Universidade de Lisboa

Faculdade de Farmácia

Genomic analysis of *Mycobacterium tuberculosis* clinical isolates from Angola: molecular determinants of resistance and integration in a global epidemiological scenario

João Sebastião Santos Nogueira

Dissertation supervised by Professor João Ruben Lucas Mota Perdigão and co-supervised by Professor Maria Isabel Nobre Franco de Portugal Dias Jordão

Master's degree in Biopharmaceutical Sciences

2022

Universidade de Lisboa

Faculdade de Farmácia

Genomic analysis of *Mycobacterium tuberculosis* clinical isolates from Angola: molecular determinants of resistance and integration in a global epidemiological scenario

João Sebastião Santos Nogueira

Dissertation supervised by Professor João Ruben Lucas Mota Perdigão and co-supervised by Professor Maria Isabel Nobre Franco de Portugal Dias Jordão

Master's degree in Biopharmaceutical Sciences

2022

Abstract

Mycobacterium tuberculosis (MTB), the causative agent of tuberculosis (TB), has evolved alongside its human host for millennia and continues to be one of the main causes of death worldwide, particularly so in developing and third-world countries. As standard drug resistance screening is both financially and logistically difficult for *M. tuberculosis*, other methods for drug resistance screening and outbreak monitoring have been developed, based on genetic markers of drug resistance and phylogenetics, supported by whole-genome sequencing (WGS). Although WGS-based methods have become standard in low-burden countries, genetics-based drug resistance screening carried by specialized software is still not 100% accurate.

Although Angola is considered a high-burden tuberculosis (TB) country by the World Health Organization, populational characterization remains relatively underdeveloped. This pilot study involves *M. tuberculosis* strains obtained in Angola, sequenced and characterized for drug resistance and phylogenetic contextualization within the scope of Portugal and the rest of the world through use of specialized software, along with confirmatory phenotypic drug resistance screening by determination of minimum inhibitory concentration, as well variant calling and single nucleotide polymorphism (SNP) clustering.

One of the studied strains presents phenotypic resistance to isoniazid that cannot be properly linked to a genetic mutation in known genes of interest, possibly indicating resistance caused by efflux pump overexpression. SNP clustering indicates that Angola strains appear to not possess a direct phylogenetic relationship with strains found in Europe, with clustering only observed at distances of 50 SNPs. At this clustering depth, some strains demonstrate an indirect link either within other African countries or with *M. tuberculosis* strains isolated in Portugal and Brazil, implying indirect chains of transmission and common ancestor, notable of which strains AO000110, AO000111, AO000113, AO000131, which cluster with strains isolated in Portugal, Brazil, Malawi, Canada and the United Kingdom. Furthermore, phylogenetic analysis of non-clustered Angola strains also indicates the existence of shared common ancestors with strains isolated in Portugal, Brazil, Malawi, Bangladesh, Australia, Canada and the United Kingdom. Taken together, this information implies a possible bridge between the community of Portuguese Speaking Countries and the Anglosphere. Efforts should continue to be undertaken in order to further characterize the phylogenetic background of *M. tuberculosis* in Angola to better understand these chains of transmission and further characterize the phylogenetic background of *M. tuberculosis* in the region.

Keywords: *Mycobacterium tuberculosis, whole genome sequencing, phylogenetics, drug resistance screening, Community of Portuguese-speaking Countries*

Resumo

Mycobacterium tuberculosis, o agente causador da tuberculose, tem evoluído com o seu hospedeiro principal, o ser humano, ao longo de milhares de anos, apresentando inclusivamente características que permitem uma demarcação regional e linhagens próprias. A tuberculose é definida pela Organização Mundial de Saúde (OMS) como uma das dez principais causas de morte a nível mundial, com particular peso em países emergentes e sub-desenvolvidos, onde o panorama socio-económico se reflecte em padrões de saúde pública aquém dos padrões encontrados em países ocidentais industrializados. Acresce a este panorama a dificuldade na obtenção de antibióticos em quantidades suficientes, resultando em regimes de tratamento incompletos, auxiliando a pressão selectiva que origina a propagação de estirpes resistentes aos fármacos antibióticos, reconhecidas pela OMS como um problema crescente nos esforços de erradicação da tuberculose. Estes padrões de resistência são definidos num número de patamares: estirpes multi-resistentes são definidas como estirpes apresentando resistência à rifampicina e isoniazida; estirpes pré-extensivamente resistentes apresentam um padrão multi-resistente, com resistência adicional a fluoroquinolonas; estirpes extensivamente resistentes são definidas como as de padrão semelhante às pré-extensivamente resistentes, desta feita com resistência adicional seja ao linezolide ou bedaquilina.

M. tuberculosis é uma bactéria com um ciclo de crescimento e propagação complexo que exige parâmetros específicos para crescimento, bem como instalações e pessoal especializado, nomeadamente compatíveis com o padrão de biosegurança de nível 3. Estas exigências, emparelhadas com a taxa de crescimento relativamente lenta de *M. tuberculosis*, levam a que a avaliação de resistência a antibióticos sejam especialmente demorados.

Por outro lado, o genoma relativamente estável de *M. tuberculosis*, estimado com uma taxa de mutação de 0.3 a 0.5 SNPs por genoma por ano, combinado com uma transmissão de mutações quase exclusivamente vertical, ou seja, de progenitor para progenia, permite uma caracterização genética eficaz de *M. tuberculosis* que, juntamente com princípios genéticos de obtenção de resistência, propicia a criação de modelos de previsão de padrões de resistência a antibióticos, ou seja, resistência genotípica, sem necessidade de instalações e pessoal com o nível de especialização exigido na testagem tradicional de padrões de resistência, dita testagem fenotípica. Adicionalmente, esta constelação de factores torna favorável a análise filogenética de *M. tuberculosis*, direccionada à monitorização de estirpes e seguimento de surtos. Actualmente, graças a avanços na área de sequenciação genética, permitindo a sequenciação de genomas de forma relativamente rápida e de baixo custo,

juntamente com avanços na compreensão das bases genéticas da resistência a antibióticos por parte de *M. tuberculosis*, culminando na criação de bibliotecas de mutações causadoras de resitências, a abordagem genotípica tornou-se um modelo de análise de padrões de resistência em alguns países onde a tuberculose não é uma doença de marcada frequência, como o Reino Unido e a Holanda.

No contexto da Comunidade de Países de Língua Portuguesa, Angola, Brasil, Guiné-Bissau e Moçambique são considerados pela Organização Mundial de Saúde como países de *high burden* (lit. peso elevado) de tuberculose. Notável também Timor-Lorosae devido à sua proximidade geográfica com a Indonésia, outro país do mesmo agrupamento. Com Portugal servindo de intermediário entre estes países e o resto da Europa, torna-se pertinente caracterizar o contexto em que *M. tuberculosis* se insere nestes países. Como tal, é importante obter amostras de doentes com tuberculose nestes países, isolar o agente causativo e sequenciar o seu genoma de forma a obter informação relativa a padrões de resistência antibiótica e contexto filogenético.

O trabalho desenvolvido neste estudo-piloto envolve o uso de 17 isolados obtidos no Hospital da Divina Providência, na zona de Kilamba-Kiaxi, em Luanda, Angola, entre Março e Junho de 2014, de acordo com os padrões definidos pela comissão de ética do ministério da Saúde Angolano e com consentimento informado dos doentes de quem estas amostras foram obtidas. Estes 17 isolados foram devidamente sequenciados, servindo a estirpe H37Rv como referência para o mapeamento efectuado. As sequenciações assim obtidas foram de seguida avaliadas em relação aos seus possíveis padrões de resistência através de *software* especializado, nomeadamente TB-Profiler, Phyres e Mykrobe, sendo os resultados obtidos comparados entre si e com resultados obtidos por testagem fenotípica, realizada de acordo com o método de referência definido pelo Comité Europeu sobre Testagem de Susceptibilidade Antimicrobiana (lit. European Committee on Antimicrobial Susceptibility Testing, EUCAST), de concentração mínima inibitória, para isoniazida, rifampicina, etambutol, estreptomicina, amicacina e ofloxacina, sendo que se verificaram um número de discrepâncias, nomeadamente os *software* testados não conseguiram prever a resistência de algumas estirpes à isoniazida, ou preveram resistências que não se verificaram em contexto fenotípico, salientando a desvantagem dos métodos genotípicos, nomeadamente o facto de dependerem de catálogos em constante actualização de mutações causadoras de resistências, cuja validade depende de despistagem por métodos fenotípicos ou de descoberta em contexto clínico.

Foi igualmente realizada uma análise filogenética das 17 estirpes em três contextos diferentes: entre si, em junção com estirpes isoladas em Portugal e, finalmente, das 17

estirpes estudadas, 13 delas, pertencentes à sub-linhagem 4.3, foram usadas em comparação com 1682 estirpes isoladas em múltiplos países, em contexto mundial, fornecidas pelo *European Nucleotide Archive* (ENA, lit. Arquivo Nucleotídico Europeu), também pertencentes à sub-linhagem 4.3. As estirpes foram agrupadas em árvores filogenéticas e comparadas para a formação de clusters através de distanciamento de SNPs, para distâncias de cinco, 12, 25, 50 e 100 SNPs, sendo que cadeias de transmissão directas são geralmente reconhecidas para distâncias de SNP entre cinco e 12 SNPs. Demonstrou-se que as estirpes isoladas não têm um relacionamento directo entre si. Quanto ao contexto a nível de Portugal, detectou-se clustering a distâncias de 50 SNP, indicativo de cadeias indirectas de transmissão. Igualmente, a nível mundial, cinco das estirpes estudadas agruparam em cluster, sendo que quatro delas (AO000110, AO000111, AO000113, AO000131) agruparam no mesmo cluster com 38 outras estirpes e outra das estirpes estudadas (AO000138) agrupou em cluster com três outras estirpes. As estirpes pertencentes a estes clusters são oriundas de Portugal, Brasil, Malawi, Canadá e do Reino Unido, indicando para uma possível ponte de ligação entre a Comunidade de Países de Língua Portuguesa e a anglosfera. Uma análise das árvores filogenéticas englobando as estirpes não agrupadas em cluster com estirpes na sua proximidade filogenética reforça esta ideia, com estas estirpes aparentando ter um ancestral comum com estirpes isoladas em Portugal, Austrália, Bangladesh, Brasil, Canadá, Malawi e o Reino Unido.

Todas as estirpes englobadas nos estudos filogenéticos foram igualmente avaliadas face aos seus perfis de padrões de resistência a fármacos antibióticos. Uma vez que não existe informação acerca de padrões de resistência fenotípica para todas as estirpes, previsõoes de padrões de resistência com base em análise por TB-Profiler foram utilizados para esta análise. Num âmbito geral, apesar da existência nos clusters formados de uma série de estirpes com padrões de multi- e resistência pré-extensiva, a maioria das estirpes agrupadas em cluster apresentam padrões de susceptibilidade a fármacos antibióticos. No entanto, uma vez que a informação aqui avaliada foi obtida através de software preditivo, é possível que um número de estirpes apresente resistência fenotípica a um número de fármacos antibióticos.

A realização deste trabalho envolveu ainda a criação de um número de *scripts* nas línguagens de programação bash e R, de forma a agilizar os processos envolvendo o software de apoio.

Os resultados obtidos neste estudo apontam para a pertinência e necessidade de caracterização da tuberculose em Angola, assegurando uma melhor compreensão das cadeias de transmissão e contexto filogenético da tuberculose na região, por conseguinte

contribuindo para melhor caracterização de cadeias de transmissão de *M. tuberculosis* noutras regiões do mundo.

Palavras-chave: *Mycobacterium tuberculosis*, sequenciação completa de genoma (WGS), filogenética, teste de resistência a antibióticos, Comunidade de Países de Língua Portuguesa.

Acknowledgements

Professor João Perdigão for the kindness and patience with which he has welcomed me. Professor Isabel Portugal, who fate's whimsy had accompanying me from one stage of my academic life into the next. Professor Cecilia Rodrigues for allowing me the opportunity to further my academic ambitions. Pedro Gomes for all the help at unpredictable times and unexpected turns. Everyone in the pathogenomics lab for welcoming me as one of their own from the very first day.

