative to their variant level (0 = no variant/ synonymous SNPs; 1 = non-synonymous SNPs; 2 = in-frame indels; 3 = frameshift/premature stop codon/big deletion) and their sub-family, for *pe* family **(A)** and *ppe* family **(C)**. **(B) and (D)** observed gene length vs expected gene length, coloured by sub-family, for *pe* family **(B)** and *ppe* family **(D)**.

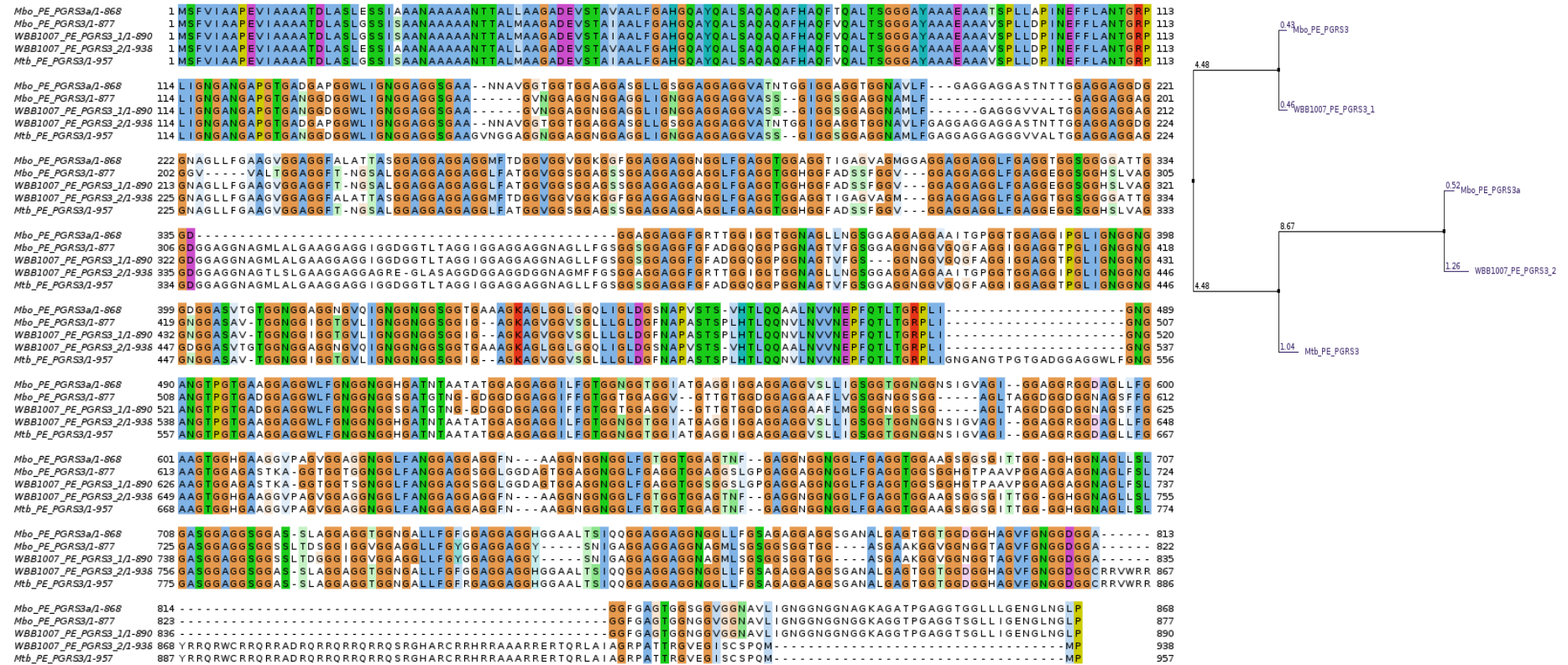


Figure S10. Protein sequence alignment of (from top to bottom) PE_PGRS3a from *M. bovis*, PE_PGRS3 from *M. bovis*, PE_PGRS3_1 from WBB1007 (*Mtb* L1), PE_PGRS3_2 from WBB1007 (*Mtb* L1) and PE_PGRS3 from H37Rv *Mtb*. Highlighted are the conserved residues across the different sequences. On the right, Neighbour joining tree using PID for the 5 sequences.

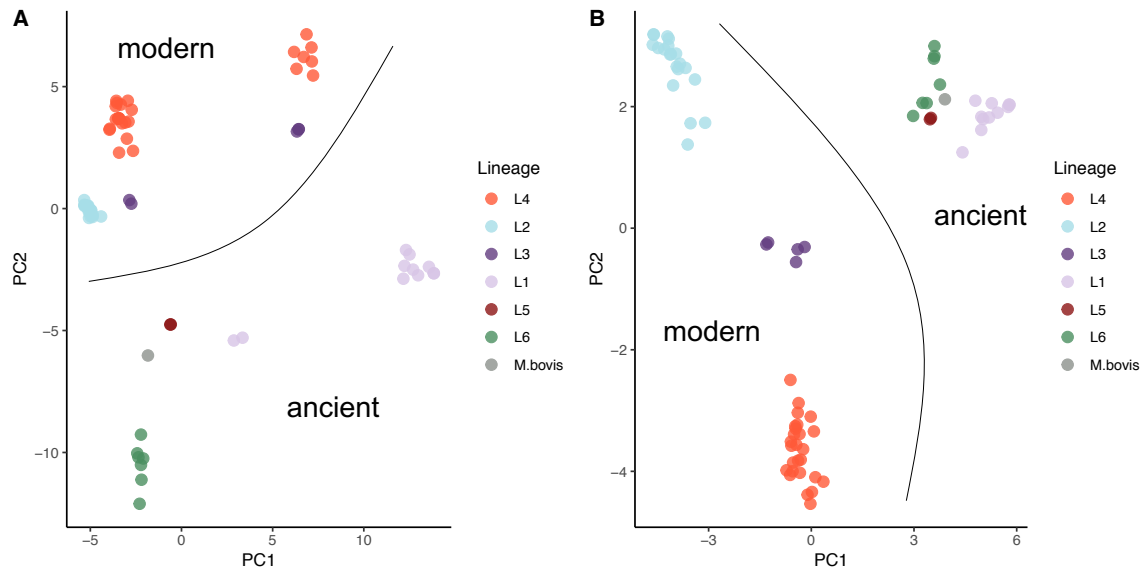


Figure S11. PCA of SNPs **(A)** and indels **(B)** with samples coloured by lineage.

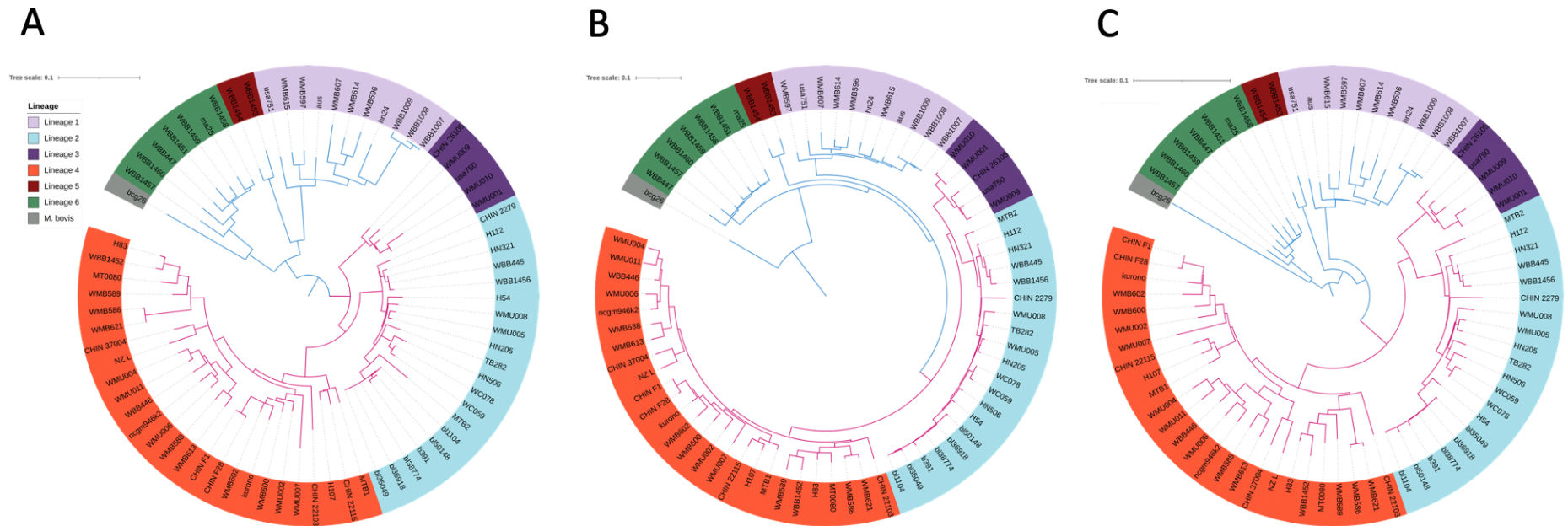


Figure S12. Maximum likelihood phylogenetic trees reconstructed with variants only in the *pe/ppe* genes as follows: (A) SNPs, (B) indels and (C) SNPs and indels.

CHAPTER 6

Portable sequencing of *Mycobacterium tuberculosis* for clinical and epidemiological applications

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	lsh1704009	Title	
First Name(s)	Paula Josefina		
Surname/Family Name	Gómez González		
Thesis Title	Analysis of Mycobacterium tuberculosis 'omics data to inform on loci linked to drug resistance, pathogenicity and virulence		
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?			
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.

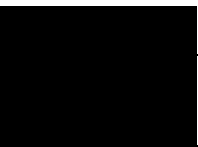
SECTION C – Prepared for publication, but not yet published


Where is the work intended to be published?	Briefings in Bioinformatics
Please list the paper's authors in the intended authorship order:	Gomez-Gonzalez, PJ; Campino, S; Phelan, JE; Clark, TG
Stage of publication	Submitted

SECTION D – Multi-authored work

<p>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)</p>	<p>I cultured and extracted DNA from clinical isolates. I received the long-read sequence data and performed the bioinformatic analysis, consisting in assembly, mapping, optimisation of variant calling process and phylogenetics. All statistical analysis and plotting were performed in R with custom scripts . I wrote the first draft of the manuscript and circulated to co-authors, and after receiving comments I edited the last version. I submitted the manuscript to the journal.</p>
---	---

SECTION E

Student Signature		
Date	28/01/2022	

Supervisor Signature		
Date	28/01/2022	

Portable sequencing of *Mycobacterium tuberculosis* for clinical and epidemiological applications

Paula J. Gómez-González¹

Susana Campino¹

Jody Phelan^{1,*}

Taane Clark^{1,2,*}

¹ Faculty of Infectious and Tropical Diseases, London School of Hygiene & Tropical Medicine, WC1E 7HT London, UK

² Faculty of Epidemiology and Population Health, London School of Hygiene & Tropical Medicine, WC1E 7HT London, UK

* Joint authors

** correspondence

taane.clark@lshtm.ac.uk

Department of Infection Biology,

Faculty of Infectious and Tropical Diseases

London School of Hygiene & Tropical Medicine, Keppel Street, London, UK

Briefings in Bioinformatics, Case Study

Abstract

With >1 million associated deaths in 2020, human tuberculosis (TB) caused by *Mycobacterium tuberculosis* bacteria remains one of the deadliest infectious diseases. A plethora of genomic tools and bioinformatic pipelines have become available in recent years to assist the whole genome sequencing of *M. tuberculosis*. The Oxford Nanopore Technologies (ONT) portable sequencer is a promising platform for cost-effective application in clinics, including to personalise treatment through detection of drug resistance associated mutations, or in the field, to assist epidemiological and transmission investigations. In this study, we performed a comparison of ten clinical isolates with DNA sequenced on both long-read ONT and (gold standard) short-read Illumina HiSeq platforms. Our analysis demonstrates the robustness of ONT variant calling for SNPs, despite the high error rate. Moreover, because of improved coverage in repetitive regions where short sequencing reads fail to align accurately, ONT data analysis can incorporate additional regions of the genome usually excluded (e.g., *pe/ppe* genes). The resulting extra resolution can improve characterisation of transmission clusters and dynamics, which is based on inferring closely related isolates. High concordance in variants in loci associated with drug resistance supports its use for rapid detection of resistant mutations. Overall, ONT sequencing is a promising tool for TB genomic investigations, particularly to inform clinical and surveillance decision making to reduce disease burden.

Word count: 216

Keywords: *Mycobacterium tuberculosis*, tuberculosis, sequencing, genomics, mutations

Introduction

Mycobacterium tuberculosis remains one of the deadliest single infectious agents, leading to 10 million human tuberculosis (TB) cases and 1.5 million associated deaths in 2020 [1]. The *Mycobacterium tuberculosis* complex is phylogeographically distributed in defined lineages that can determine the emergence of drug resistance, transmissibility, pathogenicity and host response, disease site and severity [2–4]. Drug resistant *M. tuberculosis* is one of the major threats to effectively control the disease, especially resistance to first-line rifampicin (RR-TB) and isoniazid; in combination, called multi-drug resistance (MDR-TB). MDR-TB accounted for around 150,000 cases in 2020 [1]. The acquisition of drug resistance in *M. tuberculosis* has been mainly attributed to spontaneous mutations, such as single nucleotide polymorphisms (SNPs) and small insertions and deletions (indels) in genes coding for drug-targets, drug-converting enzymes or involved in transport of small molecules such as efflux pumps [5, 6]. Phenotypic susceptibility testing is the traditional method to determine drug resistance; however, in combination with genome-wide association and convergent evolution studies, genetic variants conferring drug resistance have been validated enabling the use of genotypic methods to establish resistance through sequencing or nucleic acid amplification approaches [5]. Transmission events can be inferred through identification of variants in *M. tuberculosis* isolates sourced from different patients with (near) identical genomes [7]. Characterising the phylogeographic distribution of *M. tuberculosis* strains across regions can reveal outbreaks of more virulent strain-types, including Beijing strains [7].

Genome sequencing of *M. tuberculosis* has gained traction for both clinical and epidemiological investigations. These applications have provided insights into mutations underlying drug resistance, circulating strain-types and virulence, and transmission dynamics,

thereby with the potential to inform clinical and surveillance activities. New genomic tools allow for whole genome sequencing (WGS) with increasing opportunities to use it directly from sputum [8]. Together with new analysis methods, WGS data can be used to profile the bacteria for drug resistance [5, 9, 10], characterise ancient and modern lineages and different strain-types [11], and establish who may have transmitted to whom; thus allow targeted resources to hotspot areas to reduce transmission [12]. These genomic insights are facilitated through advances in health informatics [13].