All my colleagues in Farmácia do Largo, Farmácia Estácio and Farmácia Ibéria for moral support and flexibility, never giving up on me.

Family. Mother, father, brother and sister. Grandmother.

Zé-Carlos, Nica, Gil, São.

Thank you.

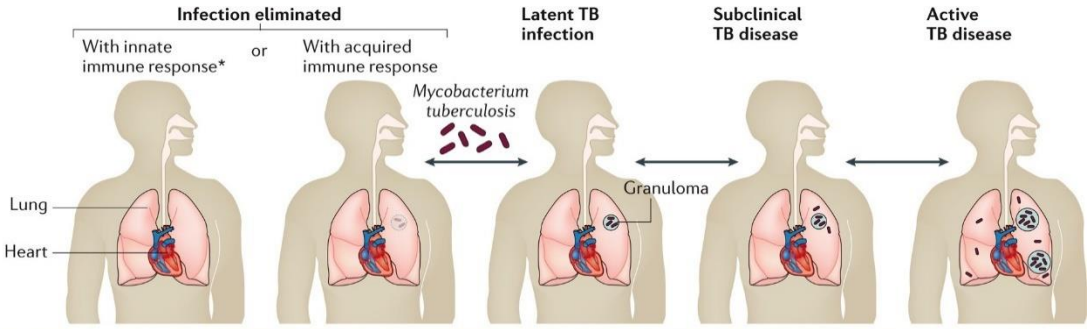
Index

1. Introduction

1.1. Tuberculosis

Tuberculosis (TB) is one of the oldest recorded infectious diseases in human history (1) and is still one of the worldwide leading causes of death to this day, accounting for an estimated 1.5 million deaths worldwide (2).

Disease presentation and symptomatology are commonly a chronic cough resulting in hemoptysis, fever and loss of appetite leading to weight loss (3). In most cases, TB is spread through inhalation of airborne particles generated by infected individuals (4). It should be noted that TB disease progression entails a clinical spectrum, and may exist in a latent state in which symptoms are not present and its infectious nature is null, with active disease capable of regressing back to latency (**Figure 1**) (3). This trait, along with a complex diagnosis protocol, makes TB particularly difficult to treat, especially in developing and third-world countries, in which healthcare resources and laboratory support are not to standard. Although it is believed a third of the world's population is infected with TB, it is a disease with a greater burden on third-world countries, with two thirds of recorded infections in 2018 occurring in India, Pakistan, Bangladesh, China, the Philippines, South Africa, Indonesia and Nigeria (2).



TST	Negative	Positive	Positive	Positive	Usually positive
IGRA	Negative	Positive	Positive	Positive	Usually positive
Culture	Negative	Negative	Negative	Intermittently positive	Positive
Sputum smear	Negative	Negative	Negative	Usually negative	Positive or negative
Infectious	No	No	No	Sporadically	Yes
Symptoms	None	None	None	Mild or none	Mild to severe
Preferred treatment	None	None	Preventive therapy	Multidrug therapy	Multidrug therapy

Figure 1 - From MTB infection to active TB disease. Note that TB progression is not a linear chain, with disease presentation varying according to patient fitness and treatment outcomes. Adapted from (3).

Although classified primarily as a lung infection, extrapulmonary tuberculosis may occur, accounting for an estimated 15% of total tuberculosis cases (5). Tuberculosis may affect organs and systems such as kidneys, gastrointestinal (GI) tract, the genitourinary tract and even the central nervous system (5).

Regardless of disease presentation, tuberculosis is caused by multiple species of microbial pathogens belonging to the *Mycobacterium tuberculosis* complex (MTBC), of which *Mycobacterium tuberculosis sensu stricto* (MTB) remains the responsible for most infections (6). MTBC are a number of mycobacterial species that exhibit particularly close genetic distance, indicating the existence of a common ancestral (6). Among MTBC, MTB and *Mycobacterium africanum*, have been recorded as infectious to human beings (7), though a minority of cases have been attributed to zoonotic members of the *M. tuberculosis* complex, such as *M. bovis* or *M. caprae* (3). Both MTB and *M. africanum* have shown an evolutionary parallel to humans', displaying adaptative mechanisms that provide optimal infectious capability according to environmental and host factors, resulting in marked phylogeographical differences (**Figure 2**) (8).

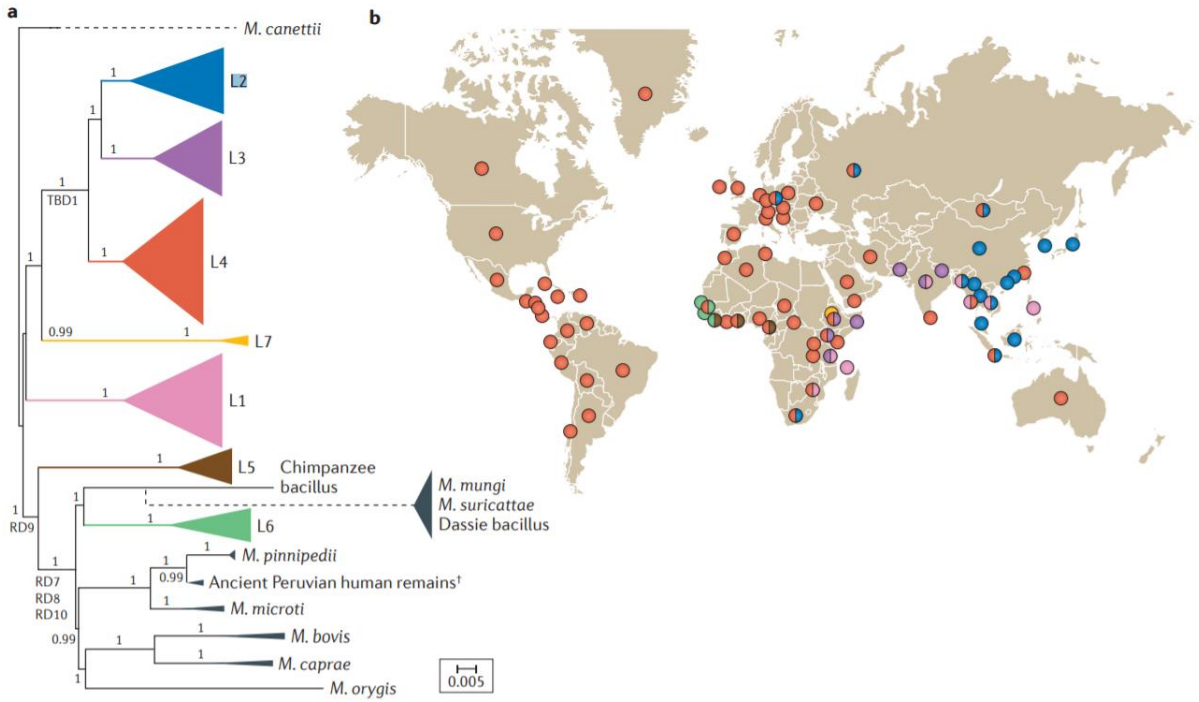


Figure 2 – Global phylogeography of MTBC. Colored lineages indicate human infectious character (9). Of note, the globally spread character of Lineage 4, presenting itself as the dominant lineage in Europe, the Americas, Australia and most of Africa.

1.2. Etiology and populational structure

Mycobacterium tuberculosis is an obligate-aerobic bacillus that, due to the high lipid content of its cell wall and presence of unique monomers (such as arabinogalactan and mycolic acids), is neither classified as gram-positive nor gram-negative (9), displaying acid-fastness that, like other mycobacteria, manifests in coloration by Ziehl-Neelsen staining (10). It is a catalase-negative, slow-growing bacteria that invades target cells, thus displaying a parasitic nature. It does not present motility nor is it spore-forming (10).

Due to its unique characteristics, particularly its cell wall structure (**Figure 3**), MTB are bacteria that display ability to survive in extreme environments while possessing intrinsic resistance to a number of antibiotics (10).

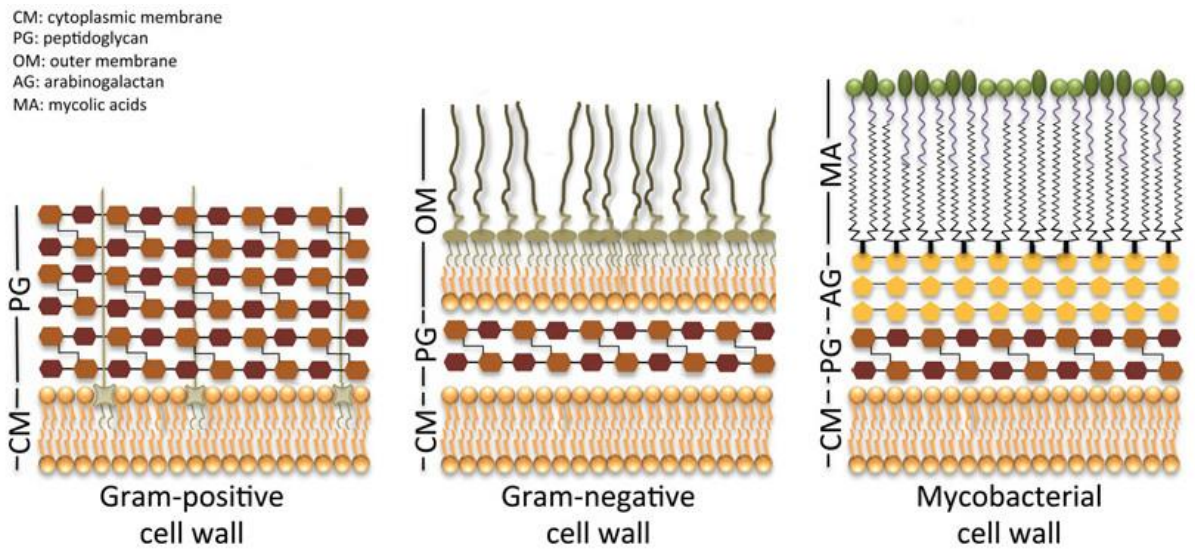


Figure 3 – Comparison of bacterial cell wall structures. From left to right: gram-positive, gram-negative and mycobacterial cell walls. Mycobacteria present a cell wall structure that sets it apart from other bacteria. Note the presence of arabinogalactan (AG) and mycolic acids (MA), the latter of particular importance in MTB survivability and as a target of multible drugs. Adapted from (12).

Other species within the MTBC exhibit the same properties, as well as infectious character towards humans despite a marked preference for specific host species. However, although infection of non-preferential species is possible, mycobacteria are seldom contagious from the newly infected organism towards other members of the same species, indicating an evolutionary pressure towards intra-species specialization (8).

As MTB has evolved alongside its human host, it stands to reason that multiple lineages with distinct regional demarcations have developed over the millennia, reflecting the selective

pressures between host, pathogen and their environment (11). MTB was first subdivided into three Principal Genetic Groups (PGG1-3) based on single nucleotide substitutions in the *katG* and *gyrA* genes (12). Today, MTB is subdivided into nine different lineages, with lineages 1 through 4 (L1-L4) being the most clinically relevant, defined along with L7 as MTB *sensu stricto* (13).

These strains are organized in accordance with their phylogenetic associations. L1 is designated as the indo-oceanic lineage, while L2 is east asian, L3 east african/indian and L4 the euro-american lineage (14). Lineages 5 and 6 are comprised of *M. africanum*, lineages endemic to west Africa (also known as west African 1 and west African 2, respectively) that show particular adaptation to the local population (15), hardly ever found in immigrant population and with outbreaks outside of Africa isolated to population of African ancestry (16). Lineage 7 appears to be of Ethiopian origin and endemic to the horn of Africa, with traits that imply an evolutionary link between L1 and L2-3-4 (17). Lineage 8, recently discovered, has also been identified in the African great lakes region (18). Finally, lineage 9 has recently been characterized, with genetic proximity to L5 and L6 but, unlike L5 and L6, with origin traced to east Africa (13). It should be noted that Africa is the only continent where all lineages can be found. Coupled with historic data and populational vulnerability to TB observed in the 19th and 20th centuries, the presence of multiple lineages has been correlated with a "virgin soil" hypothesis, theorizing the African continent has been exposed to worldwide variants due to historic migratory paths encopassing the region (19). However, other authors, supported by recent development in molecular dating, theorize an "out of Africa" model of MTB evolution, mimicking humanity's own evolutionary pattern (20).