WGS opportunities are set to revolutionise the diagnosis and clinical management of TB patients, with routine pathogen genetic characterisation applied in the UK healthcare system. Building on this success and recent COVID-19 experience, an increasing number of countries worldwide are seeking to adopt genomics as part of clinical care [14]. However, to be effective for global disease control and maximise impact, NGS platforms need to be applied in high TB burden settings, which may be resource poor. To achieve the economies of scale and cost reductions in these settings, it is possible to target a high number of genes (*e.g.*, drug resistance loci) across many samples using an amplicon-based approach on next-generation sequencing (NGS) platforms, or focus on the multiplexing of whole genomes if transmission is important.

Compared to other pathogen genomes, the *M. tuberculosis* genome (size 4.4 Mbp) is relatively clonal with no horizontal gene transfer, but was historically challenging to sequence due to its high GC content and repetitive nature [15]. The Illumina sequencing platform with its paired short reads and low error rates has been employed successfully to analyse almost the entire genome, including drug resistance loci [10], with highly variable and GC-rich *pe/ppc*

genes often excluded due to the difficulties in accurately mapping these repetitive regions [16–18]. Recently, sequencing platforms with long reads (> 1 kbp) have been applied for the construction of reference genomes and analysis of methylated base modifications [19], but are too costly for implementation as a high throughput tool. Our previous work compared the application of the Illumina MiSeq, Ion Torrent PGM™ [15] and Oxford Nanopore Technologies (ONT) platforms [13]. We observed higher sequencing error rates on the ONT platforms, but sufficient coverage to call drug resistant variants [13]. The ONT sequencing platform is portable enabling the characterisation of *M. tuberculosis* in remote and field settings, and has the potential to perform multiplexing of samples, leading to cost reductions. Future cost-effectiveness is likely by informed decision making in clinics through personalisation of treatments in drug resistance settings, as well as by determining geographical regions for the optimal targeting of TB surveillance and control activities. To assess the viability of the ONT platform for these applications, we apply the technology to DNA extracted from *M. tuberculosis* isolates. In a paired analysis, we compare the resulting WGS sequence data to those generated on an Illumina platform, finding high concordance in variant calls between methods and including regions traditionally excluded in our analysis, such as *pe/ppa* genes.

Results

Coverage

ONT long-reads and Illumina short-reads were generated from the sequencing of replicate DNA of ten clinical isolates originally sourced from Malawi (labelled S1-10; **Table S1**). These isolates covered lineages 1 (L1: 1.1.2, n=1; 1.1.3.2, n=1), 2 (L2: Beijing 2.2.1, n=3), 3 (L3: n=4), and 4 (L4: 4.9, n=1). Sequencing with the ONT platform yielded a median of 67,939 reads per sample, with a median read length of 3,806 bp. Illumina data (median number of reads:

1,687,571; read length: 75-100 bp) was generated for the same samples. Mapping to the reference genome (H37Rv GCA_000195955.2) led to high depth of coverage for all samples (average depth of coverage: Illumina 93.6-fold, ONT 72.2-fold) (**Table 1**). For all samples, median coverage normalised by four housekeeping genes (*gyrB*, *gyrA*, *rpoB*, *rpoC*) was investigated genome-wide (**Figure 1A**). Overall, across sample pairs and sequencing platforms, there was high normalised read depth with medians above 0.75 (**Figure 1B**). Normalised coverage levels in ONT data below 0.5 coincided with lineage-specific deleted regions, including known regions of difference (*e.g.*, RD152 in lineage 2). The presence of these deletions in specific lineages was independently validated in high quality PacBio whole-genome assemblies [19].

Through mapping of the ONT data against a representative PacBio assembly for each lineage, high normalised coverage was achieved genome wide. There were several peaks with normalised coverage below 0.5 belonging to insertion sequences (*e.g.*, IS6110) or deleted genes in specific strains (*e.g.*, RD152 region in sample S1) (**Figure S1**). Overall, these results suggest that ONT technology has performed well, including in repetitive regions. The genes with the lowest coverage in Illumina data were mostly *pe/ppc* genes, whose mapping accuracy with short-reads is known to be low [17, 18], due to their high GC content and repetitive regions. For the 85 *pe/ppc* genes thought to be non-conserved harbouring structural variants that disrupt their protein sequences [16], there was lower sequencing coverage in Illumina compared to ONT data (T-test adjusted $P < 0.001$, **Figure 1C**).

Although aligning to a lineage specific reference improved the coverage for Illumina (**Figure S1**), extreme GC content disproportionately reduced coverage in short-read compared to

long-read data (**Figure 1D**). Genes with the lowest average values of normalised coverage in Illumina data, had higher coverage in ONT data (T-test $P < 0.001$, **Figure S2**). Two genes had greater coverage in Illumina compared to ONT data in L2 and L3 sample pairs, coinciding with an insertion sequence (*Rv0797*) and a conserved hypothetical protein (*Rv1765c*). The latter belongs to RD152, which was deleted in all L2 isolates and one L3. However, due to the high similarity (97%) between *Rv1765c* and *Rv2015c* sequences, the Illumina platform seems to not capture the deletion.

Variant calling

Variants were called using Freebayes software retaining all sites where at least one sample had $> 50\%$ alternate reads, leading to 9,052 unique positions. For the analysis, Illumina variants with an allele depth fraction of at least 0.7 were considered as true variants. Due to the high error rate of ONT sequencing, almost all positions at which a true variant exists contain a mixture of alternate and reference alleles. To find the optimum cut-off which balances the sensitivity (true positive rate) and specificity (true negative rate), alternate-allele proportions for each site in the ONT replicates were compared to their Illumina counterparts. An optimum alternate-allele proportion value of 0.7 was chosen, keeping the true positive rate $> 97\%$ and true negative rate $> 91\%$, and the false positive rate $< 1\%$ (**Figure S3**). After refining genotype calls using the chosen minimum alternate frequency of 0.7 and removing repetitive and poorly covered regions in Illumina alignments, a final filtered dataset of 3,955 SNPs covering $> 89\%$ of the genome was retained for subsequent analysis (see **Figure S4**). The chosen frequency cut-off of 0.7 was validated using ONT sequence data for four replicates of the H37Rv reference strain [20]. After implementing the pipeline above, there was high

concordance between the H37Rv replicates, with only 4 discrepancies found among the 29 SNPs identified.

The concordance of SNPs and small indels detected by ONT and Illumina data was assessed. For all pairs, > 99% of the total SNPs identified were called in both samples, showing few combined platform discrepancies (median 3.5; range: 0-9 SNPs) (**Table 2**). Agreement between platforms for depth of coverage and alternate frequencies was assessed. Good coverage in Illumina coincided with good coverage in ONT, and the alternate frequencies were observed to be lower in ONT than Illumina, suggesting the noisier nature of the ONT technology (see **Figure S5**). Most discrepancies arose in the few SNPs called in Illumina but not in ONT data, due to alternate frequency values just below the 0.7 allele depth cut-off (see **Table S2**). In addition, every sample except S5 (L4.9) differed in the call for the (H37Rv) genomic position 55,553. This position is in a GC-rich region where ONT data had a CCG insertion followed by a nucleotide change (C -> T), whereas the variant called in Illumina data only included the SNP. The multiple CCG repeats present in the sequence leads to the Illumina data analysis not capturing the insertion. Additionally, ONT data for sample S10 showed a SNP in a GC-rich region whilst in its Illumina counterpart it was identified as a 1 bp insertion followed by the SNP, suggesting an error in the ONT call.

The majority (>87%) of small indels called at an alternate frequency of 0.7 were correctly captured by both platforms (**Table 2**). However, more discrepancies were identified with small indels than with SNPs (median 9; range 4 – 12 small indels). These discrepancies were mostly driven by small indels (1 bp) in polyC/polyG repeats which were called from Illumina but not in ONT sequence data (see **Table S3**). On the other hand, the second type of calls in

ONT that differed from Illumina were larger indels (8-10bp), in which the allele depth fraction in Illumina was slightly lower than 0.7, suggesting that these larger variants called by ONT were not spurious (see **Table S4**). Larger structural variants (>15bp) were investigated with Delly software. Long-reads allow more accurate identification of large indels. As expected, a higher number of large variants were observed in ONT (median 81) compared to Illumina (median 24) data (**Table 2**), with deletions having the highest agreement between platforms (pairwise sample overlap: median 17, range 2 – 20 large indels).

Strain typing and phylogenetics

Lineage prediction was performed by TB-Profiler using the 3,955 high quality SNPs covering >89% of the genome, and consistency between pairs was assessed. All predictions were found to be identical between Illumina and ONT platforms confirming the robust nature of the variant calling process (**Table 1**). To further investigate the use of the ONT platform to perform clustering, phylogenetic reconstruction was performed using IQ-TREE software. Clear clustering of strain-types was observed with long internal branches separating each major lineage. In addition, each sample pair formed a monophyletic clade with short terminal branch lengths indicating the near identical pattern of variation detected through both platforms (**Figure 2; Figure S6**). Two and three samples belonging to L2 (S8, S9) and L3 (S2, S3, S4) respectively were closely related, with the number of SNP differences below or equal to 20.

To increase the accuracy of the phylogenetic reconstruction, potentially for transmission analysis, base-calls were manually curated and SNPs which were called as reference with alternate depth frequencies between 0.6 and 0.7 were redesignated as alternate base calls.

Following this, the reconstruction of the phylogenetic tree with only ONT isolates was performed using the 3,955 polymorphic sites (**Figure 2; Figure S6**). Samples within a putative L2 transmission cluster (S8 and S9) differed by 2 SNPs, whilst the distance within the L3 transmission cluster (S2, S3, S4) varied between 2 and 18 SNPs. Characterisation of transmission chains is of epidemiological importance, and due to the small numbers of variants that sometimes separate closely related isolates, accurate estimation of the number of SNPs differences between samples is crucial. Previous studies have shown how long-read sequencing solves some of the traditional Illumina blind spots [21], including by the successful assembly and variant calling of *pe/ppe* genes with ONT data [22]. On this basis, 150 out of 169 *pe/ppe* genes with good coverage (> 0.7 normalised mean coverage) were included to complement the genomic regions analysed and therefore potentially achieve a deeper separation of the transmission clusters. These regions overlapped with previous studies [16, 23]. An extra 568 high quality SNPs were added, resulting in one extra SNP within the transmission cluster from L2 (S8, S9) and four extra SNPs for L3 (S2, S3, S4), thereby slightly increasing the differences obtained within highly similar samples (**Figure 2C**).

Drug resistance prediction

Drug resistance profiles were predicted by TB-Profiler using the filtered set of 3,955 SNPs. Predictions were compared across replicates and matched perfectly between platforms, leading to nine pan-susceptible isolates and one pre-MDR isolate. In addition, identical variants were found across the 42 genes analysed by TB-Profiler. Drug susceptibility test data was used to confirm these predictions with all matching, except one (**Table S1**). One inconsistency was observed in the pre-MDR isolate (sample S5), where although isoniazid resistance was genotypically and phenotypically concordant (*katG* S315T present in both ONT

and Illumina data), streptomycin resistance was observed through drug susceptibility testing but not in the genotypic prediction. Upon further inspection of non-associated variants in streptomycin resistance genes in isolate S5, a premature stop codon was observed (in both Illumina and ONT data) in *gid* (S136*), which is the likely explanation of the discrepancy between phenotypic and genotypic predictions.

Discussion

The benefits of using whole-genome sequencing (WGS) technologies in clinical and epidemiological settings, such as the characterisation of transmission networks, or for detection of drug resistance associated mutations to inform on treatment decisions, have been described [12, 13]. Nevertheless, the associated costs of WGS can limit their application, especially in remote, field or resource-poor settings. The recent development of portable sequencing devices powered from laptops, such as ONT MinION, are significantly reducing the costs and infrastructure necessary for sequencing, thereby improving accessibility [24, 25]. This accessibility would be useful for infection control in the high TB transmission setting of the Karonga District, Malawi, the source of our samples. In parallel, the possible direct sequencing from sputum samples has been successfully reported, taking up to 5 days [8, 24, 26], which will shorten the time from specimen collection to a drug resistance profile, leading to timely and personalised treatment that can be significantly delayed when culture isolation is required (up to 3 weeks).

To assess the performance of Illumina short-read and ONT long read platforms, we have carried out a comparative analysis of ten sample pairs with data from both technologies. Illumina technology with a low sequencing error rate is considered the gold standard, and

therefore has been applied to inform on drug resistance or transmission, but the performance of ONT, with its known higher error rate, is less clear. Several studies have evaluated the performance of ONT sequencing in target-sequencing approaches for drug resistance detection [26–28], finding good concordance between Illumina and ONT, or in WGS analysis [29]. For ONT sequencing data, an even coverage distribution along the chromosome was observed, with drops coinciding with deleted genes or regions, such as RD152 (*Rv1758c-Rv1765c*) in L2, or insertion sequences, whose presence/absence is variable among different strains. Coverage levels were not dependent on GC content, with high values even in the extremely GC-rich genes (> 80% GC content). Using a lineage specific genome as reference yielded an expected overall improvement in coverage across both platforms. However, Illumina replicates of L3 isolates still failed to reach similar values to those of ONT in the high GC content regions, revealing the higher susceptibility of the short-read sequencing platform to GC-rich genes. Blind spots for Illumina sequencing technologies have been previously reported [18], for which long-read sequencing technologies can assist [21, 22]. In accordance with previous studies [22], our work demonstrates that long-read data has the potential to elucidate complex regions, such as *pe/ppe* genes, which due to their GC-rich and repetitive nature have been systematically excluded from WGS analysis, losing potential phylogenetic information. Coverage of the Illumina replicates on these regions, and more specifically in the most diverse genes of these two families, was shown to be significantly lower than their ONT counterparts, suggesting a potential inclusion of these genes for the downstream analysis in WGS from ONT. This could assist with understanding the genetic diversity of *pe/ppe* genes, whose functions are largely still unknown, but some are involved in host-pathogen interactions and thereby promising targets for vaccine development [16].

The performance of the variant calling pipeline for ONT sequences was investigated and compared to the Illumina data, considering the latter as a gold standard. The ONT platform is prone to sequencing errors, whereas Illumina high sequencing accuracy makes it preferred for identification of SNPs and small indels [15]. In contrast, larger structural variants are difficult to capture with short-reads, thus applying a hybrid approach involving assembly of long-reads with correction using short-reads can improve the accuracy and completeness of variant detection. For the evaluation of the variant calling method in ONT data, an alternate allele depth fraction ≥ 0.7 was established as the optimum cut-off based on the true and false positive error rates. The exclusion of repetitive regions (*e.g.*, *pe/ppa* genes) led to good agreement between platforms for SNPs and small indels, as previously shown in other studies [26], with discrepancies often being found at an allele depth between 0.6 and 0.7, suggesting the potential use of the lower cut-off of 0.6 to include more true positive calls. With SNPs covering more than 89% of the genome, an accurate phylogenetic reconstruction was obtained, supporting the utility of ONT for variant identification and lineage profiling. Moreover, the inclusion of 150 *pe/ppa* genes with high levels of coverage, which would normally be among the regions excluded, added extra variants that have the potential of being phylogenetically informative. The possibility of including extra variants may lead to an improved resolution that would be of special interest in outbreak settings, where transmission analysis of closely related isolates can be potentially better established.

One of the most important applications of the ONT MinION portable device is the accurate detection of drug resistant variants, which can inform and assist patient management in a timelier manner than traditional phenotypic tests. A promising cost-effective approach to the high throughput evaluation of drug resistant loci in clinical isolates is target-amplicon

sequencing [30]. We validated the high quality of the variant calling process on ONT data for 42 known *M. tuberculosis* drug resistant loci, finding congruent results with their Illumina counterparts. This outcome suggests the potential identification of drug-resistant variants from ONT data, including within a target-amplicon framework.

Limitations of the study include the low number of isolates analysed, the low intra-lineage diversity, and limited number of drug resistant isolates. Whilst the latter may limit the investigation of variants in drug resistance associated loci, given the error rate of ONT including within these loci, our approach robustly characterises the sequence of drug resistance genes and it is thus reasonable to conclude that it will also accurately characterise the sequence of genes that contain variants and, by extension, predict resistance. Previous works have shown good drug resistance variant detection through different methods [26, 29], with promising results towards its use for diagnostics purposes in the clinic. However, for the complete reliance of *in silico* drug resistance prediction based on genotypes, an improved understanding of catalogue of resistance mutations is essential. A more complete characterisation of phenotype-genotype associations for certain drugs are required and the phenotypic-genotypic inconsistency observed in this analysis reflects this need. WGS facilitates a more comprehensive analysis compared to targeted gene sequencing. The use of long-reads can cover repetitive regions of the genome, and thereby help elucidate compensatory or epistatic mutations that could be crucial for the better understanding of drug resistance mechanisms in *M. tuberculosis*.

In conclusion, the data obtained through this analysis supports the use of ONT sequencing platforms for well established drug resistance variants detection and phylogenetic

reconstruction, with potential application in transmission analysis, since the underpinning SNP variant calling process appears robust. However, due to the high error rate, Illumina remains the best option for small indel analysis, suggesting, for their accurate study with ONT data, a hybrid correcting approach is warranted. Moreover, we demonstrate the possibility of including additional genomic regions in the standard variant calling pipelines, such as the *pe/ppe* genes, which due to their implications in pathogenicity and host-pathogen interactions could give insights into epidemiological implications, as well as potentially improving the resolution of transmission clusters. Furthermore, for variants in more complex gene arrangements that might fail to be captured using the H37Rv reference, the use of lineage-specific reference genomes could be practical. The portable MinION technology could therefore be implemented and is likely to gain traction for epidemiological, phylogenetic, or drug resistance detection applications, providing much needed assistance in the control of tuberculosis, especially in high burden settings where impacts will be greater.

Methods

Culture, DNA extraction and sequencing

The 10 isolates analysed in this study were sourced from TB patients in Karonga (Malawi) between 2001 and 2009, with isolates stored at the LSHTM. The bacterial culture and extraction of genomic DNA was carried out at the LSHTM Biosafety Level 3 containment facility. Briefly, *M. tuberculosis* isolates were pre-cultured in Middlebrook 7H9 supplemented with 0.05% Tween 80 and 10% albumin-dextrose-catalase (ADC) at 37°C to mid-log phase. Once reached the exponential growth, they were passaged to roller bottles until desired optical density (OD = 0.6 – 0.8). Heat inactivation (one hour at 80°C) followed by CTAB-chloroform-isoamyl alcohol method was used for genomic DNA extraction [31]. Whole-

genome sequencing of DNA samples was performed with Oxford Nanopore Technologies (ONT) (MinION Flow Cell with R10.3 nanopore chemistry; SQK-LSK109 ligation-based sequencing kit) and Illumina HiSeq 4000 (150bp paired-end) platforms through The Applied Genomics Centre at LSHTM. A further set of four DNA replicates for the reference H3Rv strain also underwent sequencing using the ONT MinION platform. All raw sequencing data is available (see **Table S1** for accession numbers).

Bioinformatics pipeline

Base calling of ONT raw sequence data was performed with bonito basecaller (model dna_r9.4.1_e8.1_sup@v3.3) [32] and reads aligned to the H37Rv reference genome (GCA_000195955.2) using minimap2 (v2.17-r941) software [33] discarding ambiguous reads. Depth of coverage along the chromosome and median coverage per annotated gene was calculated with BEDTools (v2.29.2) [34], using the alignments of data obtained by ONT and Illumina platforms. To compare between samples, median coverage per gene per sample was normalised by the coverage of four housekeeping genes (*gyrB*, *gyrA*, *rpoB*, *rpoC*) known to not be deleted or duplicated and expected to have a good “average” coverage. Lineage-specific reference genomes were selected among high quality PacBio assemblies [19] and used to assess levels of coverage. Due to the high error rate of the ONT platform, a mixture of alternate and reference alleles is often found. In order to identify an optimum cut-off for variant calling, a minimum alternate allele frequency of 0.5 was used in the variant calling process carried out using Freebayes (v1.3.2) software [35]. Variant calls obtained in Illumina data with an allele frequency of 0.7 were considered as true variants. Alternative allele frequency cut-off values of 0.5, 0.6, 0.7, 0.8, 0.9 and 1.0 for ONT variant calls were used and true and false positive and negative rates for each of the cut-offs were calculated. True

positive and false positive rates were compared and evaluated using a receiver-operator characteristic curve analysis. A final cut-off of 0.7 was determined to perform variant calling, and validated using ONT data from the H37Rv replicates.

To obtain a curated set of SNPs for the subsequent analysis, variants were filtered (see **Figure S4**). In brief, regions with repetitive sequences that generate mapping problems (see GitHub repository https://github.com/pgomezgonzalez/nanopore_tb_data_analysis), such as *pe/ppe* genes or insertion sequences, were excluded, and only SNPs were selected. Genotype calls were refined by read depth (DP) and alternate allele depth (AD) fraction, with a minimum DP of 10 required to determine a position and an $AD \geq 0.7$ needed to retain the alternate call. The resulting refined SNP dataset was used for the agreement evaluation between sample pairs and their phylogenetic reconstruction. Small indels called using Freebayes (v1.3.2) were filtered using the same pipeline as SNPs. Delly (v0.8.7) software [36] was used for large structural variants (indels with size > 15 bp). Lineage and drug resistance profiling of the sample pairs was carried out with TB-Profiler (v3.0, commit version: de4e796) [13]. Maximum likelihood phylogenetic reconstruction of the genomes was performed with IQ-TREE (v1.6.12) with a GTR+G+ASC nucleotide substitution model [37] by using genome-wide SNPs excluding repetitive regions or including the 150 *pe/ppe* genes with good coverage, and visualised together with annotations in iTOL software. Custom scripts used in the analysis pipeline are available in a GitHub repository (https://github.com/pgomezgonzalez/nanopore_tb_data_analysis).

Data availability

Raw sequencing data is available from the ENA archive (see **Table S1** for a list of accession numbers).

Ethics approval and consent to participate

The studies were approved by the Health Sciences Research Committee in Malawi (#424) and by the LSHTM ethics committee (#5067). Informed written consent was sought and obtained for all patients in the original study.

References

1. World Health Organization (WHO). Global Tuberculosis Report 2021. 2021.
2. 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.
3. 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;4:1–10.
4. Coscolla M, Gagneux S. Consequences of genomic diversity in *Mycobacterium tuberculosis*. *Semin Immunol.* 2014;26:431–44.
5. Coll F, Phelan J, Hill-Cawthorne GA, Nair MB, Mallard K, Ali S, *et al.* Genome-wide analysis of multi- and extensively drug-resistant *Mycobacterium tuberculosis*. *Nat Genet.* 2018;50:307–16.
6. Villellas C, Coeck N, Meehan CJ, Lounis N, de Jong B, Rigouts L, *et al.* Unexpected high prevalence of resistance-associated *Rv0678* variants in MDR-TB patients without documented prior use of clofazimine or bedaquiline. *J Antimicrob Chemother.*

2016;72:dkw502.

7. 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. *Microb Genomics*. 2020;6.

8. 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. *J Clin Microbiol*. 2018;56:1–11.

9. 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 Med*. 2015;7:51.

10. Gómez-González PJ, Perdigao J, Gomes P, Puyen ZM, Santos-Lazaro D, Napier G, *et al*. Genetic diversity of candidate loci linked to *Mycobacterium tuberculosis* resistance to bedaquiline, delamanid and pretomanid. *Sci Rep*. 2021;11.

11. 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.

12. Guerra-Assunção J, Crampin A, Houben R, Mzembe T, Mallard K, Coll F, *et al*. Large-scale whole genome sequencing of *M. tuberculosis* provides insights into transmission in a high prevalence area. *Elife*. 2015;4:1–17.

13. 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.

14. Ferdinand AS, Kelaher M, Lane CR, da Silva AG, Sherry NL, Ballard SA, *et al*. An

implementation science approach to evaluating pathogen whole genome sequencing in public health. *Genome Med.* 2021;13:1–11.

15. Phelan J, O’Sullivan DM, Machado D, Ramos J, Whale AS, O’Grady J, *et al.* The variability and reproducibility of whole genome sequencing technology for detecting resistance to anti-tuberculous drugs. *Genome Med.* 2016;8:132.

16. 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.

17. 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. *Nat Rev Microbiol.* 2019;17:533–45.

18. Modlin SJ, Robinhold C, Morrissey C, Mitchell SN, Ramirez-Busby SM, Shmaya T, *et al.* Exact mapping of Illumina blind spots in the *Mycobacterium tuberculosis* genome reveals platform-wide and workflow-specific biases. *Microb Genomics.* 2021;7.

19. Gomez-Gonzalez PJ, Andreu N, Phelan JE, de Sessions PF, Glynn JR, Crampin AC, *et al.* An integrated whole genome analysis of *Mycobacterium tuberculosis* reveals insights into relationship between its genome, transcriptome and methylome. *Sci Rep.* 2019;9:1–11.

20. 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.

21. Elghraoui A, Modlin SJ, Valafar F. SMRT genome assembly corrects reference errors, resolving the genetic basis of virulence in *Mycobacterium tuberculosis*. *BMC Genomics.* 2017;18:302.

22. Bainomugisa A, Duarte T, Lavu E, Pandey S, Coulter C, Marais BJ, *et al.* A complete high-

- quality MinION nanopore assembly of an extensively drug-resistant *Mycobacterium tuberculosis* Beijing lineage strain identifies novel variation in repetitive PE/PPE gene regions. *Microb Genomics*. 2018;4.
23. Marin M, Vargas R, Harris M, Jeffrey B, Epperson LE, Durbin D, *et al*. Benchmarking the empirical accuracy of short-read sequencing across the *M. tuberculosis* genome . *Bioinformatics*. 2022;38:1781–7.
24. Votintseva AA, Bradley P, Pankhurst L, Del C, Elias O, Loose M, *et al*. Same-Day Diagnostic and Surveillance Data for Tuberculosis via Whole-Genome Sequencing of Direct Respiratory Samples. *J Clin Microbiol*. 2017;55:1285–98.
25. Runtuwene LR, Tuda JSB, Mongan AE, Suzuki Y. On-Site MinION Sequencing. In: Suzuki Y, editor. *Advances in Experimental Medicine and Biology*. Springer New York LLC; 2019. p. 143–50.
26. Mariner-Llicer C, Goig GA, Zaragoza-Infante L, Torres-Puente M, Villamayor L, Navarro D, *et al*. Accuracy of an amplicon-sequencing nanopore approach to identify variants in tuberculosis drug-resistance- associated genes. *Microb Genomics*. 2021;7.
27. Tafess K, Ng TTL, Lao HY, Leung KSS, Tam KKG, Rajwani R, *et al*. Targeted-Sequencing Workflows for Comprehensive Drug Resistance Profiling of *Mycobacterium tuberculosis* Cultures Using Two Commercial Sequencing Platforms: Comparison of Analytical and Diagnostic Performance, Turnaround Time, and Cost. *Clin Chem*. 2020;66:809–20.
28. Chan WS, Au CH, Chung Y, Leung HCM, Ho DN, Wong EYL, *et al*. Rapid and economical drug resistance profiling with Nanopore MinION for clinical specimens with low bacillary burden of *Mycobacterium tuberculosis*. *BMC Res Notes*. 2020;13:1–7.
29. Smith C, Halse TA, Shea J, Modestil H, Fowler RC, Musser KA, *et al*. Assessing nanopore sequencing for clinical diagnostics: A comparison of Next-Generation Sequencing (NGS)

methods for mycobacterium tuberculosis. J Clin Microbiol. 2021;59:1–14.

30. Gliddon HD, Frampton D, Munsamy V, Heaney J, Pataillot-Meakin T, Nastouli E, *et al.* A Rapid Drug Resistance Genotyping Workflow for *Mycobacterium tuberculosis*, Using Targeted Isothermal Amplification and Nanopore Sequencing. Microbiol Spectr. 2021;9:1–12.

31. Somerville W, Thibert L, Schwartzman K, Behr MA. Extraction of *Mycobacterium tuberculosis* DNA: A question of containment. J Clin Microbiol. 2005;43:2996–7.

32. Xu Z, Mai Y, Liu D, He W, Lin X, Xu C, *et al.* Fast-bonito: A faster deep learning based basecaller for nanopore sequencing. Artif Intell Life Sci. 2021;1 November:100011.

33. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34:3094–100.

34. Quinlan AR, Hall IM. BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841–2.

35. Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing. 2012;:1–9.

36. 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;28:333–9.

37. Nguyen LT, 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.

Acknowledgements

PJG-G is funded by an MRC-LID PhD studentship. JEP is funded by a Newton Institutional Links Grant (British Council, no. 261868591). TGC was funded by the Medical Research Council UK (Grant no. MR/M01360X/1, MR/N010469/1, MR/R025576/1, and MR/R020973/1) and BBSRC (Grant no. BB/R013063/1). SC was 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

SC, JEP and TGC conceived and directed the project. PJG-G undertook sample processing and DNA extraction. PJG-G performed bioinformatic and statistical analyses under the supervision of SC, JEP and TGC. PJG-G, SC, JEP and TGC interpreted results. PJG-G 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 manuscript. PJG-G, JEP, and TGC compiled the final manuscript.