Different lineages have been associated with differences in disease presentation, with L1, considered the oldest lineage, showing lower rates of virulence when compared with L4, the most globally spread lineage, likely due to migration pressures and historic trade routes to and from Europe (14,21). Lineages 5 and 6 present slower growth rates as well as slower, attenuated disease progression, showing preference for latency (22). Lineage 7 shows slower disease progression, akin to *M. africanum* (23). All lineage 8 strains found show resistance to isoniazid, ethionamide as well as rifampicin (17), though it should be noted that not enough strains of both L8 and L9 have been isolated to estimate defining characteristics insofar as disease presentation. Furthermore, some sub-lineages appear to display preference towards specific clinical phenotypes. A worldwide study based on 490 strains associated L5, L6, L1.1.1, L1.2.1, L4.1.2.1 and ancient Beijing sublineages with extra-pulmonary forms of disease, while L4.3.1, L4.3.3, L4.5, Asian African 2 and Europe/Russia B0/W148 modern Beijing sublineages appeared associated with pulmonary disease. Additionally, L3 and modern Beijing sublineages

appeared associated with antibiotic resistance (24). A multi-center study based on 2092 MTB isolates collected from Riyadh, Saudi Arabia, associated *M. africanum* with extra-pulmonary disease, L3 with central nervous system disease phenotype, as well as an association between Uganda-I and gastrointestinal tuberculosis (25).

1.3. Infection cycle

In general terms, as already mentioned, the main route of TB infection is via inhalation of airborne particles, typically of 1 to 3μm in size (26). After invasion of the pulmonary alveoli, the bacteria are first phagocytized by resident macrophages, consequently confined to phagosomes (27). The bacteria then block maturation of the phagosome by preventing its fusion with lysosomes, avoiding the generation of an environment unfit for bacterial survival (28). Nevertheless, infected macrophages are able to recruit other immune cells to help combat the infection through cytokine production. As mycobacterial growth proceeds unhindered, necrosis of infected macrophages allows the release of bacteria into the extracellular space, wherein a new cycle of macrophage invasion may begin (29). Depending on host fitness, the disease may progress to other areas of the body through migration of infected macrophages either by translocation through the alveolar wall, migration across alveolar barriers and invasion via the lymphatic system (30), or remain contained to its primary sites of infection in its attenuated, latent form. The immune cascade generated by MTB infection eventually culminates into the generation of granulomae characteristic of this infection and usually found in chest x-rays of infected individuals (31).

1.4. Tuberculosis treatment

The World Health Organization (WHO) has published guidelines for treatment and management of TB that are regularly updated according to new findings and advances in treatment and diagnosis. Treatment options are divided according to illness progression and drug susceptibility, as well as the patient's age.

1.4.1. Latent TB

The WHO recommends ruling out active TB diagnosis through a clinical algorithm. Those not manifesting symptoms are unlikely to have active TB. Testing for latent TB *per se* is recommended through tuberculin skin test or interferon-gamma release assay (IGRA).

The standard, daily isoniazid monotherapy for six months is recommended both for adults and children in both high-incidence and low-incidence countries. Alternative treatment options for latent TB vary according to TB incidence in the patient's country: in high incidence countries, weekly rifapentine and isoniazid for three months may be offered to both adults and children, while children under 15 may also be offered daily rifampicin plus isoniazid for three months; in low incidence countries, treatment options include extending the isoniazid treatment to nine months, combined weekly isoniazid and rifapentine for three months and either daily rifampicin on its own or in combination with isoniazid for three to four months (32).

1.4.2. Active TB

Treatment of active TB is divided into an initial, intensive phase, followed by a continuation phase, and depends on patient history, namely if the patient is presenting TB for the first time (new TB case) or if the patient has been previously treated for TB.

Treatment of new cases begins with two months combined therapy of daily isoniazid, rifampicin, pyrazinamide and ethambutol, followed by four months of daily isoniazid and ethambutol (2HRZE/4HR). However, in population where it is known a priori or suspected that TB presents isoniazid resistance, ethambutol may be added to the continuation phase. Alternatively, the continuation phase may be administered three times weekly rather than daily, so long as each dose is directly observed. If the patient has not contracted human immunodeficiency virus (HIV) or lives in an HIV-prevalent setting, the intensive phase may also be administered thrice weekly (33).

Previously treated patients who remain with active TB or relapse after treatment are presupposed to be infected with drug-resistant variants.

1.4.3. Drug-resistant TB

Drug-resistant TB is classified by tiers, depending on the type of antibiotic drug a strain may be resistant to. The WHO has recently updated the typology of TB drug resistance in order to better differentiate stages of drug resistance, as well as to improve treatment methodology - TB may be classified as susceptible, when it presents no resistance to any anti-TB drug, rifampicin-resistant (RR-TB) when resistance to rifampicin occurs, multi-drug resistant (MDR-TB) when presenting resistance to both rifampicin and isoniazid, pre-extensively drug-resistant (pre-XDR-TB) when resistance to either rifampicin, or both isoniazid and rifampicin, presents itself along with resistance to any fluoroquinolone, and extensively-drug resistant (XDR-TB) when the pre-XDR resistance pattern is observed along with resistance to either linezolid or bedaquiline (34).

Treatment of active, drug-resistant TB is wholly dependent on available resources, specifically concerning the availability of effective drugs and sophistication of drug susceptibility testing (DST).

In contexts in which DST is not available, an empiric approach is recommended, based on data characterizing TB in the region, such as surveys of treatment effectivity of similar patients and whether or not the drug is commonly used in the area. The WHO recommends the use of at least four drugs certain to be effective, to avoid drugs for which cross-resistance may play a role, as well as to guide drug selection by a potency-based hierarchy. Whenever it becomes available, DST should be performed and the treatment regimen changed according to the results obtained.

Whenever DST is routinely available, treatment should be guided by the above-stated principles while results are pending. Current treatment may be continued or changed, depending on results.

Whenever rapid DST is available, empirical or individualized treatment may be provided. Considering rapid DST takes about two days to provide results, there is little to no necessity to switch drugs mid treatment (33).

In accordance with the WHO's recommendations (**Table 1**), supported by research concerning each anti-TB drugs' efficacy, the WHO has organized drugs by priority classes (groups/steps) based on their effectivity against drug-resistant forms of TB (35).

Table 1 - WHO recommendations on treatment of tuberculosis. Anti-MTB drugs are presented in order, according to their effectivity against multi-drug and extensively-drug-resistant MTB.

Group/steps	*Drugs (abbreviations)*
Group A – **Include all three medicines**	• Levofloxacin/Moxifloxacin (Lfx/Mfx) • Bedaquiline (Bdq) • Linezolid (Lzd) • Pyrazinamide (Z) • Rifabutin (Rfb) • Rifapentine (Rpt)
Group B – add one or both medicines	• Clofazimine (Cfz) • Cycloserine/Terizidone (Cs/Trd) • Amikacin (Am) • Capreomycin (Cm)
Group C – add to complete the regimen when medicines from previous groups cannot be used	• Ethambutol (E) • Delamanid (Dlm) • Pyrazinamide (Z) • Meropenem/Imipenem-cilastatin (Mpm/Ipm-Cln) • Amikacin/Streptomycin (Am/S) • Ethionamide/Prothionamide (Eto/Pto) • P-aminosalycilic acid (PAS)

1.5. Anti-MTB drugs and resistance mechanisms

Along the years, a combination of widespread use of anti-TB compounds, incomplete treatments, relapses and migratory flows among other selective pressures have culminated in the generation of drug resistance mechanisms, rendering standard treatment strategies ineffective in controlling outbreaks. Accumulation of these resistance mechanisms has led to the emergence of MDR- or XDR-TB, which, if left unmonitored and untreated, may lead to the exhaustion of healthcare resources and loss of life, particularly in countries with socioeconomic issues reflecting in debilitated healthcare infrastructure. Below, the mechanism of action of the most widely used anti-TB drugs for treatment of both susceptible and multidrug resistant is explained, along with drug resistance mechanisms employed by resistant TB strains.

1.5.1. Isoniazid

Used as an anti-TB drug since 1952, isoniazid is a prodrug that enters the mycobacterial cytoplasm via passive diffusion. Although it was known isoniazid is converted to its active form by the mycobacterial catalase and peroxidase (KatG) (36), the specific mechanism of action was unclear. It was known that mycolic acid synthesis, paramount in the upkeep of cell envelope permeability, MTB virulence and overall antibiotic resistance was inhibited by INH (37). However, only in 2006 was it understood that inhibition of mycolic acid synthesis was the main driver of INH toxicity (38). It became understood that isoniazid inhibits mycolic acid synthesis, by selectively targeting enoyl acyl carrier protein reductase (InhA) (39). Furthermore, it is known that INH's structural simplicity leads to indiscriminate incorporation into, and inactivation of multiple processes, such as nucleic acid synthesis (40), protein synthesis (41), and carbohydrate synthesis (42), boosted by generation of free radical species. It should be noted that isoniazid first exerts bacteriostatic activity and only after a period of approximately 24 hours does it manifest bactericidal activity (43). Further, isoniazid only exhibits bactericidal activity against actively dividing bacteria in aerobic conditions (43).

Known resistance mechanisms against isoniazid action are related to this drug's targets and pro-drug activation pathways. Genomic mutations leading to conformation changes which compromise binding of INH, or promoter region mutations leading to overexpression of enzymes that counter INH's direct effects on MTB are the most common findings. The most frequently INH resistance-conferring mutation found in clinical isolates involves mutations in the gene *katG*, causing conformation changes in the protein it codifies, KatG (44). This gene is found in a region of high variability containing repetitive DNA sequences, leading to point mutations and deletions. The most common *katG* mutation found in clinical isolates is a substitution of serine by threonine in position 315 (45).

Apart from mutations in *katG*, mutations in the promoter region of *inhA*, leading to InhA overexpression, also confer resistance to INH (46). Likewise, overexpression of AhpC caused by mutations in *ahpC*'s promoter region also lead to resistance by compromising INH activation-derived free radical activity (47). Both these mutations aren't as common and provide lower-level resistance comparative to *katG* mutations, with *inhA* mutations in particular often associated with *katG* mutations as a possible compensatory mechanism *vis-a-vis* loss of catalase peroxidase activity derived from *katG* mutations (48).

Another mutation which compromises the redox pressure exerted by activated INH occurs in the *ndh* gene, coding NADH dehydrogenase, compromising the drug's action against InhA, resulting in low level resistance (49). Other mutations have been found associated with

INH resistance, such as *nat*, *fbpC* or *fabD*, but have yet to be found unassociated with other INH-resistance conferring mutations (50).

1.5.2. Rifampicin

Rifampicin interacts with the beta subunit of RNA-polymerase in MTB, thus interfering with DNA transcription. It enters MTB via passive diffusion (51). Hence, any changes in the beta subunit of RNA-polymerase which affect its bonding with RIF will influence this drug's antibiotic effect. Concordantly, mutations in the gene *rpoB*, paramount in RIF-RNApol binding, have been found in over 95% of RIF resistant strains (52). Of these mutations, about 90% occur in a region that has been defined as the RIF resistance-determining region (RRDR) (45). Other mutations of disputed significance have first been reported in treatment relapse patients in Kinchasa and Bangladesh (53) and since related to first-time tuberculosis patients (52). These mutations are especially noteworthy for remaining undetected by WHO-endorsed, gold standard drug resistance assays (54). As a result of *rpoB* mutation, the RIF-RNApol bonding is neutralized, transcription is not inhibited and the bacteria survives. Finally, about 5% of RIF resistant bacteria have shown no mutations in *rpoB*, instead exhibiting overexpression of efflux pumps, coded by genes *Rv0783*, *Rv2936* and *Rv0933* (55).