Additional Information

Table S1. ENA accession number for study samples undergoing sequencing using Illumina and Oxford Nanopore Technologies platforms.

Table S2. Discrepancies between Illumina and Oxford Nanopore Technologies SNP calls.

Table S3. Discrepancies between Illumina and Oxford Nanopore Technologies indel calls.

Table S4. Large structural variants identified in Illumina and Oxford Nanopore Technologies sequence data.

Figure S1. Genome-wide normalised coverage.

Figure S2. Correlation of normalised coverage between Illumina and Oxford Nanopore Technologies platforms.

Figure S3. Receiver-operator characteristic curve for the error rate of Oxford Nanopore Technologies data.

Figure S4. Analysis pipeline.

Figure S5. Depth of coverage and alternate allele depth fraction correlation between Illumina and Oxford Nanopore Technology for SNPs called in both platforms.

Figure S6. Cladogram of Oxford Nanopore Technology and Illumina sequenced isolates.

Competing interests

No potential conflict of interest was reported by the authors.

Key points

- Robust variant calling following Oxford Nanopore Technologies sequencing.
- Suitability of Oxford Nanopore Technology sequencing to detect variants in drug resistance associated loci.
- Enhanced transmission analysis by deeper resolution from long-read sequence data.

Table 1. Summary of ten sample pairs (S1-S10) sequenced using Illumina and Oxford Nanopore Technology (ONT) platforms.

Sample	Lineage	Platform	Mean read length	Number of reads	% reads mapped	Mean depth	No. SNPs*
S1	3	ONT	4,496	97,949	95.77	94	1144
		Illumina	100	2,000,955	99.39	78	1146
S2	3	ONT	5,421	75,742	95.79	87	1154
		Illumina	100	1,593,992	99.52	67	1157
S3	3	ONT	4,204	113,137	97.45	102	1158
		Illumina	75	11,239,186	99.32	251	1160
S4	3	ONT	4,784	72,196	96.49	74	1156
		Illumina	75	6,929,436	99.31	152	1158
S5	4.9	ONT	6,958	46,188	94.49	69	259
		Illumina	100	1,320,558	99.78	55	259
S6	1.1.2	ONT	4,997	60,416	95.63	64	1741
		Illumina	100	2,116,280	99.35	90	1746
S7	1.1.3.2	ONT	4,411	75,431	96.81	72	1763
		Illumina	100	1,127,055	99.22	48	1772
S8	2.2.1	ONT	4,296	63,528	96.98	59	1154
		Illumina	100	1,334,916	99.53	55	1158
S9	2.2.1	ONT	5,395	43,239	97.29	51	1154
		Illumina	100	1,781,150	99.46	76	1158
S10	2.2.1	ONT	3,468	63,682	97.78	48	1115
		Illumina	100	1,510,044	99.57	65	1119

* High quality SNPs obtained at an alternate frequency of 0.7

Table 2. Concordance of variants found using Illumina and Oxford Nanopore Technology (ONT) platforms.

Sample pair	SNPs			Small indels			Large structural variants*		
	ONT only	Illumina only	Both	ONT only	Illumina only	Both	ONT only	Illumina only	Both
S1	0	2	1144	3	4	94	58	9	20
S2	0	3	1154	3	8	88	64	6	17
S3	0	2	1158	5	7	84	62	6	14
S4	0	2	1156	4	7	84	66	4	14
S5	0	0	259	2	2	28	14	0	4
S6	0	5	1741	0	9	115	67	5	20
S7	0	9	1763	3	5	108	61	6	19
S8	0	4	1154	3	6	97	68	8	14
S9	0	4	1154	2	7	95	67	9	16
S10	1	5	1114	2	5	97	72	12	18

* includes insertions and deletions (indels) > 15 bp

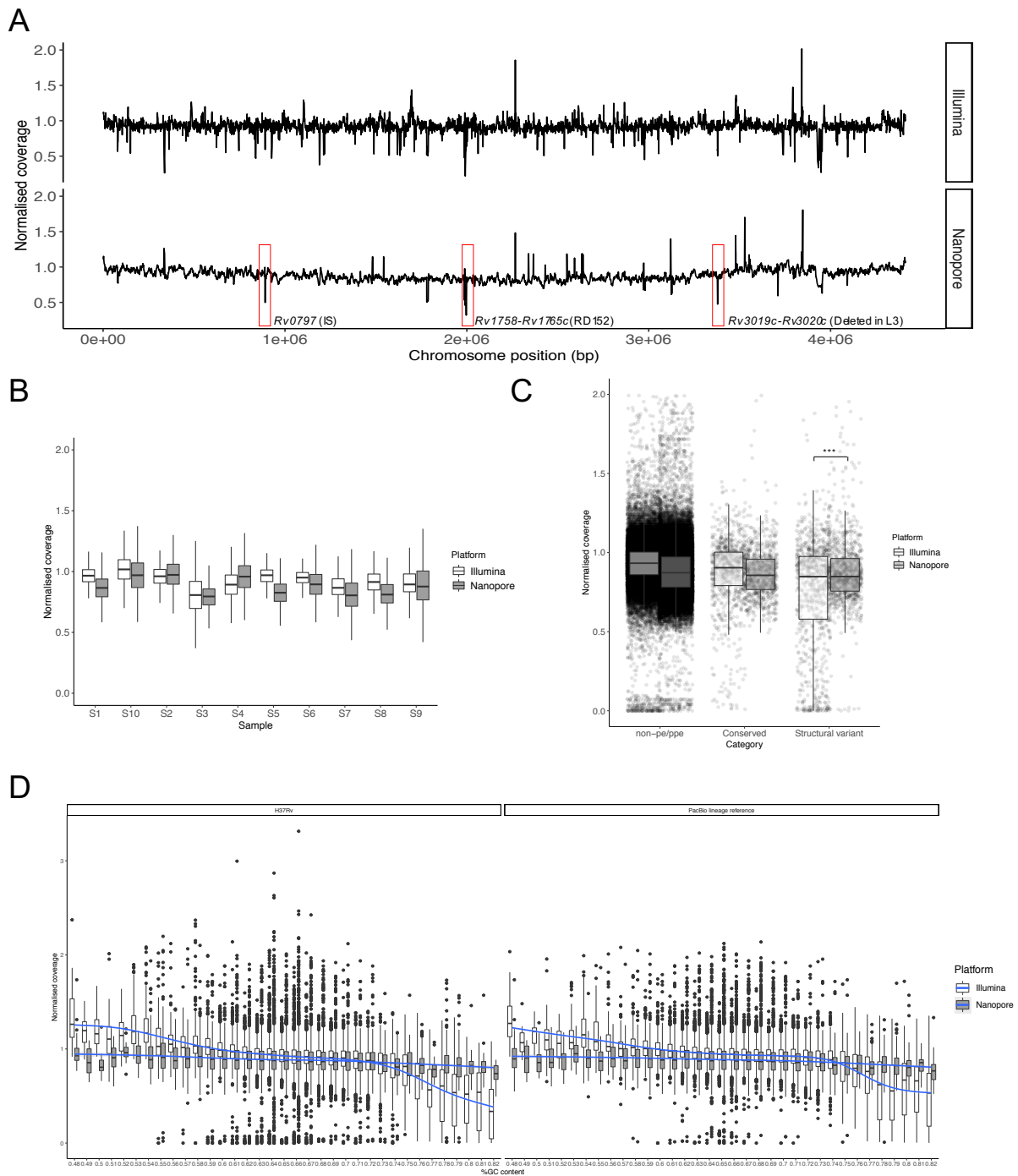


Figure 1. Coverage analysis for Illumina and Oxford Nanopore Technologies (ONT) data

Coverage analysis for ONT and Illumina data across the 10 sample pairs (S1-S10). **(A)** Average median normalised coverage along the chromosome across all samples for both technologies (top Illumina, bottom ONT). Genes with average median coverage < 0.5 for ONT platform are annotated: *Rv0797* corresponds to an insertion sequence; *Rv1758-Rv1765c* corresponds

to RD152, deleted in L2 and one isolate from L3; and *Rv3019c-Rv3020c* is a genomic region deleted in L3 isolates. The vertical axis shows the median coverage normalised by four house-keeping genes. The horizontal axis shows the position along the chromosome aligned to H37Rv. **(B)** Boxplots of normalised coverage per gene per sample for the 10 pairs. **(C)** Normalised coverage per gene per sample by group as follows: non-*pe/ppe* genes, conserved *pe/ppe* genes and *pe/ppe* genes with structural variants; *** adjusted P value < 0.001. **(D)** Normalised coverage distribution per sample per gene by GC content for each sequencing platform. On the left, coverage obtained aligning to H37Rv reference; on the right coverage obtained aligning to PacBio lineage-specific assemblies. PCA of SNPs.

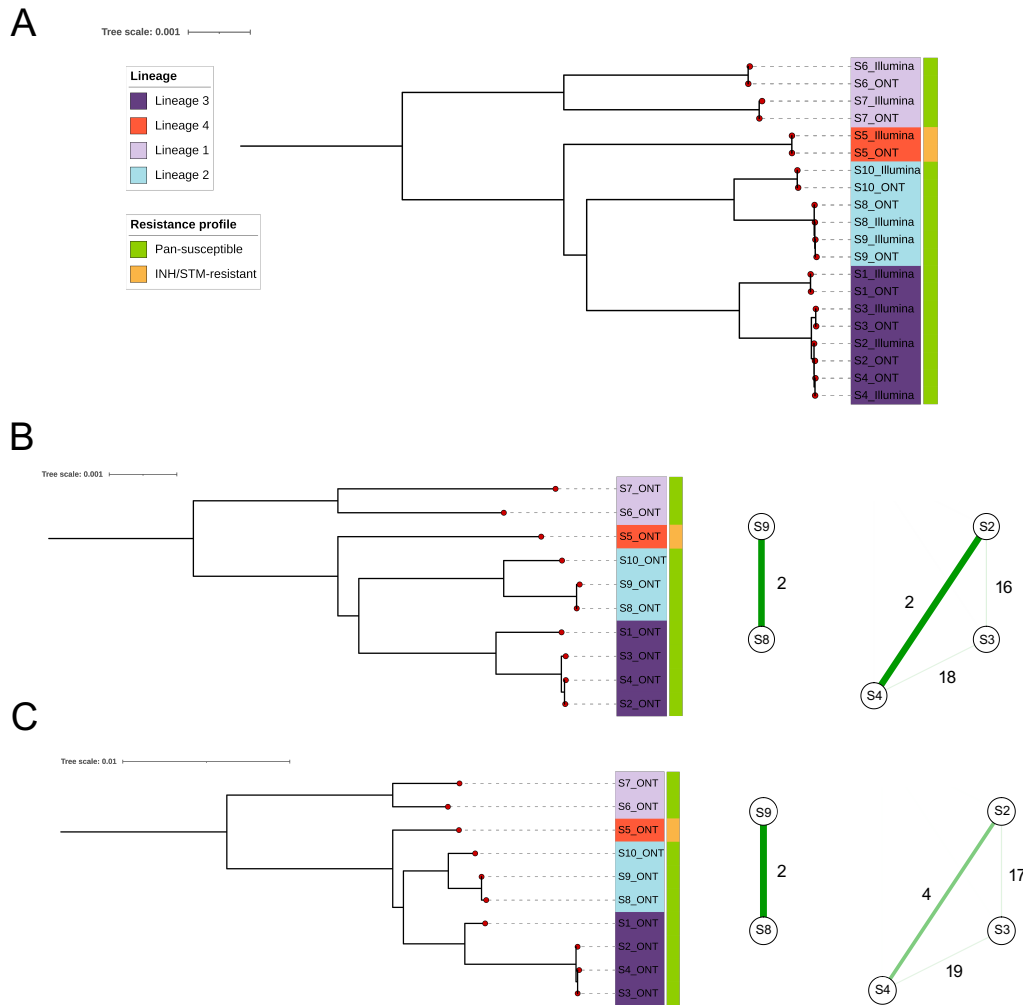


Figure 2. Phylogenetic trees and transmission networks

Maximum likelihood phylogenetic trees and transmission networks for the samples studied. Isolates are coloured by lineage. Drug resistance profile obtained by phenotypic drug susceptibility testing is shown in the strip labels on the trees. **(A)** Phylogenetic tree reveals high degree of concordance and clustering of replicates sequenced using Oxford Nanopore Technologies (ONT) and Illumina platforms, reconstructed with 3,955 SNPs excluding genomic repetitive regions. **(B)** Phylogenetic tree of ONT sequenced samples using the 3,955 SNPs, as well as transmission networks for lineage L2 (S8 and S9) and L3 (S2, S3 and S4) clusters showing SNP

distances. **(C)** Phylogenetic tree of ONT sequenced samples using the 3,955 SNPs in addition to 568 more polymorphic sites located in *pe/ppc* genes, as well as transmission networks for lineage L2 (S8 and S9) and L3 (S2, S3 and S4) clusters with SNP distances.

Portable sequencing of *Mycobacterium tuberculosis* for clinical and epidemiological applications

Paula J. Gómez-González¹

Susana Campino¹

Jody Phelan^{1,*}

Taane Clark^{1,2,*}

¹ Faculty of Infectious and Tropical Diseases, London School of Hygiene & Tropical Medicine, WC1E 7HT London, UK

² Faculty of Epidemiology and Population Health, London School of Hygiene & Tropical Medicine, WC1E 7HT London, UK

* Joint authors

** correspondence

taane.clark@lshtm.ac.uk

Department of Infection Biology,

Faculty of Infectious and Tropical Diseases

London School of Hygiene & Tropical Medicine, Keppel Street, London, UK

Briefings in Bioinformatics, Case Study

Additional File

Table S1. ENA accession number for study samples undergoing sequencing using Illumina and Oxford Nanopore Technology (ONT) platforms.

Sample	Lineage	Drug resistance genotypic profile	Phenotypic DST profile	Illumina Sequencing	ONT
S1	3	Pan-susceptible	Pan-susceptible	ERR161062	ERR8170869
S2	3	Pan-susceptible	Pan-susceptible	ERR182032	ERR8170870
S3	3	Pan-susceptible	Pan-susceptible	ERR245682	ERR8170871
S4	3	Pan-susceptible	Pan-susceptible	ERR245678	ERR8170872
S5	4.9	INH resistant	INH and STR resistant	ERR181826	ERR8170873
S6	1.1.2	Pan-susceptible	Pan-susceptible	ERR181951	ERR8170874
S7	1.1.3.2	Pan-susceptible	Pan-susceptible	ERR181929	ERR8170875
S8	2.2.1	Pan-susceptible	Pan-susceptible	ERR181821	ERR8170876
S9	2.2.1	Pan-susceptible	Pan-susceptible	ERR221538	ERR8170877
S10	2.2.1	Pan-susceptible	Pan-susceptible	ERR221573	ERR8170878
H37Rv	4	Pan-susceptible	Pan-susceptible	-	ERR8441303, ERR8441304, ERR8441305, ERR8441306

INH = Isoniazid, STR = Streptomycin; DST = drug susceptibility testing

Table S2. Discrepancies between Illumina and Oxford Nanopore Technology (ONT) SNP calls.

Sample pair	POS REF	Gene	ONT alternative allele (depth*)	Illumina alternative allele (depth*)
S1	55553 CCG	<i>Rv0050</i>	TCG (0.37) CCGTTCG (0.60)	TCG (0.88)
	1608276 A	<i>Rv1431</i>	C (0.69)	C (1)
S2	55553 CCG	<i>Rv0050</i>	TCG (0.44) CCGTTCG (0.53)	TCG (0.82)
	4027914 C	<i>Rv3586</i>	T (0.69)	T (1)
	4323831 C	<i>Rv3849</i>	T (0.68)	T (1)
S3	55553 CCG	<i>Rv0050</i>	TCG (0.36) CCGTTCG (0.61)	TCG (0.85)
	1608276 A	<i>Rv1431</i>	C (0.65)	C (1)
S4	55553 CCG	<i>Rv0050</i>	TCG (0.25) CCGTTCG (0.69)	TCG (0.87)
	4027914 C	<i>Rv3586</i>	T (0.65)	T (1)
S6	50906 C	<i>Rv0046c</i>	T (0.68)	T (0.98)
	55553 CCG	<i>Rv0050</i>	TCG (0.06) CCGTTCG (0.92)	TCG (0.74)
	1585283 A	<i>Rv1409</i>	C (0.61)	C (0.98)
	1798355 G	<i>Rv1597</i>	A (0.61)	A (1)
	2663210 G	<i>Rv2380c</i>	A (0.69)	A (0.96)
	55553 CCG	<i>Rv0050</i>	TCG (0.12) CCGTTCG (0.79)	TCG (0.82)
S7	1585283 A	<i>Rv1409</i>	C (0.60)	C (1)
	1798355 G	<i>Rv1597</i>	A (0.64)	A (1)
	2092970 C	<i>Rv1843c</i>	T (0.68)	T (0.98)
	2093715 T	<i>Rv1843c</i>	C (0.61)	C (0.97)
	2827111 C	<i>Rv2510c</i>	T (0.68)	T (0.94)
	3220048 C	<i>Rv2913c</i>	T (0.57)	T (1)
	3479561 G	<i>Rv3111</i>	A (0.69)	A (0.99)
	3653225 C	<i>Rv3271c</i>	T (0.66)	T (0.91)
S8	55553 CCG	<i>Rv0050</i>	TCG (0.23) CCGTTCG (0.69)	TCG (0.83)
	460413 C	<i>Rv0384c</i>	T (0.65)	T (0.94)
	1831219 CAC	<i>Rv1629</i>	CCC (0.19) CC (0.78)	CCC (1)
	3010993 C	<i>Rv2693c</i>	T (0.68)	T (1)
S9	55553 CCG	<i>Rv0050</i>	TCG (0.09) CCGTTCG (0.91)	TCG (0.73)
	460413 C	<i>Rv0384c</i>	T (0.65)	T (0.94)
	1097220 C	<i>Rv0981</i>	T (0.69)	T (1)
	1831219 CAC	<i>Rv1629</i>	CCC (0.19) CC (0.78)	CCC (1)
S10	39030 C	<i>Rv0035</i>	T (0.36)	T (0.80)

55553 CCG	<i>Rv0050</i>	TCG (0.20) CCGTCC (0.74)	TCG (0.73)
549361 CGC	<i>Rv0457c</i>	CGG (0.95) CGGG (0)	CGG (0.14) CGGG (0.84)
1608276 A	<i>Rv1431</i>	C (0.68)	C (1)
1831219 CAC	<i>Rv1629</i>	CCC (0.3) CC (0.67)	CCC (1)
4359165 G	<i>Rv3879c</i>	C (0.63)	C (0.99)

ONT = Oxford Nanopore Technology; * Allele depth; in bold, platform where alternate allele was called (alternative depth ≥ 0.7 ; alleles with indels not considered). Note, there were no discrepancies between calls in Illumina and ONT data for sample S5.

Table S3. Discrepancies between Illumina and Oxford Nanopore Technology (ONT) indel calls.

Sample pair	POS	Gene	ONT alternative allele (depth*)	Illumina alternative allele (depth*)
S1	293628	<i>Rv0243</i>	insC (0.61)	insC (1)
	854252	<i>Rv0759c</i>	delC (0.42) delCC (0.57)	delC (1)
	1365837	<i>Rv1222</i>	insGG (0.45) insG (0.40)	insGG (1)
	2320329	<i>Rv2062c</i>	delC (0.75)	delC (0)
	2631009	<i>Rv2351c</i>	insTGCCG (0.47)	insTGCCG (0.93)
	2850856	<i>Rv2525c</i>	delG (0.71)	delG (0)
	3296371	<i>Rv2947c</i>	insCGCGGCC (0.71)	insCGCGGCC (0.69)
S2	293628	<i>Rv0243</i>	insC (0.63)	delC (1)
	691887	<i>Rv0592</i>	insC (0.55)	insC (0.98)
	830868	<i>Rv0739</i>	insCG (0.67)	insCG (1)
	854252	<i>Rv0759c</i>	delC (0.44) delCC (0.54)	delC (1)
	1365837	<i>Rv1222</i>	insGG (0.37) insG (0.48)	insGG (1)
	2320329	<i>Rv2062c</i>	delC (0.80)	delC (0)
	2536625	<i>Rv2264c</i>	insGG (0.21)	insGG (1)
	2631009	<i>Rv2351c</i>	insTGCCG (0.42)	insTGCCG (0.89)
	2850856	<i>Rv2525c</i>	delG (0.82)	delG (0)
	3131469	<i>Rv2823c</i>	insTCGGCGATG (0.85)	insTCGGCGATG (0.64)
3296371	<i>Rv2947c</i>	insCGCGGCC (0.65)	insCGCGGCC (0.74)	
S3	125830	<i>Rv0107c</i>	insA (0.68)	insA (1)
	293628	<i>Rv0243</i>	insC (0.65)	insC (1)
	691887	<i>Rv0592</i>	insC (0.50)	insC (0.98)
	854252	<i>Rv0759c</i>	delC (0.58) delCC (0.38)	delC (1)
	1365837	<i>Rv1222</i>	insG (0.45) insGG (0.43)	insG (1)
	2536625	<i>Rv2264c</i>	insGG (0.25)	insGG (1)
	2631009	<i>Rv2351c</i>	insTGCCG (0.40)	insTGCCG (0.73)
	2320329	<i>Rv2062c</i>	delC (0.78)	delC (0)
	2850856	<i>Rv2525c</i>	delG (0.79)	delG (0)
	3059811	<i>Rv2747</i>	delT (1)	delT (0.04)
	3059829	<i>Rv2747</i>	insA (0.92)	insA (0)
	3131469	<i>Rv2823c</i>	insTCGGCGATG (0.90)	insTCGGCGATG (0.48)
	S4	293628	<i>Rv0243</i>	insC (0.61)
691887		<i>Rv0592</i>	insC (0.67)	insC (1)
830868		<i>Rv0739</i>	insCG (0.66)	insCG (1)
854252		<i>Rv0759c</i>	delC (0.50)	delC (1)

			delCC (0.37)	
	1365837	<i>Rv1222</i>	insG (0.49)	insG (1)
			insGG (0.47)	
	2536625	<i>Rv2264c</i>	insGG (0.38)	insGG (1)
	2631009	<i>Rv2351c</i>	insTGCCG (0.39)	insTGCCG (0.80)
	3059811	<i>Rv2747</i>	delT (0.99)	delT (0.02)
	3059829	<i>Rv2747</i>	insA (0.94)	insA (0)
	3131469	<i>Rv2823c</i>	insTCGGCGATG (0.93)	insTCGGCGATG (0.52)
	3296371	<i>Rv2947c</i>	insCGCGGCC (0.76)	insCGCGGCC (0.49)
			delC (0.48)	delC (1)
	854252	<i>Rv0759c</i>	delCC (0.48)	
S5	2059780	<i>Rv1817</i>	insG (0.26)	insG (0.97)
	2320329	<i>Rv2062c</i>	delC (0.77)	delC (0)
	3190145	<i>Rv2880c</i>	delC (0.84)	delC (0)
	125830	<i>Rv0107c</i>	insA (0.69)	insA (1)
	191391	<i>Rv0161</i>	insC (0.2)	insC (0.95)
	293628	<i>Rv0243</i>	insC (0.59)	insC (0.96)
	854252	<i>Rv0759c</i>	delC (0.40)	delC (0.99)
			delCC (0.56)	
S6	919284	<i>Rv0825c</i>	insG (0.30)	insG (0.96)
	1365837	<i>Rv1222</i>	insG (0.59)	insG (1)
			insGG (0.29)	
	2547529	<i>Rv2275</i>	insG (0.58)	insG (0.97)
	2730151	<i>Rv2434c</i>	insC (0.17)	insC (1)
	3723901	<i>Rv3337</i>	insT (0.69)	insT (0.94)
	293628	<i>Rv0243</i>	insC (0.64)	insC (1)
	854252	<i>Rv0759c</i>	delC (0.57)	delC (0.97)
			delCC (0.41)	
	1365837	<i>Rv1222</i>	insG (0.51)	insG (1)
			insGG (0.34)	
S7	2090400	<i>Rv1841c</i>	insCCAACGCCACCG (0.86)	(0.67, **DP=24)
	2547529	<i>Rv2275</i>	insG (0.68)	insG (0.91)
	3131469	<i>Rv2823c</i>	insTCGGCGATG (0.88)	insTCGGCGATG (0.63)
	3296371	<i>Rv2947c</i>	insCGCGGCC (0.70)	insCGCGGCC (0.68, *DP=22)
	3723901	<i>Rv3337</i>	insT (0.69)	insT (0.98)
	125830	<i>Rv0107c</i>	insA (0.65)	insA (1)
	293628	<i>Rv0243</i>	insC (0.69)	insC (1)
	799136	<i>Rv0698</i>	delC (0.73)	delC (0)
S8	854252	<i>Rv0759c</i>	delC (0.36)	delC (0.98)
			delCC (0.56)	
	964001	<i>Rv0866</i>	insG (0.44)	insG (0.98)
	987585	<i>Rv0888</i>	insG (0.22)	insG (0.98)

	1365837	<i>Rv1222</i>	insG (0.38) insGG (0.43)	insG (1)
	2320329	<i>Rv2062c</i>	delC (0.71)	delC (0.02)
	2850856	<i>Rv2525c</i>	delG (0.79)	delG (0)
	125830	<i>Rv0107c</i>	insA (0.68)	insA (0.98)
	293628	<i>Rv0243</i>	insC (0.60)	insC (1)
	854252	<i>Rv0759c</i>	delC (0.4) delCC (0.6)	delC (0.98)
	964001	<i>Rv0866</i>	insG (0.44)	insG (0.85)
S9	987585	<i>Rv0888</i>	insG (0.16)	insG (0.94)
	1365837	<i>Rv1222</i>	insG (0.46) insGG (0.46)	insG (1)
	1753519	<i>Rv1549</i>	insC (0.64)	insC (1)
	2850856	<i>Rv2525c</i>	delG (0.77)	delG (0)
	3131469	<i>Rv2823c</i>	insTCGGCGATG (0.90)	insTCGGCGATG (0.69)
	125830	<i>Rv0107c</i>	insA (0.65)	insA (1)
	809840	<i>Rv0712</i>	insC (0.24)	insC (1)
	987585	<i>Rv0888</i>	insG (0.27)	insG (1)
S10	1365837	<i>Rv1222</i>	insG (0.55) insGG (0.31)	insG (1)
	2338194	<i>Rv2081c</i>	delC (0.37) insC (0.06)	insC (0.93)
	2850856	<i>Rv2525c</i>	delG (0.72)	delG (0)
	3131469	<i>Rv2823c</i>	insTCGGCGATG (0.86)	insTCGGCGATG (0.67)

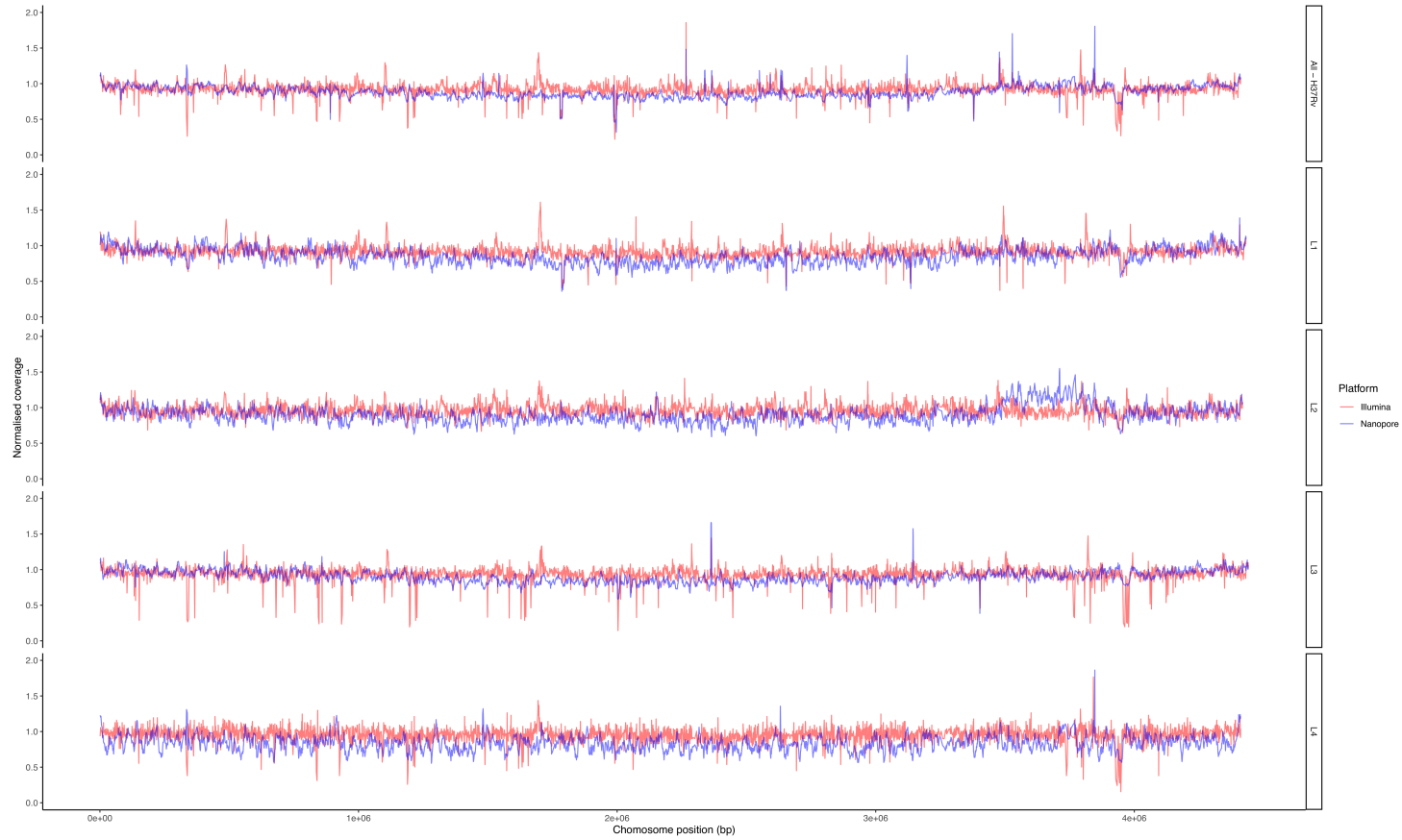
ONT = Oxford Nanopore Technology; *Allele depth; **DP = low total read depth at locus; in **bold**, platform where alternate allele was called (allele depth ≥ 0.7).