1.5.3. Pyrazinamide

Pyrazinamide's exact mechanism of action remains to be fully understood. It's known to be a prodrug metabolized into pyrazinoic acid by pyrazinamidase and that it only shows activity in acidic pH. It's hypothesized that PZA enters MTB by passive diffusion, is then converted to its active form, pyrazynoic acid (POA), is then expelled to the extracellular media where, upon protonation, it is once again absorbed by MTB by passive diffusion, afterwards exerting its effect on the bacteria (**Figure 4**) (56). PZA is particularly effective in MTB's dormant state, showing inferior toxicity against actively multiplying bacteria. Although its mechanism of action is not clear, resistance to PZA has been observed and attributed to mutations in specific genes. Mutations in *pncA* are the most common, representing almost 99% of resistant isolates studied. This gene is responsible for pyrazinamidase (PncA) expression, the key enzyme in activating PZA into POA. Mutations in this gene compromise PZA conversion to its active form, ergo stopping POA from exerting its effect in the bacteria. Mutations affecting other genes are rarely found and have only been defined in *rpsA*, which may cause changes in PZA's binding site (57), as well as in *panD*, involved in the synthesis of pantothenate and coenzyme-A, possible targets of the active form of PZA (58).

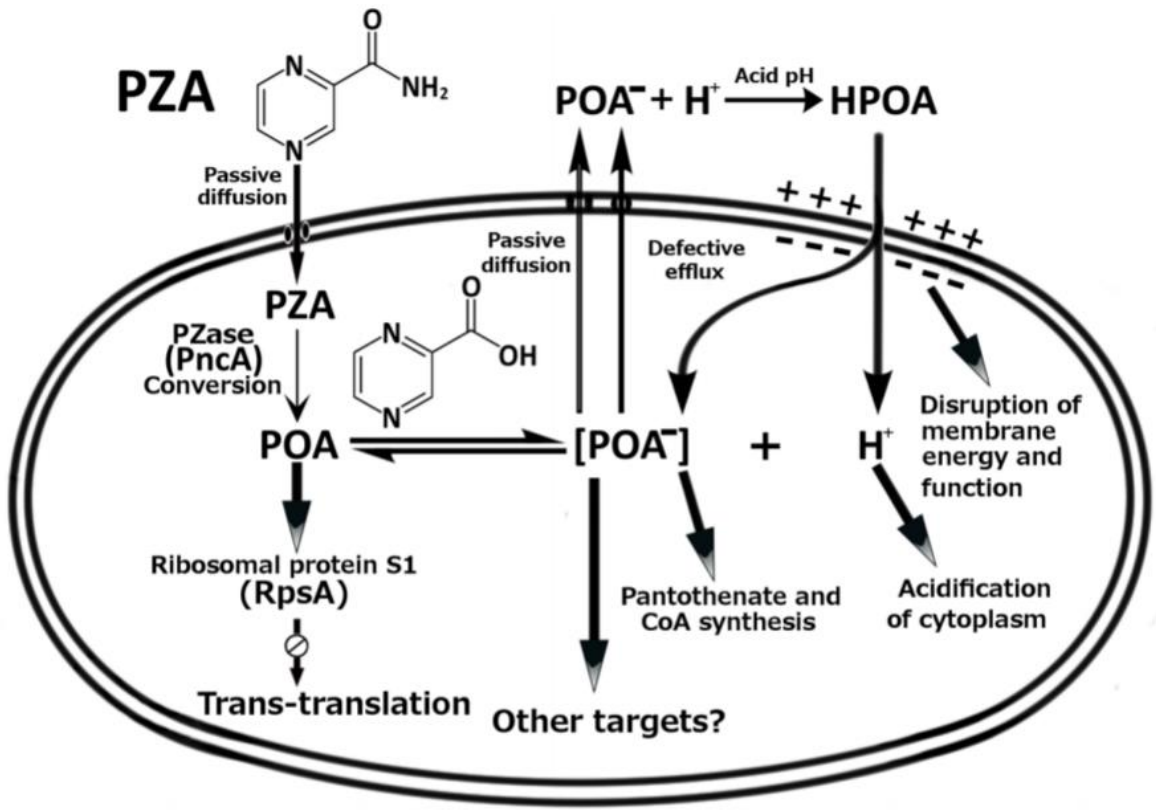


Figure 4 – Proposed mechanism of action of pyrazinamide. The prodrug enters MTB through passive diffusion, wherein it is converted to its active form, pyrazinoic acid (POA), by PncA. It then diffuses back to the extracellular space through passive diffusion. The acidic pH of the mycobacterial periphery stabilizes POA, allowing its re-entry in MTB (56).

1.5.4. Ethambutol

Ethambutol inhibits membrane-embedded arabinosyl transferases EmbA, EmbB and EmbC. All these enzymes work in unison in synthesis of arabinogalatan and lypoarabinomannan synthesis, both mycobacterial-specific cell wall components, with lipoarabinomannan also acting as an immunomodulator via interaction with host dendritic cell-specific intracellular adhesion molecule 3 (ICAM-3), facilitating dendritic cell infection (59). It is understood that EmbA and EmB are involved in arabinogalactan synthesis (60), while EmbC is involved in the synthesis of the arabinan portion of lipoarabinomannan (61). Inhibition of these arabinosyl transferases compromises cell wall synthesis, ultimately causing accumulation of mycolic acid (62). Consequently, mycobacterial immune escape and cell wall integrity are compromised, not only leading to cell death on its own but also allowing easier entry of other anti-MTB drugs.

Resistance to ethambutol arises from mutations leading to overexpression and/or conformational changes of EMB targets. The most common mutations are found in the *embCAB* gene cluster, which code a number of arabinosyltransferases. Both overexpression

and structural mutations lead to EMB resistance, with mutations most commonly found in *embB* (63). Furthermore, compromise of EmbCAB activity leading to overexpression of decaprenyl-phosphate 5-phosphoribosyltransferase (DPPR synthase, also known as UbiA), another enzyme involved in cell wall synthesis, has been found to cause EMB resistance (64).

Mutations in the *pknH-embR* axis also lead to EMB resistance, as *embR* regulates *embCAB* expression while *pknH* fosforilates the product of *embR*, leading to its activation. Mutations in this axis aren't common and are typically associated with mutations in *embB* (65). A number of other mutations have also been associated with EMB, among which efflux pump mutations. However, these mutations are rare and mostly associated with mutations in *embCAB* (66).

1.5.5. Aminoglycosides and Tuberactinomycins

Aminoglycosides used in MTB therapy include streptomycin, kanamycin and amykacin. This antibiotic drug class acts by binding to the 16S rRNA of the 30S ribosomal subunit, interfering with protein synthesis by inducing mistranslation on tRNA delivery (67). Capreomycin, an injectable tuberactinomycin, shares the same mechanism of action and route of administration, leading to its association as a drug class with aminoglycosides (68).

Cross-resistance among this drug class is incomplete, with some mutations leading to resistance to some drugs of this class while others remain active against MTB. The resistance mechanism affecting all drugs of this class is related to the structural conformation of the 30S subunit, compromising the binding of these drugs with their target, caused by mutations in *rrs*, a gene which codifies 16S rRNA (69). Another mutation related with the ribosome's structure occurs in *rpsL*, which codes the ribosomal protein S12, which confers resistance only to streptomycin (69). Capreomycin resistance may also be caused by mutations in *tlyA*, which codifies rRNA methyl-transferase, particularly loss-of-function mutations (70). Finally, kanamycin may also be ineffective against MTB due to mutations in the promoter region of *eis*. Eis is an acetyl-transferase, which causes multiacetylation of aminoglycosides, with greater affinity towards kanamycin. Mutations in its promoter lead to its overexpression and subsequent drug inactivation (71).

1.5.6. Fluoroquinolones

Levofloxacin and moxifloxacin are the most commonly used fluoroquinolones in TB treatment. These antibiotics inhibit the action of topoisomerase IIA, an enzyme which causes transient double-strand breaks in replicating DNA, hence essential in successful DNA replication (72). Any structural changes to this protein that incapacitate its binding to fluoroquinolones will ultimately lead to resistance. Mutations in *gyrA* and *gyrB* account for fluoroquinolone resistance, with a specific region in both genes accounting for most FQ-resistant clinical isolates. This region is defined as the fluoroquinolone-resistance-determining region A and B (QRDR-A, QRDR-B), respectively for *gyrA* and *gyrB*, with QRDR-A mutations being the most commonly found (73). QRDR-B mutations appear less frequently, with fluoroquinolone-resistance conferring mutations in *gyrA* and *gyrB* outside these regions being even less frequent (74).

1.5.7. Ethionamide

Ethioniamide is a second-line anti-TB drug which, much like isoniazid and pyrazinamide, is a pro-drug which is converted to its active state by the EthA enzyme, also known as EtaA. Like isoniazid, it forms ETO-NAD adducts which inhibits InhA function, preventing mycolic acid synthesis (75). ETO is particularly effective in *katG* mutation-derived INH-resistant strains. However, it also shares a number of resistance mechanisms with isoniazid, namely mutations in *inhA* and *ndh* (76). Mutations to *ethA* are also an important source of ETO resistance, with high mutational diversity found in clinical isolates in multiple regions of the gene showing no discernible pattern (76). Overexpression of EthR through mutations in *ethR* inhibit expression of EthA, causing ETO to remain in its inactive state (77). Finally, mutations in *mshA* and *mshC*, involved in mycothiol synthesis, playing a role in detox pathways, are found to lead to ETO resistance (78). Though the resistance mechanism remains unclear, it is understood these mutations cause a decrease in mycothiol production. In experimental conditions, *nudC* was found to degrade INH-NAD and ETO-NAD adducts, but it's important to note that *nudC* is an inactivated gene in MTB H37Rv (79).

1.5.8. Linezolid

Linezolid binds to the peptidil transferase center within the 50S subunit of the bacterial ribosome, inhibiting translation by impeding binding of tRNA (80). Resistance to linezolid is attributed to mutations in *rrl* and *rplC* (81). Mutations in *rrl*, the gene coding 23S rRNA, and in

rplC, coding for the L3 ribosomal protein, both part of the 50S ribosomal subunit, avert linezolid binding due to structural alteration to its binding site (80,82)

1.5.9. Bedaquiline

Approved as an anti-TB drug for conditional use by the food and drug administration (FDA) in 2012, bedaquiline was the first such drug in 40 years. Resistance to bedaquiline involves alterations in ATP and efflux pump gene expression. Missense mutations in *atpE*, which encodes subunit C of ATP synthase, has been found to produce marked increases in bedaquiline's MIC (84). Mutations in *Rv0678*, a transcriptional repressor the MmpS5-MmpL5 efflux pump gene expression, consequently leading to overexpression of this efflux pump, has also shown to worsen bedaquiline's effectiveness (85). Less understood is the mechanism of drug resistance generated by mutations in the *pepQ* gene, understood to code a putative Xaa-Pro aminopeptidase. However, mutations in *pepQ* lead only to low-level resistance, not completely compromising bedaquiline effectiveness (86).

1.6. Drug resistance screening

There are currently multiple drug resistance screening approaches available for MTB, with varying degrees of specificity and sensitivity. These screening methods can be divided in three groups: phenotypic drug susceptibility testing (pDST), genotypic drug susceptibility testing (gDST) and *in-silico* methods based on whole genome sequencing (WGS).

1.6.1 Phenotypic drug susceptibility testing

According to the newest guidelines, the WHO defines both liquid and solid media assays for DST based on critical concentrations (CC) (87). The main point of contention in regard to these methods is that CCs are based on experimental results by comparison of *a priori* defined "resistant" and "wild-type" strains according to clinical outcomes, rather than

based on pharmacokinectic and/or pharmacodynamic principles. Furthermore, these CCs vary according to the type of pDST performed, mostly due to the medium used (**Table 2**) (88).