Table S4. Large structural variants* identified in Illumina and Oxford Nanopore Technology (ONT) sequence data

Sample pair	Insertions			Deletions		
	ONT only	Illumina only	Both	ONT only	Illumina only	Both
S1	41	1	1	17	8	19
S2	46	0	0	18	6	17
S3	44	0	0	18	6	14
S4	46	0	0	20	4	14
S5	11	0	0	3	0	4
S6	50	2	1	17	3	19
S7	47	1	1	14	5	18
S8	48	0	0	20	8	14
S9	49	0	0	18	9	16
S10	52	0	1	20	12	17

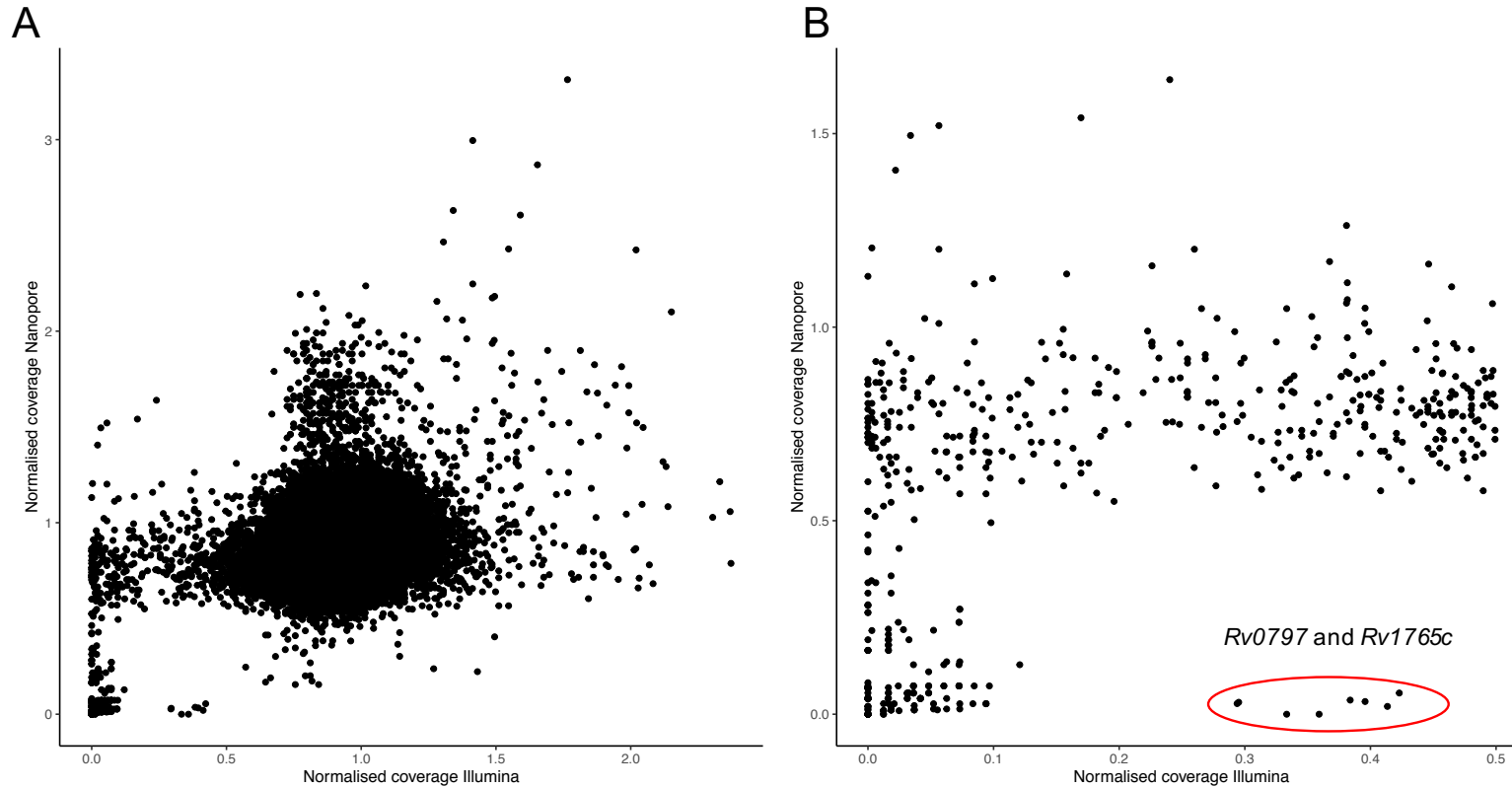
*Insertions and deletions over 15 bp identified using Delly software.

Figure S1. Genome-wide normalised coverage



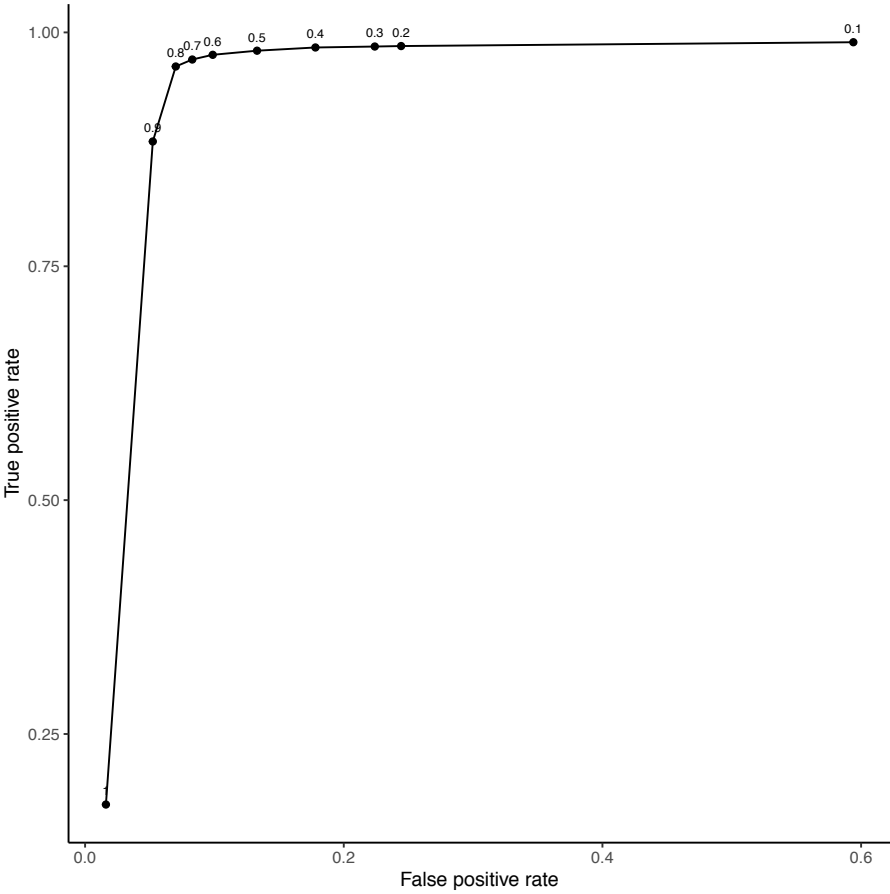
Normalised median coverage (vertical axis) along the chromosome (horizontal axis) when using H37Rv as a reference (top track) or PacBio lineage-specific assemblies (second to fifth track (L1-L4)). Coverage from Illumina data (red) and ONT data (blue). The large region spanning 3.5 Mbp to 4 Mbp with increased coverage in L2 isolates corresponds to the *DosR* regulon duplication.

Figure S2. Correlation of normalised coverage between Illumina and Oxford Nanopore Technology (ONT) platforms



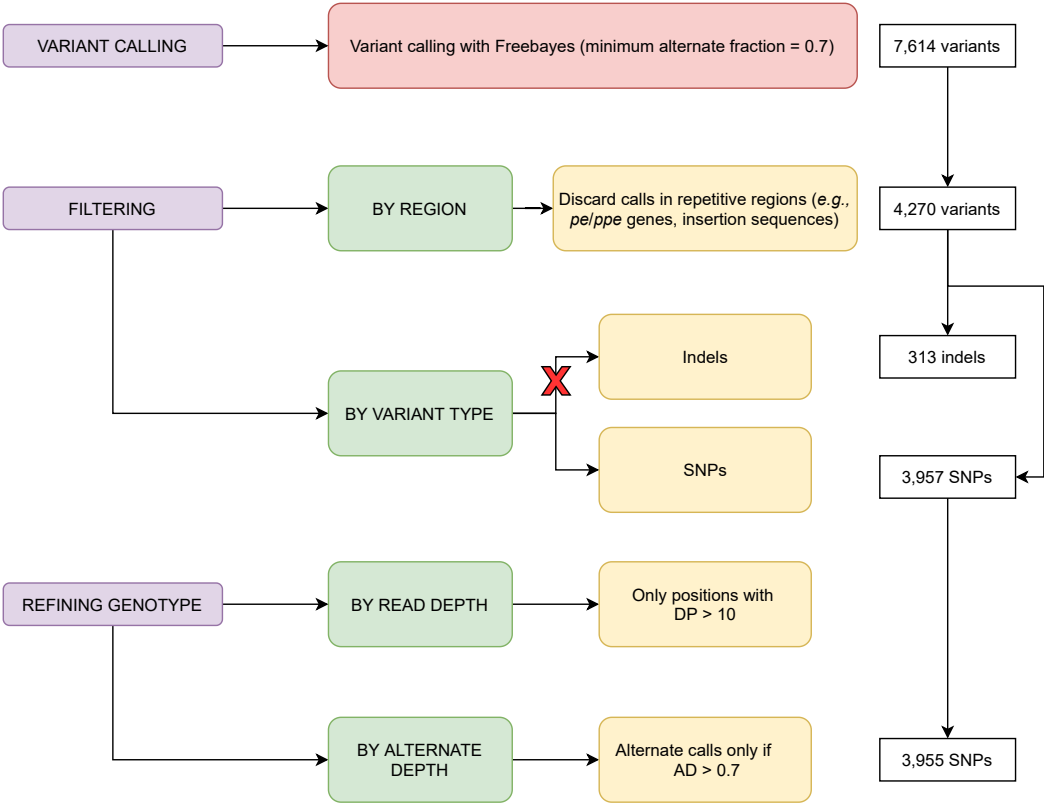
Correlation of normalised median coverage per gene per sample in both sequencing platforms (vertical axis ONT, horizontal axis Illumina) for (A) all genes and (B) genes with a median normalised coverage < 0.5 in Illumina data in at least one sample. Overall, (A) shows a good correlation of coverage between both platforms. In (B), most genes show higher coverage in ONT data. Genes with normalised coverage < 0.1 in both Illumina and ONT represent true deletions. Annotated genes (*Rv0797* and *Rv1765c*) highlight two cases where coverage was higher in Illumina data due to repetitive regions (insertion sequence and highly similarity of a deleted gene belonging to RD152 to *Rv2015c* respectively).

Figure S3. Receiver-operator characteristic curve for the error rate of Oxford Nanopore Technology (ONT) data



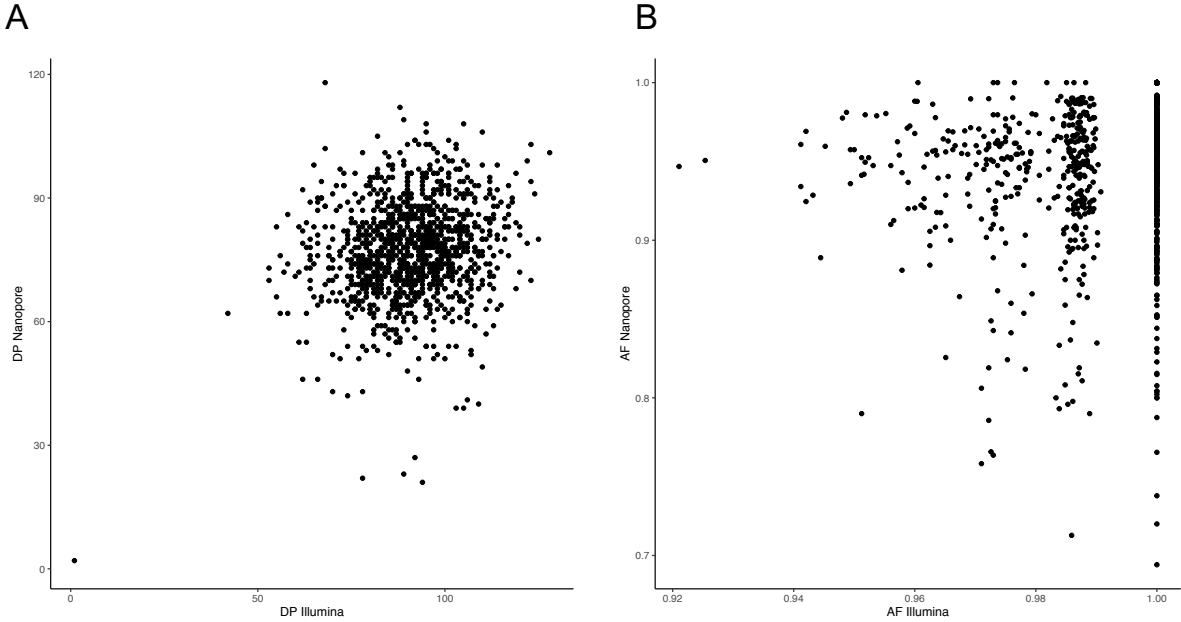
ROC curve showing the True Positive Rate on the vertical axis with the False Positive Rate on the horizontal axis. All cut-off points studied are annotated in the curve.