Table 2 - Critical concentrations (mg/mL) used in various pDST techniques for MTB. Adapted from (88). LJ - Lowenstein-Jensen; 7H10, 7H11 - synthetic Middlebrook media; MGIT - mycobacteria growth indicator tube; eXiST – extended individual drug susceptibility testing; qAST – quick antimicrobial susceptibility testing

Antimicrobial agent	LJ	7H11	7H10	MGIT	MGIT TB eXiST qAST		
					Low	Intermediate	High
Isoniazid	0.2	0.2	0.2	0.1	0.1	0	10
Rifampicin	40	1	1	1	1	4	20
Ethambutol	2	7.5	5	5	5	12.5	50
Streptomycin	4	2	2	1	1	4	20
Kanamycin	30	6	5	2.5	0	0	0
Amikacin	30	0	4	1	1	4	20
Capreomycin	40	0	4	2.5	2.5	5	25
Ofloxacin	4	2	2	2	1	2	10
Moxifloxacin	0.5	0	0	0	0.25	0.5	2.5

The WHO recommends protocols both for solid, as well as liquid media. In regards to liquid media, the BACTEC MGIT system, based on UV fluorometry (89), is recommended (87). As for solid media, the WHO recommends the proportion method, which relies on comparison of bacterial growth via colony-forming unit (cfu) count between bacteria-free media and growth in media containing bacteria as well as anti-TB agents at varying concentrations (90). However, although both methods are susceptible to interpretation errors, the proportion method is moreso due to possible clumping of mycobacteria in suspensions (91).

Although the WHO has issued a set of recommendations for pDST, there is still no clear consensus between laboratories as to a reference MIC method (88). The European Committee on Antimicrobial Susceptibility Testing (EUCAST) has recently put forward a reference method based on MIC determination in order to harmonize epidemiologic cut-off values and standardize procedures at the european union level (92).

In general, pDST methods are considered the gold standard for drug resistance screening, as they are an empirical method of determining a MTB strain's drug resistance/susceptibility profile (90). However, these methods are time-consuming as they are dependent on MTB's growth rate and require tight biosafety measures, highly trained, specialized staff, and facilities to be performed, rendering their use limited in least wealthy

countries, which are usually also high-burden TB countries (87). Other options for DST are replacing pDST, particularly molecular-based approaches, specifically for first and second-line anti-MTB drugs (87).

1.6.2. Molecular drug susceptibility testing

Molecular-based approaches focus on detecting drug resistance through DNA amplification of known genetic markers of drug resistance. These tests can be categorized as either DNA line probe assays (LPA) or real-time PCR assays. Line probe assays make use of multiple probes to detect mutations through DNA-DNA hybridization (93). The WHO recommends the usage of GenoType MTBDRplus version 2 assay (Hain Lifescience, Nehren, Germany) and Nipro NTM+MDRTB detection kit 2 (Tokyo, Japan) for detection of resistance to first-line drugs, as both kits include probes for testing of isoniazid (*katG*, *inhA*) and rifampicin (*rpoB*) resistance; for second-line resistance testing, the WHO recommends the GenoType MTBDRsl version 2 assay (Hain Lifescience, Nehren, Germany), which includes probes for detection of drug resistance to fluoroquinilones (*gyrA*, *gyrB*) and aminoglycosides (*eis* promoter, *rrs*) (87).

Real-time PCR (qPCR) assays are capable of directly detecting DNA mutations associated with drug resistance by using sequence-specific probes containing a fluorescent reporter for detection of hybridation with the sample sequence (93). The WHO recommends the usage of Xpert MTB/RIF and MTB/RIF Ultra assays, wherein the main difference is heightened sensitivity of MTB/RIF Ultra in detection of rifampicin resistance as it uses two probes instead of one. The Truenat MTB, MTB Plus and MTB-RIF Dx assays are similarly recommended for detection of the same type of drug resistance, with the advantage of requiring minimal volumes of sputum sample to be performed (87, 94).

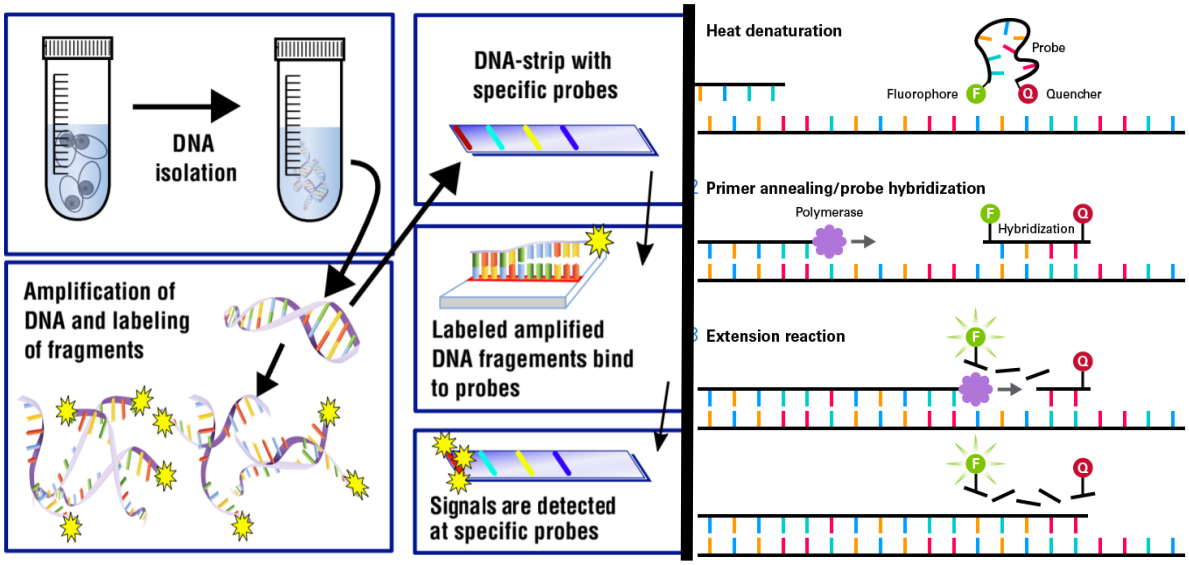


Figure 5 - Schematic of a line probe assay procotol (left) (94), and principle of qPCR assays (right) (95).

Figure 5 provides a general outline of the principles underlying the methods described above (94, 95). Unlike pDST, molecular-based methods do not require particularly specialized staff and can be performed in the same biosafety level conditions as sputum-smear microscopy. Furthermore, results can be obtained as soon as one hour after completing the procedure. However, false susceptibility results may be obtained if the resistance-conferring mutations are found in regions outside the used probes' scopes. It should be noted that the WHO currently only approves of qPCR testing in the detection of *rpoB*-related rifampicin resistance. Additionally, LPA sensitivity is inferior to pDST and require better infrastructure as well as higher degree of personnel specialization when compared to real-time PCR methods (87).

1.6.3. Drug resistance prediction through DNA sequencing

In 1977, Sanger created the first method of DNA sequencing, opening the way for full scale sequencing of genomes. This technique combined the dideoxy chain termination technique with sequencing of a DNA strand by synthesis with a complementary radioactive DNA strand, using polyacrylamide gel electrophoresis to analyze the resulting fragments (96). In its infancy, the technique was costly and slow, but improvements to the technique, such as replacement of radioactive nucleotides with fluorescent labeling (97), use of capillary gel electrophoresis (98) and PCR amplification techniques (99) allowed for increased yield at lower costs, culminating in the feasibility of commercialization of DNA sequencing of "first-generation" sequencing technology and the whole genome sequencing (WGS) of organisms (94).

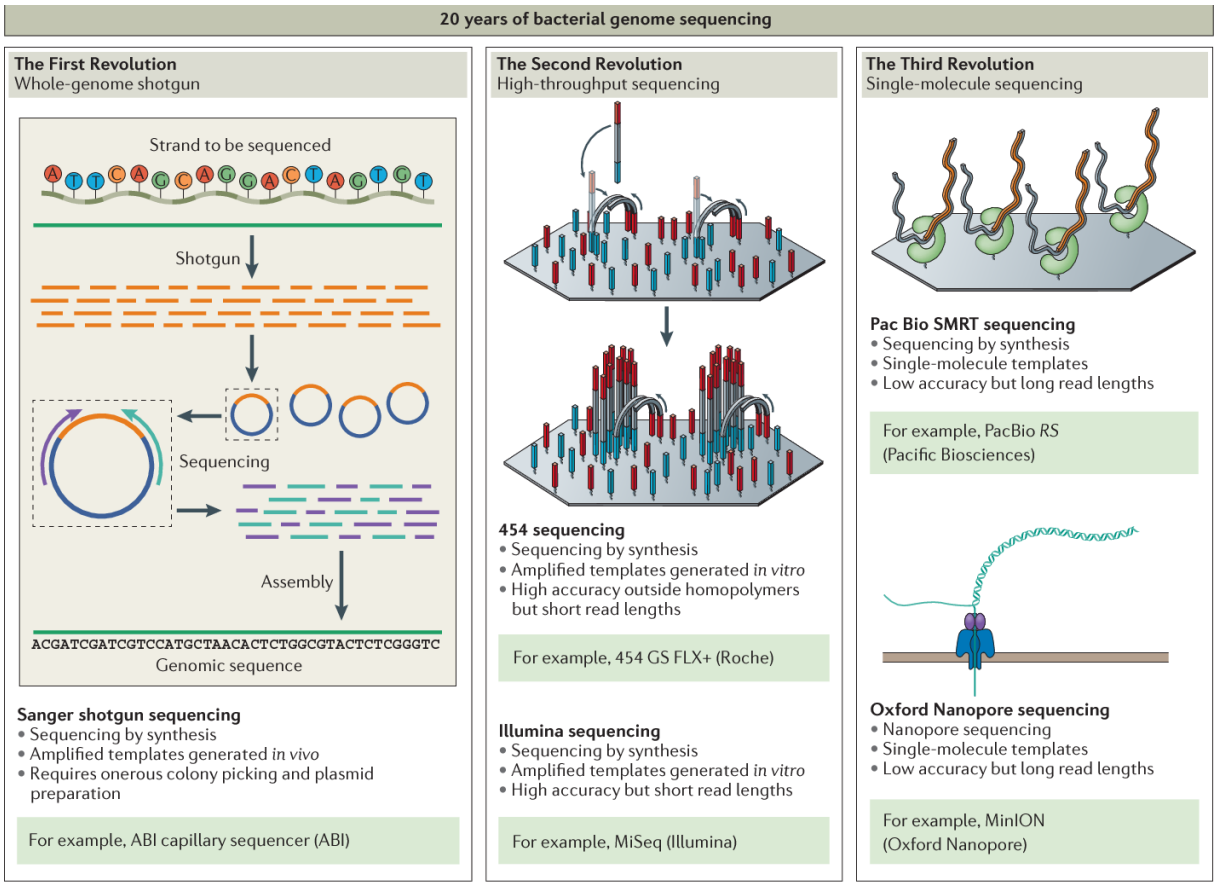


Figure 6 - The evolution of DNA sequencing technologies, with examples of commercial kits (100).

Figure 6 relays the progress made in DNA sequencing technology (100). Next-generation sequencing (NGS) technology allows for DNA sequencing of multiple DNA fragments simultaneously, without separation of each procedure. Such technique allows further increased yields at even lower costs. These techniques can be divided according to the template used for sequencing: amplification techniques based on DNA libraries, and single-molecule sequencing, with the former currently being most widespread (101). This is mostly due to the fact that, currently, DNA library-based sequencing is more accurate than single molecule-based alternatives, even though these produce longer read lengths that are computationally less demanding to assemble. Although each technique has its own advantages and challenges, they all share the same global steps of library preparation, amplification and sequencing, as summarized in **Figure 7** (101).

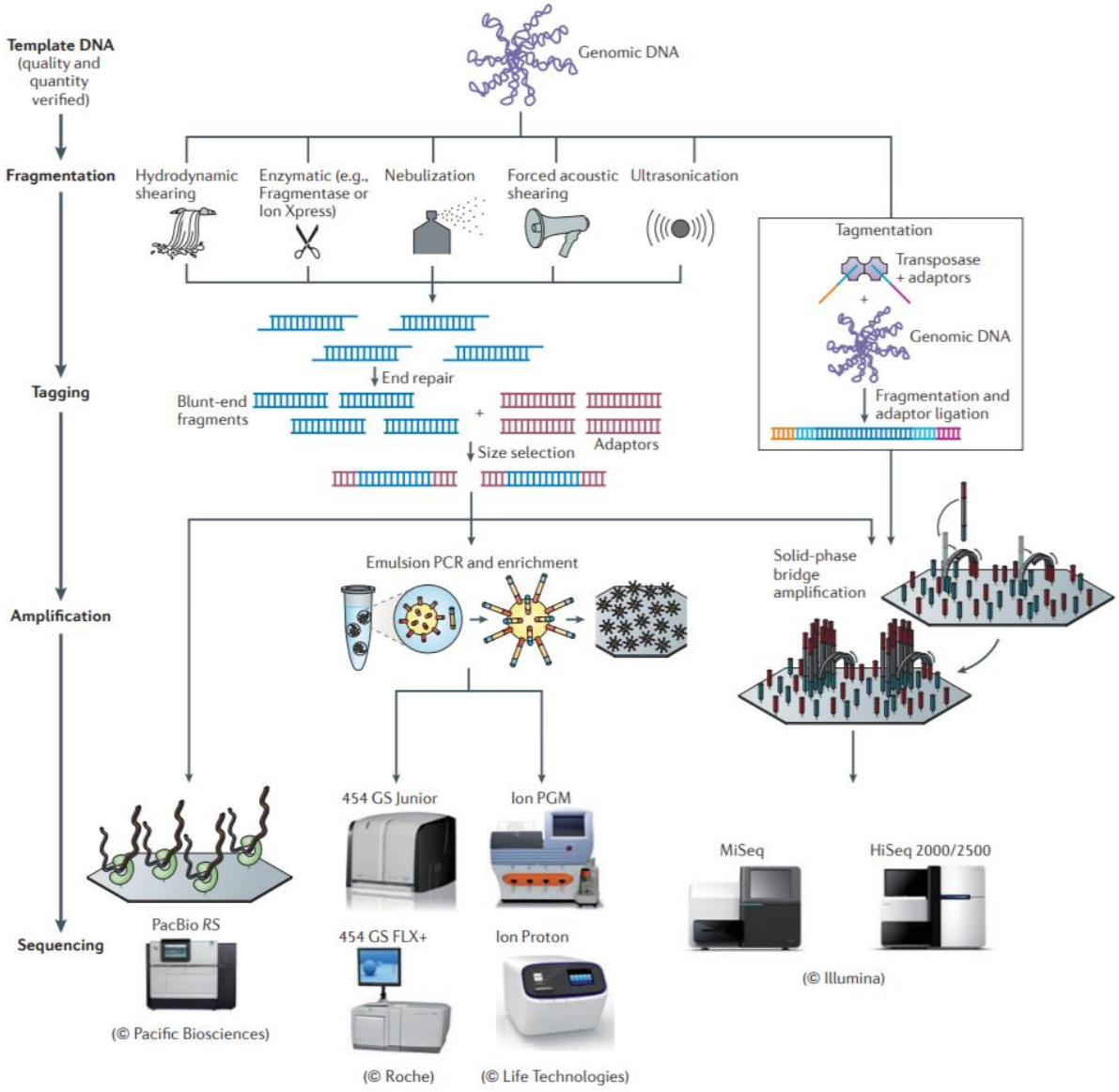


Figure 7 - Workflow of NGS techniques (58). Genomic DNA is fragmented by one of many procedures, either chemically or physically (fragmentation); DNA fragments are then selected for size and prepared for amplification (tagging); Fragments are amplified through PCR (amplification) and finally sequenced (58).

In the context of tuberculosis, the sequencing of an isolate's genome may be useful in the prediction of that sample's drug resistance profile. Furthermore, the data obtained can be used to unveil phylogenetic profiles, which allow for epidemiologic surveilance and monitoring of outbreaks. Through NGS, drug resistance moves from an *in vitro* confirmation test to an *in silico* predictive model, which reduces working times in as restrictive biosafety requirements as pDST and expands upon the principle behind LPA and PCR-based techniques. However, these techniques are limited to the genomic data obtained from previously isolated clinical strains with observable, phenotypic drug resistances and efforts are still underway in creating a global knowledge base of phenotypic drug resistance-conferring mutations (102).

Furthermore, some presumed resistance-conferring mutations occur in association with other, more common genetic markers of phenotypic resistance, such as *nat*, *fbpC* or *fabD* mutations occurring in association with *katG* mutations in isoniazid-resistant strains, complicating the process of defining true drug resistance-conferring mutations (50). NGS also requires specialized personnel and infrastructure and may be less cost-efficient in high-burden countries. Nevertheless, as DNA sequencing technology improves, so does its commercial availability. Multiple robust drug resistance-conferring catalogues, both commercial and open-source are available for use, with the WHO having published their own catalog of MTBC mutations (103). Drug resistance screening via WGS already produces sufficiently accurate results and the public health systems of the Netherlands, England and New York are increasingly favouring it over pDST (104).

1.6.4. *In silico* prediction of drug resistance

The process of assessing the presence of drug resistance-conferring mutations occurs after WGS of the sample under study. Commonly, a WGS procedure culminates in the alignment of contiguous sequences (contigs) in order to map the sample's full genome, usually in the form of one or multiple FASTQ files that can be further interpreted. This procedure entails a number of phases in which errors which may compromise variant calling (the process of identifying indels and small nucleotide polymorphisms - SNPs), such as amplification biases, mapping artifacts and software errors (105). Similarly, quality control of the WGS product includes several steps. First is the mapping of the generated reads to a reference genome, usually performed using the Burrows-Wheeler Alignment Tool (BWA), a publicly available software package (106). In MTB's case, the reference genome used is the full genome of the H37Rv strain, the first of its species to be fully sequenced (107). This procedure generates a SAM file to be further formatted for input in the genome analysis toolkit (GATK), another publicly available toolkit (108). Prior to GATK processing, the generated SAM file should be cleared of duplicate reads, which can be performed through packages such as Picard tools. Indels must also be taken into consideration when performing the proper realignment of WGS data, both tasks possible through GATK. This realignment can be further refined by filtering for base read quality. This process culminates in the creation of a BAM file ready for variant calling.

In order to produce reliable results, technicians must have access to curated libraries of documented mutations to guide their study. In recent years, in order to streamline the bioinformatic process previously described and allow access to this methodology to those with underdeveloped bioinformatics backgrounds, software has been created based on optimal

presets for mapping and variant calling of MTB genome data that compares mutations found with libraries of drug resistance-conferring mutations.

1.7. Tuberculosis in the community of Portuguese language countries

The community of Portuguese language countries (Comunidade de Países de Língua Portuguesa - CPLP) is a commonwealth comprised of nine member-states (Portugal, East Timor, São Tomé and Príncipe, Mozambique, Cape Verde, Brazil, Guinea-Bissau, Equatorial Guinea and Angola), sharing a common history and the Portuguese language as a national spoken language. Additionally, it comprises multiple associate and consultative observer states. The CPLP functions as a regulatory institution that has ratified a number of resolutions facilitating the flow of goods, services and people among member-states, including temporary visas for medical treatment.

Within the scope of the CPLP, the WHO has defined Angola, Mozambique, Brazil and Guinea-Bissau as high-burden countries for TB within the period encompassing 2016 to 2020 (2). Further, with Africa being the continent with the highest prevalence of TB, other African member-states may be at greater risk of outbreaks. East Timor, neighboring Indonesia, another high-burden country for TB, is in a similar situation. It should be noted that Portugal is the only country in western Europe which has not registered a TB incidence rate of less than 10 in 100000 population in 2019 (2). Migration patterns and delays in economic development may be the cause.

As a consequence of limited resources, although recents efforts have allowed the isolation of strains and phylogenetic categorization of MTB in the CPLP (109), the characterization of the genetic background of MTB in the CPLP is still a work in progress. While the CPLP has established cooperation agreements towards managing HIV and malaria, there is still no joint resolution on fighting TB.

This work will set out to study seventeen strains isolated in Angola and genetically sequenced in regards to their resistance to anti-MTB drugs and contextualize their phylogenetic makeup in order to integrate them on both the context of the CPLP as well as globally. Thus contributing to a better understanding of MTB strain migrations and connections across the globe.

2. Materials and methods

2.1. *Mycobacterium tuberculosis* clinical isolates

The studies include 17 MTB clinical isolates obtained from sputum-smear positive patients clinically diagnosed with TB at the Hospital da Divina Providência (HDP) in the Kilamba-Kiaxi municipality of Luanda, Angola. This sample set was selected for WGS from a larger sample (n=89) obtained between March to June 2014 (110). Due to time and logistical constraints, it was unfeasible to sequence all strains obtained by Perdigão et al. pDST data was already available for enabling the selection by prioritizing strains with resistance profiles to first-line anti-TB drugs, while maintaining a wide enough array of isolates so as to be representative of the genetic diversity observed for the original sample which was conveyed by classical typing methods.

Whole genome sequencing was carried out in collaboration with the London School of Hygiene and Tropical Medicine from DNA extracted using the cetyltrimethylammonium bromide (CTAB) method. Sequencing was carried out on an Illumina HiSeq 2500 sequencing platform, according to the protocols and instructions recommended by the manufacturers.

2.2. *In silico* DST

Drug resistance profile prediction was performed on the raw reads obtained from all 17 strains using three different software tools: Mykrobe (111), PhyResSE (112) and TB-Profiler (113). An in-house script combining R and bash programming languages was created to further streamline the process and integrate data from the different tools. The obtained data was then compared for similarities and differences in prediction of drug susceptibility and resistance.

2.3. Quantitative drug susceptibility testing

Phenotypic DST was performed for isolates showing discordant results between previously available DST data and the results obtained *in silico*. pDST was carried out in accordance with biosafety standards for the handling of MTB (biosafety level 3), through MIC determination based on the EUCAST reference method (92). Three-to-four week cultures on

Lowenstein-Jensen solid medium slopes were used to prepare bacterial suspensions adjusted to a 0.5 MacFarland standard. This was carried out by harvesting bacterial growth, followed by homogenization via vortexing in glass tubes with glass beads and resuspension in sterile distilled water. The suspension was adjusted to a MacFarland standard in a new tube containing 10 mL of sterile distilled water and using a Grant Instruments DEN-1B densitometer. A subsequent 1:100 dilution in Middlebrook 7H9-10% OADC medium was performed as to achieve a final inoculum concentration of ~10^5 CFU/mL.

Broth microdilution was carried out in U-shaped 96-well polystyrene plates containing two-fold dilutions of each antimicrobial agent at twice the desired concentration, according to Figure 8. Plates were inoculated with the inoculum prepared as described above at a 1:1 ratio (final volume per well of 200 µL). A negative control (sterility control) column was included in each plate along with a positive (no drug) and 1% diluted inoculum positive control. The plates were incubated under normal atmosphere at a temperature of 36±1 ºC and growth assessed at days 10, 14 and 21 using an inverted mirror and confirmed on an inverted microscope. Final results were recorded upon detection of growth at the 1% positive control. Drug susceptibility is determined by each antimicrobial agent's MIC, defined as the lowest concentration that inhibits visual growth. The CCs herein considered were those defined according to the WHO's Technical manual for drug susceptibility testing of medicines used in the treatment of tuberculosis (90) and Technical Report on critical concentrations for drug susceptibility testing of medicines used in the treatment of drug-resistant tuberculosis (114) as follows, in µg/mL: INH: 0.1; RIF: 1; EMB: 5; STR: 1; AMK: 1; OFX: 2.