Figure S4. Analysis pipeline



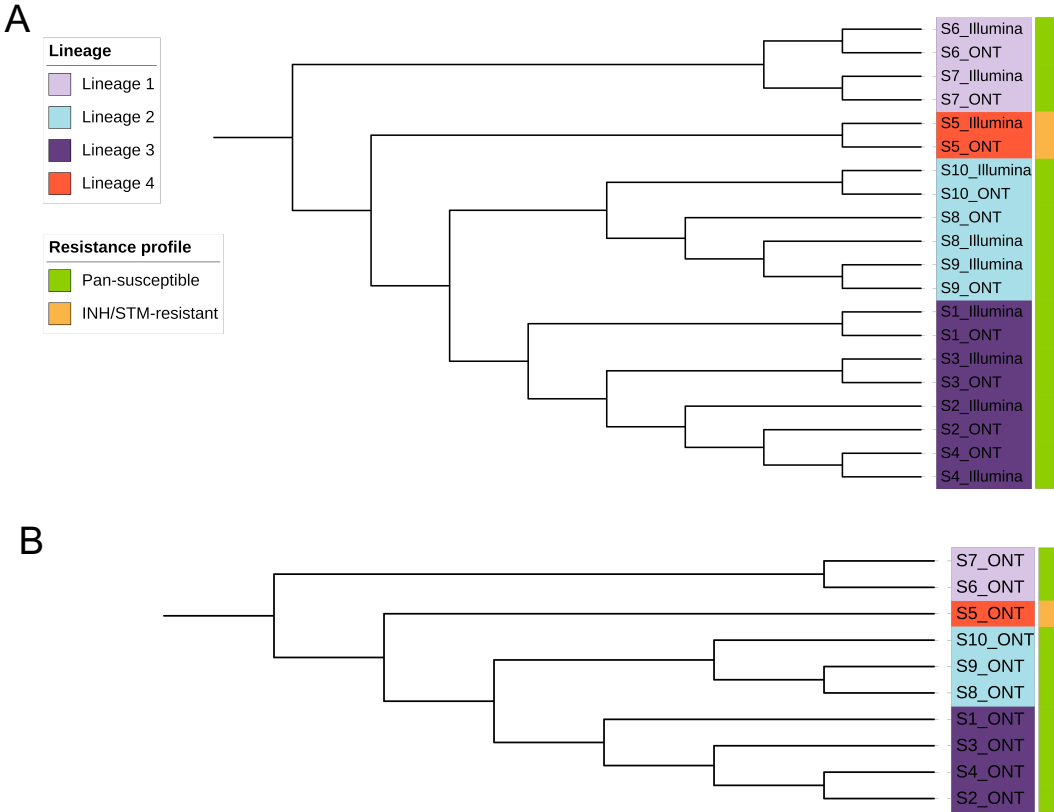
Summary pipeline of the variant calling, filtering and genotype refining steps carried out to obtain a set of high-quality SNPs. DP = read depth at a locus; AD = alternate allele depth fraction; indels = insertions and deletions.

Figure S5. Depth of coverage and alternate allele depth fraction correlation between Illumina and Oxford Nanopore Technology (ONT) for SNPs called in both platforms.



Correlation plots showing **(A)** read depth (RD) and **(B)** alternate allele depth fraction (AF) between Illumina and ONT for positions with SNPs called in both platforms, with Illumina on the horizontal axis and ONT on the vertical axis. **(A)** shows a good correlation between read depth in positions with concordant SNPs; **(B)** shows how ONT reads are noisier than the corresponding Illumina reads, with the fraction of alternate allele depth for ONT with lower values (0.7-1) than the Illumina platform (>0.92).

Figure S6. Cladogram of Oxford Nanopore Technology (ONT) and Illumina sequenced isolates



Cladogram representing the branching order with equal branch lengths for the 10 pairs of Illumina and ONT isolates **(A)** and only the 10 ONT isolates **(B)**; INH = Isoniazid, STR = Streptomycin.

CHAPTER 7

Discussion

7.1. General discussion

An understanding of the biology of *Mycobacterium tuberculosis* (*Mtb*) is a crucial aspect for the control of an infection that still causes more than a million deaths per year. Since various members of the *Mycobacterium tuberculosis* complex are responsible for human tuberculosis disease, and differences among them have implications in pathogenicity or acquisition of drug resistance, it is of importance to understand the biological mechanisms resulting in these phenotypic changes. The recent development and expansion of whole-genome sequencing technologies have provided the means to carry out high throughput genomic analysis that can help to understand the genomic differences that exist between strains and its potential relevance in different biological processes. Thus, this thesis focuses on the use of whole-genome sequencing data as a means to interrogate the genome of the different lineages of *Mtb* to study diverse aspects of its biology.

During *Mtb* infection, transcriptional changes occur in response to environmental cues as a mechanism of adaptation to the changing conditions, such as the expression changes provoked by the dormancy survival regulator DosR that affects transcription of more than 50 genes [1]. However, different core transcriptomes have also been described among *Mtb* clinical isolates under same conditions [2]. Differential expression between strains can result in various phenotypes that can ultimately impact infection and clinical outcomes. On the other hand, N⁶-adenine methylation has been proposed as a mechanism responsible for phenotypic variation among strains, where loss of function mutations in the methyltransferases (MTases), often associated with lineage, have been suggested to explain gene expression differences [3–5]. Gene expression can be influenced by genomic diversity, such as variants in transcriptional regulators or promoters [2, 6], and mutations inactivating MTases that alter the methylation pro-

file [3,4]. The aim of **Chapter 3** was to interrogate the differential gene expression of a sample set representing three of the major *Mtb* lineages in relation to their genomic and methylation patterns, by combining three levels of 'omics data: DNA, RNA and methylation. In support of previous findings [2], different transcriptomes between the ancient and the modern strains were observed. To investigate the underlying mechanisms responsible for these expression differences, two types of variants were considered, including SNPs within the promoter regions and transcriptional start sites of the differentially expressed genes, together with non-synonymous mutations potentially leading to functional impairment of transcriptional regulators. An expression quantitative trait loci (eQTL) study was performed to establish the associations between the variants and the level of gene expression, revealing numerous *cis*- and *trans*-eQTL candidates. The same approach was used for the association of the methylated and un-methylated motifs identified along the genome with gene expression levels. This analysis revealed diverse modification patterns, from which a correlation of absence of methylation as a consequence of loss of function mutations, and down-regulation of specific genes was found, consistent with previous work [3]. Besides the previously reported variants, novel mutations in *mamA* (e.g., G152S) were identified, which could explain the lack of methylation of the CTCCAG motif. Additionally, the partial activity of MTases caused by specific mutations behind intracellular stochastic methylation patterns has been suggested by recent work [7]. This insight, together with the corroboration of promoter DNA methylation influence in transcription, leads to the hypothesis of heterogeneous phenotypes as a result of heterogeneous methylation [7]. Despite the advances in methylation analysis in *Mtb*, in part due to the more accessible modification detection pipelines by PacBio or Oxford Nanopore Technologies (ONT), the physiological consequences of this epigenetic regulator are still unknown, and more research is necessary in order to gain insights into the implications of the different transcriptomes or

methylation patterns identified across different strains may have in pathogenicity or acquisition of drug resistance, to ultimately inform in drug or vaccine development.

As *Mtb* lacks horizontal gene transfer, acquisition of drug resistance is mainly caused by SNPs or indels in drug targets, drug-activating enzymes or genes coding for proteins involved in transport of small molecules like efflux pumps [8, 9]. Thereby, the investigation of variants in loci known to interact with anti-TB drugs can provide insights into the emergence of drug resistance. **Chapter 4** comprises a large-scale study of variants in candidate genes for resistance to three of the most recently introduced anti-TB drugs: bedaquiline (BDQ), delamanid (DLM) and pretomanid (PTM), used for the treatment of MDR- and XDR-TB cases. In a large data set ($n = \sim 30k$ Illumina genomes) with all *Mtb* lineages represented, the frequency and distribution of variants in 9 candidate loci were investigated. More than one thousand different mutations including non-synonymous SNPs and small indels were identified, most of them being found in isolates collected prior to the introduction of BDQ, DLM and PTM as an anti-TB treatment. Through phylogenetic and convergent evolution analysis, together with the available drug susceptibility testing (DST) data, some of these mutations could be determined as phylogenetically informative and unlikely to be associated with resistance. However, there were several other variants, including nonsense SNPs and frameshifts, that could imply intrinsic resistance in naïve strains. Interestingly, some of these variants were fixed in populations with high allele frequencies observed within a sub-lineage, others were part of transmission clusters, or showed simultaneous occurrence in phylogenetically distant isolates. These findings are in line with previous identification of spontaneous mutations in BDQ/DLM-naïve isolates [8, 10–13], and even to the most recent PTM [14], which raises concerns due to the complications that intrinsic resistance can pose for future treatment of MDR- and XDR-TB cases. In some situations, the use of clofaz-

imine (CFZ) or azoles can explain the emergence of cross-resistance to BDQ through mutations in the transcriptional regulator *mmpR5*. Additionally, in a drug resistance context, it is also important to note the possible epistatic interactions [15], where mutations in a different gene can counteract the resistance effect of another mutation (*e.g.*, *mmpL5* deletion and *mmpR5* frameshift), explaining genotype-phenotype discrepancies. Genome-wide association studies (GWAS) with DST data and a better understanding of the mechanisms of action can help to elucidate such effects. Moreover, protein stability software, such as SUSPECT-BDQ, are useful to predict likely phenotypes based on mutations. Nevertheless, more MIC data is necessary to determine the clinical relevance of the frequent mutations associated in the literature with low level of resistance, below BDQ/DLM resistance breakpoints [8, 16, 17], which could lead to treatment failure due to suboptimal regimens.

Variation among *Mtb* strains is also reflected in the *pe* and *ppe* gene families, where, due to their role in host-pathogen interactions, it could cause differences in pathogenicity. The two gene families are known hot spots of recombination and polymorphisms [18, 19] and have been suggested to be involved in antigenic variation and immune evasion [20]; although, conservation among T-cell epitopes does not support the theory of immune selection of these proteins [21, 22]. However, their GC-rich and repetitive nature has resulted in their systematic exclusion from whole-genome sequencing (WGS) analysis owing to the lack of accuracy in mapping short reads to these regions [23]. With the purpose of characterising these complex gene families, **Chapter 5** describes the successful use of long-read sequencing data to generate alignments for the 169 *pe/pppe* genes and study their diversity using representatives of the main *Mtb* lineages. Newly cultured and sequenced clinical isolates together with a set of publicly available complete PacBio genomes were included to a final data set of 72 genomes

to cover ancient, modern and *M. bovis* strains. A conservative approach was used to classify the *pe/ppe* genes based on their structural variants across the different lineages, revealing a significant number of conserved genes, and when assessed per sample, > 50% of these genes were also found conserved relative to the H37Rv reference. SNP and indel diversity per site were higher in *pe/ppe* genes than in the rest of the genome, with a predominance of indel diversity among the genes classified as non-conserved, and in the *pe_pgrrs* sub-family, more specifically, after the PE domain. In contrast to this observation, SNPs were the main source of diversity in the conserved genes and within the *ppe* and remaining *pe* genes. Inter-lineage variation was expected within these two families, as it occurs genome-wide. Indeed, the presence of several lineage-specific variants, including indels leading to disrupted proteins, was identified and validated in a larger data set of short-read data, suggesting a possible lineage-specific host-pathogen interaction. Supported by PGAP annotation and protein structure prediction by AlphaFold where possible, duplication events, gene fusions or integration of *IS6110*, often following lineage patterns, were also among the structural variants identified, demonstrating the complexity of the *pe/ppe* gene arrangements. Interestingly, inconsistencies between the clinical isolates analysed and the annotation of the H37Rv reference genome highlight a potential pitfall to accurately capture variants in these complex genes using a reference-based approach. For instance, a second copy of *pe_pgrrs3* similar to that found in *M. bovis* or *M. canetti* was identified in most of the samples, indicating that recombination events could have resulted in the possible loss of a copy in H37Rv and related strains. This consequently leads to the erroneous identification of numerous variants when H37Rv is used as a reference. Another interesting finding was the observation that several of the genes annotated as pseudogenes in H37Rv due to premature stop codons were annotated in clinical strains as likely functional genes. Overall, different degrees of variation, including lineage patterns, were found among

these two families. Considering the significant number of structurally conserved genes, but yet with a certain degree of variation, it is possible that these genes could have a phylogenetically informative value if included in WGS analysis. Moreover, due to their immunogenic nature, PE/PPE proteins have been often targeted as vaccine candidates, for which a better understanding and characterisation of their function and strain variation is necessary. Overall, the *pe/pppe* work has provided with new insights and processed data, including a list of conserved genes, to assist follow-up investigations, including laboratory functional work.

Among the advantages of the use of WGS technologies, it is important to highlight the clinical and epidemiological applications. The use of WGS to analyse pathogen DNA/RNA has been recently implemented in countries like the UK, including for COVID-19 insights and TB management. For TB, the current accessibility to these sequencing platforms and bioinformatic pipelines has the potential to significantly improve patient management, with faster detection of drug resistance associated mutations through direct sequencing from sputum [24], circumventing the laborious and time-consuming culture steps. Nevertheless, for the implementation of these technologies in high burden TB settings, reduced costs and infrastructure are necessary and are now achievable with devices such as MinION from ONT. In **Chapter 6**, a pair-wise comparison between the gold standard Illumina data and ONT long-reads from cultured and sequenced clinical isolates was carried out, to evaluate the applicability of the latter technology in drug resistance and transmission analysis. Good genome-wide coverage was obtained with ONT data, without the apparent GC content biases that can affect Illumina data output. On the premises of the better characterisation of the *pe/pppe* genes with PacBio long-reads observed in **Chapter 5**, the coverage of these genes in the ONT replicates was investigated, showing a significantly improved read depth compared to their Illumina counterparts, espe-

cially among the non-conserved genes. Despite the higher sequence error rate of ONT, good concordance between SNPs identified through both platforms was found at an alternate allele depth fraction ≥ 0.7 , supporting the reliability of ONT data. However, for small indels, a more accurate characterisation is obtained with Illumina data. The robust SNP identification with ONT reinforces its possible application to elucidate transmission clusters, and in order to improve the resolution, it is plausible to include up to 150 *pe/ppc* genes with good coverage across the different lineages. Additionally, although the samples analysed were mostly pan-susceptible, high quality variants were called at all drug resistance loci. Therefore, this study supports the application and implementation of ONT, such as the MinION portable sequencer, for drug resistance detection or epidemiological and transmission dynamics investigations. Recent target amplicon sequencing of drug resistance loci approaches have also been described using MinION technologies for the accurate and cost-effective characterisation of drug resistance markers in *Mtb* [25], moving towards a more realistic and affordable application of WGS technologies to enable a prompt and accurate diagnosis and inform decision making in the context of drug-resistant TB.

In summary, the implications of the work presented on this thesis on the field of TB control are varied. The differences between lineages of *Mtb* at different levels (genomic, expression or methylation differences) could imply phenotypic diversity. And the understanding of phenotypic diversity in *Mtb* is crucial to achieve more accurate diagnostic tools and treatments. The complexity observed in genes involved in host-pathogen interactions (*e.g.*, *pe/ppc* genes) and the differential expression between lineages points towards potential different behaviours that could be of importance when developing diagnostics or treatments. Well conserved targets among *Mtb* lineages should be of choice to ensure their application. On the other hand,

drug-resistance poses a real threat to the control of TB. The existence of mutations potentially associated with resistance to the new drugs in naïve MDR or XDR isolates leads to reduced choices for treatment with increased side effects. This highlights the need for better therapeutic options to improve patient management and adherence. Moreover, a better understanding of the drug mechanisms of action and the biological mechanisms responsible of phenotypic drug resistance could assist in drug development. The availability of fast and accurate detection of drug resistance through portable sequencing technologies is a great advance in diagnostics. Nevertheless, certain limitations such as the actual cost or the lack of reliable genotypic-phenotypic data for certain drugs makes it difficult to fully implement as the gold-standard method to use, especially in high-burden settings.

7.2. Conclusions

This thesis presents an analysis of *Mtb* sequence data to inform on diversity across various lineages at different levels, such as methylation and gene expression (**Chapter 3**), acquisition and distribution of drug resistant mutations (**Chapter 4**), or diversity within protein families involved in host-pathogen interactions (**Chapter 5**), to gain insights into the differences that can be observed and the biological implications that they might have. The combined application of different ‘omics has shown the potential to decipher more complex biological mechanisms. Moreover, the use of long-read WGS data (*e.g.*, PacBio) can resolve complicated gene morphologies, like the *pe* and *ppe* genes, with high GC content and repetitive regions, where traditional short-read sequencing may encounter difficulties. The suitability of portable and cost-effective sequencers, such as MinION from Oxford Nanopore Technology, is supported by the robustness of the variant detection pipeline (**Chapter 6**), thereby with promising applica-

tions of epidemiological and clinical relevance. Finally, the relative high frequency of mutations potentially conferring drug resistance to the most recent anti-TB drugs highlights the necessity of further efforts in drug discovery and vaccine development to assist control of the disease and move towards eradication. In summary, this thesis provides a comprehensive analysis of different *Mtb* lineages by using various 'omics approaches in order to contribute towards a better understanding of its biology and diversity.

7.3. The future of TB 'Omics

Despite the exponential growth in knowledge on *Mtb* infection and disease epidemiology since the discovery of Koch's bacillus in 1882, over the last decades, progress on TB control has been modest and human tuberculosis is yet not close to being eradicated. Hence, further research to tackle drug resistance, improve treatment regimens and develop effective vaccines is necessary to ultimately control and hopefully eradicate the disease. There is growing evidence on diversity across members of the *Mycobacterium tuberculosis* complex reflected in different phenotypes and likely to have implications in host-pathogen interactions [26]. For instance, **Chapter 5** describes the complexity and diversity of the *pe* and *ppe* genes, not only driven by SNPs, but also by indels and larger complex structural variants. Functional and other experimental data that reflects this diversity is necessary to understand how these differences may affect clinical outcomes or pathogenicity. Further transcriptomics and proteomics analysis including various strains could reveal additional insights into the mechanisms by which *Mtb* interacts with the host, and potentially provide information for vaccine development. Moreover, in view of the lack of protein structures available, the development of *in silico* prediction tools, such as AlphaFold [27] are of great value. The impact of epigenetic regulation on gene

expression based on lineage-specific profiles and its consequences in pathogenicity is another example of the importance of strain diversity. Molecular techniques, such as CRISPR/Cas9-genome editing, enable the controlled targeting of mutations. Together with the more accessible use of WGS technologies, this allows the characterisation and understanding of the effects of 'omic diversity. The possibility of combining and integrating different 'omics, which ultimately brings together various levels of information, can help to understand complex biological processes of *Mtb* in a more comprehensive manner in a systems biology approach [28]. Application of integrated 'omics analysis, including genomics, transcriptomics, proteomics or metabolomics, can provide insights into, for example, the dormancy state, as well as assist in the identification of new drug targets or the mechanisms of action and potential resistance for lead compounds during drug discovery pipelines.

Leveraging off the development and availability of cost-effective WGS platforms and the current knowledge in genotype-phenotype association for drug resistance, the fast and accurate detection of resistance associated variants by WGS to inform decision making in the clinic has already been implemented in countries like the UK. In recent years, efforts to advance in culture-free techniques for drug resistance characterisation have been successful [24], including target amplicon sequencing using the portable MinION platform [25], which opens the door to its use in high TB-burden settings. Nevertheless, the continuous surveillance and DST-genotypic association studies are necessary to ensure the accurate and reliable *in silico* drug resistance prediction by tools like TB-Profiler [29]. Moreover, the clinical repercussion of mutations conferring low-level of resistance should be investigated. Large genome-wide 'omics studies can also help to identify possible epistatic interactions, disentangling mutation effects and minimising erroneous interpretation of *in silico* drug resistance predictions.

In addition, the feasibility of direct sequencing will enable the characterisation of intra-host diversity and prevent detection of variants introduced or acquired during culture. This direct sequencing brings the possibility to better capture variants in outbreak settings or, for example, methylation patterns. Finally, the recent COVID-19 experience has showed how global WGS for surveillance and rapid availability of genomic data can generate useful epidemiological information to help with the control of an infection. Additionally, in line with diagnostic developments achieved for COVID-19, further efforts should be made towards point-of-care TB tests, including, for instance, more accessible sample collection methods [30]. In conclusion, currently available technologies and methodologies, as well as future related technological developments, should lead to advances in TB research that ultimately will assist the development of tools for the control of the disease, particularly in high burden settings.

The implementation of better diagnostic tools, such as WGS and 'omic technologies in high burden settings is key to improve patient management, especially in drug resistant cases. However, the lack of infrastructure and resources often hinders the availability of better diagnostic tools. Moreover, one limitation of the TB data currently accessible is the number of sequences available of ancient lineages, *e.g.*, *M. africanum*, which are scarce possibly due to sourcing bias. The future research priorities in order to scale-up the implementation of WGS as diagnostic tools and address the problem of drug resistance should focus on (i) culture-free portable detection of *Mtb* including drug resistance loci, for which (ii) a better understanding of the genotypic-phenotypic relationship in drug resistance and (iii) the availability of affordable methods in high burden settings are necessary; (iv) a better study of the diversity of the MTBC and the potential existence of intrinsic resistance mutations, including epistatic interactions; and (v) the development of new drugs for the treatment of MDR/XDR-cases that can

shorten the treatment regimens.

References

- [1] Park, H.-d. *et al.* *Rv3133c/dosR* is a transcription factor that mediates the hypoxic response of *Mycobacterium tuberculosis*. *Molecular microbiology* **48**, 833–43 (2003).
- [2] Homolka, S., Niemann, S., Russell, D. G. & Rohde, K. H. Functional Genetic Diversity among *Mycobacterium tuberculosis* Complex Clinical Isolates: Delineation of Conserved Core and Lineage-Specific Transcriptomes during Intracellular Survival. *PLoS Pathogens* **6**, e1000988 (2010).
- [3] Shell, S. S. *et al.* DNA Methylation Impacts Gene Expression and Ensures Hypoxic Survival of *Mycobacterium tuberculosis*. *PLoS Pathogens* **9**, 24–28 (2013).
- [4] Zhu, L. *et al.* Precision methylome characterization of *Mycobacterium tuberculosis* complex (MTBC) using PacBio single-molecule real-time (SMRT) technology. *Nucleic Acids Research* **44**, 730–743 (2016).
- [5] Phelan, J. *et al.* Methylation in *Mycobacterium tuberculosis* is lineage specific with associated mutations present globally. *Scientific Reports* **8**, 160 (2018).
- [6] Rose, G. *et al.* Mapping of Genotype–Phenotype Diversity among Clinical Isolates of *Mycobacterium tuberculosis* by Sequence-Based Transcriptional Profiling. *Genome Biology and Evolution* **5**, 1849–1862 (2013).
- [7] Modlin, S. J. *et al.* Drivers and sites of diversity in the DNA adenine methylomes of 93 *Mycobacterium tuberculosis* complex clinical isolates. *eLife* **9**, 1–33 (2020).

- [8] Villellas, C. *et al.* Unexpected high prevalence of resistance-associated *Rv0678* variants in MDR-TB patients without documented prior use of clofazimine or bedaquiline. *Journal of Antimicrobial Chemotherapy* **72**, 684–690 (2017).
- [9] Coll, F. *et al.* Genome-wide analysis of multi- and extensively drug-resistant *Mycobacterium tuberculosis*. *Nature Genetics* **50**, 307–316 (2018).
- [10] Xu, J. *et al.* Primary clofazimine and bedaquiline resistance among isolates from patients with multidrug-resistant tuberculosis. *Antimicrobial Agents and Chemotherapy* **61**, 1–8 (2017).
- [11] Zimenkov, D. V. *et al.* Examination of bedaquiline- and linezolid-resistant *Mycobacterium tuberculosis* isolates from the Moscow region. *Journal of Antimicrobial Chemotherapy* **72**, 1901–1906 (2017).
- [12] Fujiwara, M., Kawasaki, M., Hariguchi, N., Liu, Y. & Matsumoto, M. Mechanisms of resistance to delamanid, a drug for *Mycobacterium tuberculosis*. *Tuberculosis* **108**, 186–194 (2018).
- [13] Nimmo, C. *et al.* Population-level emergence of bedaquiline and clofazimine resistance-associated variants among patients with drug-resistant tuberculosis in southern Africa: a phenotypic and phylogenetic analysis. *The Lancet Microbe* **1**, e165–e174 (2020).
- [14] Reichmuth, M. L. *et al.* Natural Polymorphisms in *Mycobacterium tuberculosis* Conferring Resistance to Delamanid in Drug-Naive Patients. *Antimicrobial Agents and Chemotherapy* **64**, 1–5 (2020).

- [15] Vargas, R. *et al.* Role of Epistasis in Amikacin, Kanamycin, Bedaquiline, and Clofazimine Resistance in *Mycobacterium tuberculosis* Complex. *Antimicrobial Agents and Chemotherapy* **65** (2021).
- [16] Kadura, S. *et al.* Systematic review of mutations associated with resistance to the new and repurposed *Mycobacterium tuberculosis* drugs bedaquiline, clofazimine, linezolid, delamanid and pretomanid. *The Journal of antimicrobial chemotherapy* **75**, 2031–2043 (2020).
- [17] Peretokina, I. V. *et al.* Reduced susceptibility and resistance to bedaquiline in clinical *M. tuberculosis* isolates. *Journal of Infection* **80**, 527–535 (2020).
- [18] Karboul, A. *et al.* Frequent Homologous Recombination Events in *Mycobacterium tuberculosis* PE/PPE Multigene Families: Potential Role in Antigenic Variability. *Journal of Bacteriology* **190**, 7838–7846 (2008).
- [19] Phelan, J. E. *et al.* Recombination in *pe/ppe* genes contributes to genetic variation in *Mycobacterium tuberculosis* lineages. *BMC Genomics* **17**, 151 (2016).
- [20] Akhter, Y., Ehebauer, M. T., Mukhopadhyay, S. & Hasnain, S. E. The PE/PPE multigene family codes for virulence factors and is a possible source of mycobacterial antigenic variation: Perhaps more? *Biochimie* **94**, 110–116 (2012).
- [21] Copin, R. *et al.* Sequence Diversity in the *pe_pgrs* Genes of *Mycobacterium tuberculosis* Is Independent of Human T Cell Recognition. *mBio* **5**, 1–11 (2014).
- [22] De Maio, F., Berisio, R., Manganelli, R. & Delogu, G. PE_PGRS proteins of *Mycobacterium tuberculosis*: A specialized molecular task force at the forefront of host-pathogen interaction. *Virulence* **11**, 898–915 (2020).

- [23] Meehan, C. J. *et al.* Whole genome sequencing of *Mycobacterium tuberculosis*: current standards and open issues. *Nature Reviews Microbiology* **17**, 533–545 (2019).
- [24] Doyle, R. M. *et al.* Direct Whole-Genome Sequencing of Sputum Accurately Identifies Drug-Resistant *Mycobacterium tuberculosis* Faster than MGIT Culture Sequencing. *Journal of Clinical Microbiology* **56**, JCM.00666–18 (2018).
- [25] Gliddon, H. D. *et al.* A Rapid Drug Resistance Genotyping Workflow for *Mycobacterium tuberculosis*, Using Targeted Isothermal Amplification and Nanopore Sequencing. *Microbiology Spectrum* **9**, 1–12 (2021).
- [26] Coscolla, M. & Gagneux, S. Consequences of genomic diversity in *Mycobacterium tuberculosis*. *Semin Immunol* **26**, 431–444 (2014).
- [27] Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
- [28] Johnson, E. O. & Hung, D. T. A Point of Inflection and Reflection on Systems Chemical Biology. *ACS Chemical Biology* **14**, 2497–2511 (2019).
- [29] Phelan, J. E. *et al.* Integrating informatics tools and portable sequencing technology for rapid detection of resistance to anti-tuberculous drugs. *Genome Medicine* **11**, 41 (2019).
- [30] Ruhwald, M., Carmona, S. & Pai, M. Learning from COVID-19 to reimagine tuberculosis diagnosis. *The Lancet Microbe* **2**, e169–e170 (2021).