DW	DW	DW	DW	DW	DW	DW	DW	DW	DW	DW	DW
NC	PC	1:1	1:2	1:4	1:8	1:16	1:32	1:64	1:128	1:256	DW
NC	PC	1:1	1:2	1:4	1:8	1:16	1:32	1:64	1:128	1:256	DW
NC	PC	1:1	1:2	1:4	1:8	1:16	1:32	1:64	1:128	1:256	DW
NC	1%PC	1:1	1:2	1:4	1:8	1:16	1:32	1:64	1:128	1:256	DW
NC	1%PC	1:1	1:2	1:4	1:8	1:16	1:32	1:64	1:128	1:256	DW
NC	1%PC	1:1	1:2	1:4	1:8	1:16	1:32	1:64	1:128	1:256	DW
DW	DW	DW	DW	DW	DW	DW	DW	DW	DW	DW	DW

Figure 8 - Schematic of the 96-well plates containing each strain to study for phenotypic drug resistance. Each cell represents one well in the plate. **DW** - Distilled Water; **NC** - Negative Control (no mycobacteria present); **PC** - Positive Control (no antibiotic present); **1%PC** - Positive Control at 1% concentration of mycobacteria; **1:1, 1:2, etc** - antibiotic concentrations. Each row is used to assess resistance to one specific antibiotic. Antibiotic agents studied: INH, RIF, EMB, STR, AMK and OFX.

2.4. Variant calling

In order to further assess the genetic basis of pDST findings and elucidate any discrepancies between it and *in silico* DST results, further drug resistance-associated variants were investigated by mapping raw reads to the MTB H37Rv genome (NC000962.3). Quality control, filtering and trimming of raw reads was previously carried out using Trimmomatic and mapped to the reference genome using the BWA-MEM algorithm. Variant calling was carried out on BAM files using both the Genome Analysis Toolkit (HaploType Caller) and SAMtools/BCFtools, with the ensuing VCF files functionally annotated using SnpEff. Artemis was used for visualization of variants along the genome of MTB.

2.5. Phylogenetic analysis

The phylogenetic analysis included in this study was performed across two datasets: i) the Angolan dataset composed by the isolates herein sequenced and analysed from the Hospital da Divina Providência in Luanda, Angola; and ii) a global L4.3 dataset including the newly sequenced L4.3 strains from Angola. The latter was assembled for a precise phylogenetic positioning and contextualization of strains circulating in Angola in a wider macroepidemiological scenario and includes L4.3 isolates selected from the in-house FFUL/iMed.ULisboa TBAtlas database. This database is assembled from MTB genome-wide

publicly available data at the European Nucleotide Archive and includes variant information called from reference assembly of raw reads and *de novo* assembled draft genomes for 85983 isolates whose genomes were made available until August 1st, 2017. Isolate selection was carried out from a subset of the TBAtlas database (n=25794), a downsized dataset with decreased clonality (distance > 0 SNPs) and from which only isolates belonging to sub-lineage 4.3 (L4.3/LAM) and with available metadata on the country and date of origin were selected.

Phylogenetic reconstruction was done based on variant calling and identification of high-quality SNPs, i.e., only concordant variants between both GATK and SAMtools/BCFtools were retained followed by coverage-based validation. SNP positions and isolates showing an excess of 10% missed calls were excluded from the analysis along with SNP positions associated with PE/PPE or resistance-associated genes. The final genome-wide SNP datasets for the Angolan and global L4.3 datasets were composed, respectively, of 17 and 1695 isolates and 987 and 10255 high-quality SNP sites. Retained SNPs were concatenated into a DNA pseudo-molecule for each isolate and used in a tree inference analysis using IQ-TREE software version 2.1.3 (115) using the best-fit nucleotide substitution model as determined by the software's inbuilt Modelfinder module (116), with branch support assessed by the approximate likelihood ratio test with 1000 replicates. For the Angolan dataset, the model chosen was TVM+F in concordance with Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). For the global L4.3 dataset, the model chosen was TVM+F+R2 in concordance with BIC and AIC.

Genomic clustering of strains was initially assessed using SNP distances thresholds of five and 12 SNPs using the dna.dist and cutree functions from the ape and stats packages, respectively, in R and implemented in in-house developed scripts (117). Extended phylogroups involving Angolan L4.3 isolates were analyzed using a cutoff distance of 200 SNPs.

Phylogenetic tree visualization and extraction of specific subtrees was performed using MEGA version 11 (118). Tree annotation was performed using the Interactive Tree Of Life (ITOL) version 6.5.2 (119). For the Angolan dataset, annotation data was derived from pDST results for drug resistance type and specific drug resistance and lineage annotation derived from TB-Profiler data. For global L4.3 dataset, drug resistance type, specific drug resistance and lineage data were all derived from TB-Profiler. Clustering annotation for all datasets were derived from SNP distance data obtained as detailed above.

Spoligotyping of Angola strains used in the global L4.3 dataset, as well as spoligotyping of strains in phylogenetic proximity of clustered Angola strains was performed using the Spotyping tool (120)

2.6. In-house scripting

In order to streamline workflows, a number of in-house scripts were developed using the R programming language and in part supported by the bash programing language. Specifically, a script for automatization of *in-silico* drug susceptibility testing was created, involving the reading of strain sequencing data in FASTQ format by different software, namely Mykrobe and TB-Profiler for drug resistance screening, as well as the SpoTyping tool (120) for spoligotyping of studied strains. The prerequisites for successful usage of this script are a linux environment with the bash UNIX shell and the Comprehensive R Archive Network (CRAN) installed. The script has been made available in github (https://github.com/jssn-BFC/mtbread). Further scripts were developed in R to clean-up and reformat reference strain data as well as streamline the SNP distance clustering process and creation of ITOL annotation files.

3. Results and Discussion

3.1. Genomic characterization of Angola strains

The present study includes a total of 17 isolates obtained from patients diagnosed with tuberculosis in Luanda, Angola. The isolates herein studied are part of a larger dataset and were selected for WGS analysis based on genetic diversity and drug resistance (as detailed in Materials and Methods). Genomic characterization of Angola strains is summarized in Table 3.

Regarding lineage analysis, results were obtained from *in silico* classification by TB-Profiler as well as by mutation analysis of region of differentiation (RD). All strains belong to Lineage 4, displaying consistency with previously obtained results by classical spoligotyping performed by Perdigão et al (110). SNP-based classification allowed for a finer sub-lineage classification: seven strains belong to sub-lineage 4.3.4.1; four to sub-lineage 4.3.4.2; three to sub-lineage 4.1.1; two to sub-lineage 4.3.3; and one isolate was found to belong to sub-lineage 4.8.

Further analysis of RD corroborates SNP-based sub-lineage classification: the strain found to belong to sub-lineage 4.8 was positive for RD247; all 11 strains found to belong to lineage 4.3.4 are positive for RD174, hallmark of this sub-lineage; RD115 was detected on the two strains of sub-lineage 4.3.3; strains belonging to sub-lineage 4.1.1.1 were positive for RD183; and the strain belonging to lineage 4.1.1.3 was positive for RD193. This distribution of RD regions is, as expected, consistent with the sub-lineage delineation proposed by Coll et al (121). Moreover, 11/17 isolates were positive for RD-Rio, which has been associated, along with RD-174, to higher rates of transmissibility (122). Furthermore, RD-Rio has also been associated with both cavitary tuberculosis (123) and multidrug resistance in the past (124), indicating the studied Angola strains may show proclivity for drug resistance acquisition and a disease progression towards cavitary tuberculosis. As expected both RD-Rio and RD174 were found to occur concomitantly with one another (125).

Regarding spoligotype analysis, the isolates studied show a predominance of the LAM family with six of the 17 strains herein identified as belonging to SIT20-LAM1, and four others were identified as SIT42-LAM9. The remaining seven strains are each identified as an individual spoligotype: SIT2025-HAARLEM3:T1, SIT53-T1, SIT64-LAM6, SIT244-T1, Orphan-T3:X3, Orphan-LAM1:LAM9 and SIT244-T1. The predominance of LAM lineages and the

presence of T1, X3 and HAARLEM lineages is to be expected, as these spoligotype families are associated with Lineage 4 (126–128).

Table 3 - Summary of phylogenetic information for Angola strains.

Strain	Lineage (TB-Profiler)	RD mutations	SIT	Clade
AO000110	4.3.4.1	RD149, RD167, RD174, RD3, RDrio	20	LAM1
AO000111	4.3.4.1	RD149, RD167, RD174, RD3, RD6, RDrio	20	LAM1
AO000113	4.3.4.1	RD149, RD167, RD174, RD3, RDrio	20	LAM1
AO000115	4.3.4.2	RD149, RD152, RD174, RD3, RD6, RDrio	42	LAM9
AO000121	4.1.1.1	RD149, RD183, RD252, RD198a, RD3, RD6, RD11	244	T1
AO000123	4.3.4.1	RD149, RD174, RD3, RD6, RDrio	20	LAM1
AO000125	4.8	RD219, RD247, RD247b	2025	HAARLEM3:T1
AO000127	4.1.1.3	RD149, RD193, RD3, RD6	53	T1
AO000131	4.3.4.1	RD149, RD167, RD174, RD3, RDrio	20	LAM1
AO000132	4.3.4.2	RD149, RD152, RD174, RD3, RD6, RDrio	42	LAM9
AO000134	4.3.4.2	RD149, RD152, RD174, RD3, RD6, RDrio	64	LAM6
AO000138	4.3.3	RD115, RD149, RD3, RD6	42	LAM9
AO000142	4.3.4.2	RD149, RD152, RD174, RD3, RD6, RDrio	42	LAM9
AO000149	4.3.4.1	RD149, RD174, RD3, RDrio	20	LAM1
AO000169	4.1.1.1	RD149, RD183, RD252, RD3, RD6	244	T1
AO000176	4.3.3	RD115, RD149, RD145a, RD3, RD6	Orphan	T3:X3
AO000177	4.3.4.1	RD149, RD174, RD3, RDrio	Orphan	LAM1:LAM9

To further characterize the genomic background of the studied strains, a quantitative distribution of genomic mutations per strain was performed by cataloguing genomic variants through variant calling from reference assembly mapping data. The variants detected were classified as indels or SNPs, with SNPs further divided by whether they were found to occur in inter- or intragenic regions. Mutations in intra-genetic regions were then functionally classified as synonymous mutations, non-synonymous or other mutations (complex mutations or nonsense). On average, non-synonymous mutations were found to occur at a ratio of 1.5 against synonymous mutations, which is to be expected considering the redundancy of the genetic code (**Table 4**). Interestingly, the highest NS/S ratio was observed for the isolate belonging to sub-lineage 4.8 (AO000125), also concomitantly with a lesser number of total

mutations and indels. The latter likely denotes a closer proximity to the reference strain (*M. tuberculosis* H37Rv, L4.9) and possibly different fixation rates of both synonymous and non-synonymous substitutions.

Table 4 - Summary of genetic mutation information for Angola strains. NS/S ratio represents the ratio of non-synonymous mutations *versus* synonymous mutations.

			Intragenic SNPs				
Strain	**Indels**	**Intergenic SNPs**	**Non-synonymous SNPs**	**Synonymous SNPs**	**Other SNPs**	**NS/S ratio**	**Total mutations**
AO000110	79	7	472	312	113	1.51	983
AO000111	74	7	463	312	113	1.48	969
AO000113	84	7	459	310	110	1.48	970
AO000115	79	5	456	310	102	1.47	952
AO000121	101	8	506	338	110	1.5	1063
AO000123	76	10	464	320	125	1.45	995
AO000125	67	7	339	192	72	1.77	677
AO000127	105	7	544	347	116	1.57	1119
AO000131	80	8	472	314	108	1.5	982
AO000132	82	6	470	311	104	1.51	973
AO000134	85	6	475	303	103	1.57	972
AO000138	75	8	473	294	113	1.61	963
AO000142	77	6	476	321	103	1.48	983
AO000149	82	7	470	320	106	1.47	985
AO000169	99	9	503	343	113	1.47	1067
AO000176	78	7	480	290	116	1.66	971
AO000177	80	9	494	330	121	1.5	1034

3.2. Drug sensitivity testing

3.2.1. Drug resistance prediction profiles

In order to characterize the studied strains in regarding drug resistance profile, *in silico* drug resistance prediction analysis was performed using PhyResSE, Mykrobe and TB-Profiler.

The results of each software tool were concordant for 15 of 17 isolates. Of the 15 concordant isolates, 13 were determined to be sensitive to all anti-MTB drugs and the remaining two were predicted as resistant to INH. All three software tools attributed INH resistance in both strains to a substitution of serine by threonine in codon 315 of gene *katG* (*katG* S315T).

Of the two non-concordant strains, one strain, AO000110, showed similar drug resistance prediction profiles to RIF, INH and EMB on all three software. However, TB-Profiler predicted an additional resistance to ET. All software attributed the same mutations as the origin of resistance to INH (*katG* S315T), RIF (*rpoB* S450L) and EMB (*embB* M306I). TB-Profiler attributed resistance to ETO to a deletion of nucleotides 735 and 741 in coding DNA of gene *ethA* (*ethA* c.735 741del).

The remaining non-concordant strain, AO000111, showed variability between all three drug resistance prediction tools: Phyres, Mykrobe and TB-Profiler all detect resistance to INH (*katG* S315T) and EMB (*embB* M306V); however, Phyres further detects resistance to RIF (*rpoB* S450L) and STR (*gidB* V88A); Mykrobe is concordant with all resistances predicted by TB-Profiler as well as resistance to PZA, attributed to a nucleotide mutation in gene *pncA* (*pncA* GAG430TAG). Furthermore, Mykrobe STR resistance is attributed to a substitution of threonin by tryptophan at codon 137 of gene *gid* (*gid* T137W).

Comparison of software results with pDST previously performed by Perdigão et al for these strains is concordant for all but four strains: AO000110, which is already discordant between software, displayed resistance to RIF, INH, STR; AO000111 is concordant with Mykrobe's prediction; AO000132, considered sensitive by all software, was found resistant to INH and AO000177, which was found resistant to INH by all software, presented phenotypic resistance to STR. Discrepancies between software have their origin in the mutation libraries that serve as basis for each software's drug resistance prediction accuracy.

3.2.2. Phenotypic drug sensitivity testing

Validation of *in silico* data by pDST was performed for six of the 17 strains. Strain selection was performed based on *in silico* drug resistance profiles, as well as taking in account any strains where inter-software discrepancies were found.

pDST corroborates the data obtained from the *in silico* studies for two of the six strains tested (summarized in **Table 5**). AO000134 showed to be susceptible to all anti-MTB agents tested, while AO000169 displayed MIC values compatible with resistance to INH.

Table 5 - Minimum inhibitory concentrations for all strains subject to pDST. The six strains subject to pDST were selected based on drug resistance patterns and inconsistensies between drug resistance prediction software.

	Anti-TB drugs MIC (µg/mL)					
	INH	RIF	EMB	STR	AMK	OFX
AO000110	4	16	2	8	0.5	0.25
AO000111	4	16	4	32	1	1
AO000134	0.06	0.13	1	0.25	0.5	0.5
AO000132	0.25	0.25	0.5	0.5	0.5	0.5
AO000169	4	0.25	0.5	0.5	1	0.5
AO000177	4	0.13	0.5	8	0.5	0.25

Of the four strains with resistance profiles not completely corroborated by *in* silico testing, AO000177 displayed additional resistance to STR; AO000132, predicted as pansuceptible by all tested software, displayed resistance to INH; for AO000111 the Phyres based prediction was corroborated by phenotypic discernible resistance to RIF and STR; for AO000110, only resistance to RIF and INH was verified, with pDST detecting additional resistance to STR (**Table 6**). All pDST results obtained in this study corroborate the findings of Perdigão et al. Additional tables encompassing complete information concerning resistance-conferring genes as well as other genetic mutations detected by TB-Profiler are available as supplementary material (supplementary tables ST1 and ST2).

Table 6 - Summary of mutations detected in pDST as well as *in silico* testing. All software associated genetic mutations to each resistance detected (data not shown). Besides lineage information, TB-Profiler also displays other mutations found (data not shown).

Strain	**pDST**	**Phyres**	**Mykrobe**	**TB-Profiler**
AO000110	RIF, INH, STR	RIF, INH, EMB	RIF, INH, EMB(r)	RIF, INH, EMB, ETO
AO000111	RIF, INH, EMB, STR	RIF, INH, EMB, STR	RIF, INH, EMB, STR(r), PZA	INH, EMB
AO000113	Sensitive	Sensitive	Sensitive	Sensitive
AO000115	Sensitive	Sensitive	Sensitive	Sensitive
AO000121	Sensitive	Sensitive	Sensitive	Sensitive
AO000123	Sensitive	Sensitive	Sensitive	Sensitive
AO000125	Sensitive	Sensitive	Sensitive	Sensitive

AO000127	Sensitive	Sensitive	Sensitive	Sensitive
AO000131	Sensitive	Sensitive	Sensitive	Sensitive
AO000132	INH	Sensitive	Sensitive	Sensitive
AO000134	Sensitive	Sensitive	Sensitive	Sensitive
AO000138	Sensitive	Sensitive	Sensitive	Sensitive
AO000142	Sensitive	Sensitive	Sensitive	Sensitive
AO000149	Sensitive	Sensitive	Sensitive	Sensitive
AO000169	INH	INH	INH	INH
AO000176	Sensitive	Sensitive	Sensitive	Sensitive
AO000177	INH, STR	INH	INH	INH

3.2.3. Elucidation of discrepancies between *in silico* and phenotypic DST

Discrepancies between resistance predictions tools and pDST were further resolved by manual inspection and variant visualization of genomic variants identified via reference assembly and variant calling.

Concerning AO000110, a substitution of an adenine by a cytosine at nucleotide 907 (n.907A>C) of the *rrs* gene, may explain phenotypic resistance to STR and has been reportedly found in STR-resistant isolates in Mexico and China (69,129). There are also two SNPs found in *gidB*: first, a substitution of leucine by arginine at aminoacid 16 (p.Leu16Arg), which has been found in both resistance and susceptible isolates; the second mutation translates in a substitution of alanine by proline at aminoacid 138 (p.Ala138Pro), which has been found associated with STR-resistant MTB strains (130). The simultaneous presence of *rrs* and *gidB* mutations with STR resistance-conferring character is noteworthy, as studies have shown that *gidB* mutations causing STR resistance occur in strains without mutations in *rrs* (131,132).

Similarly, AO000177 presents two mutations in *gidB*: p.Leu16Arg and a substitution of a guanine by an adenine at nucleotide 419, which has not been associated with phenotypic resistance thus far. The presence of the p.Leu16Arg substitution in both isolates is expected as they both share the LAM clade. Studies have shown that this particular mutation is also a marker for the LAM lineage and therefore unrelated to drug resistance (131), (133).

Strain AO000132 presents resistance to INH not predicted by any software. Variant calling of known markers of genetic resistance as detailed in the literature did not produce conclusive results as to the genetic basis of INH resistance. As INH is a known substrate for

efflux pumps, which have shown *in vitro* to be able to mediate transient high-level INH resistance (134), it is possible the observed resistance is partly related with overexpression of efflux pumps. Variant calling of genes associated with overexpression of efflux pumps in the literature indicates AO000132 to possess mutations in *Rv1410c* (135) (synonymous mutation) and *pstB* (non-synonymous mutation, p.Thr61Met). The gene *whiB7*, which mediates expression of efflux pumps and whose overexpression is related to oxidative stress and intracellular accumulation of fatty acids (136), which are also consequences of INH action, shows no mutations.

Nevertheless, these mutations do not present conclusive evidence as most studies concerning efflux pump expression have been limited to quantitative PCR assays. Furthermore, mutations in *pstB* have also been correlated with resistance to other anti-MTB drugs such as RIF and EMB (137) and, in general, genome-wide association studies studies have shown a relation between efflux pump mutations and patterns of XDR (138). Regardless, the interplay between mutations in direct targets of anti-MTB drugs and efflux pumps is still a factor bearing consideration, as it may entail an unexplored source of target mutations. Mutation screening should also not focus exclusively on each of these genes, as mutations in promoter regions leading to overexpression of these factors may also translate into a cause of drug resistance. Further studies are needed involving the characterization of the strains basal efflux levels as well as the modulating effect of efflux pump inhibitors on the INH resistance level for this strain.

3.3. Phylogenetic analysis

A phylogenetic tree for Angola strains was created and annotated based on resistance type, lineage and resistance to which anti-TB drugs as inferred by pDST (**Figure 9**). The overall tree topology reflects the sub-lineage classification denoting the formation of monophyletic clades for each sub-lineage. No clear drug resistance pattern was associated with any specific tree branch or clade and even though three L4.3.4.1 INH-resistant strains, two of which MDR, are present in this sample. The tree topology enables confirmation that these have emerged through independent resistance acquisition events. The high-resolution power provided by genome-wide phylogenetic analysis therefore enables the differentiation between distinct acquisition events which is especially relevant between strains from the same sub-lineage bearing the same mutations underlying drug resistance.

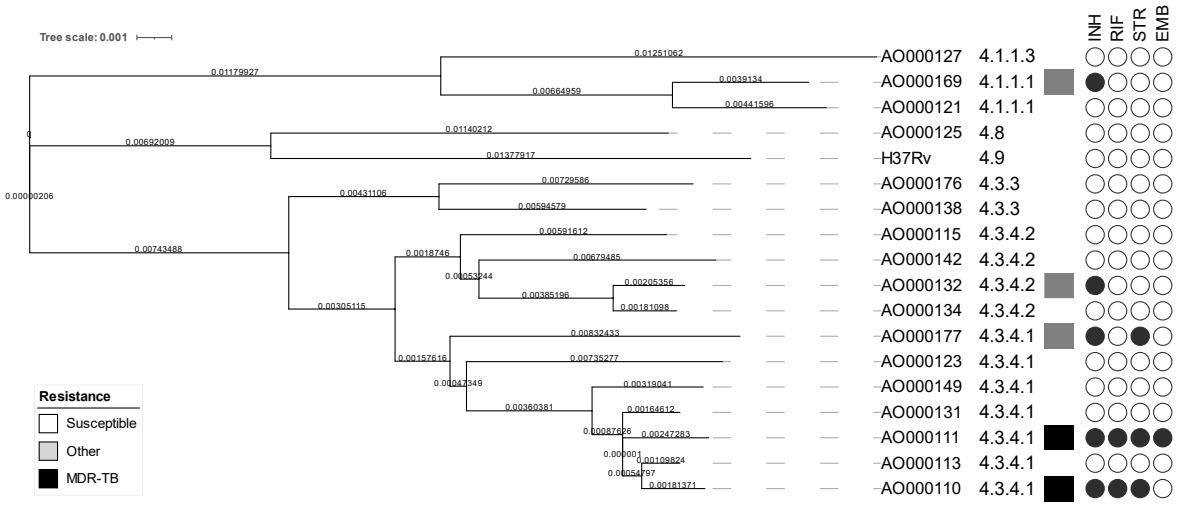


Figure 9 - Phylogenetic tree for Angola strains.

In order to further analyze genetic proximity between strains we sough to characterize the distribution pairwise SNP distances in this sample set along with clustering-based analysis. Congruently with the moderate to deep branching patterns observed in the tree, clustering analysis revealed no clustering of strains below distances of 25 SNPs (**Figure 10**). This absence of genomic clusters below this cut-off level, which is clearly above the cutoffs proposed by Walker et al of 5/12 SNPs, enables us to confirm that these isolates reflect distinct transmission chains involving recent TB transmission events. Nonetheless, the existence of a common and recent ancestor may be inferred, as some strains cluster at distances below 50 SNPs, within the same sub-lineages, and may represent long-standing presence of a given phylogroup in the community. Still regarding the absence of recent transmission events in this dataset, and given that Angola is a high-incidence setting, this study reinforces the need for follow-up studies that include large sample sizes with enough power to capture the high clonal diversity along with its recent transmission networks. Ideally, and given recent reports highlighting the problem of MDR-TB in Angola, MDR-TB-focused molecular epidemiological analysis is likely at this stage the most pressing priority.

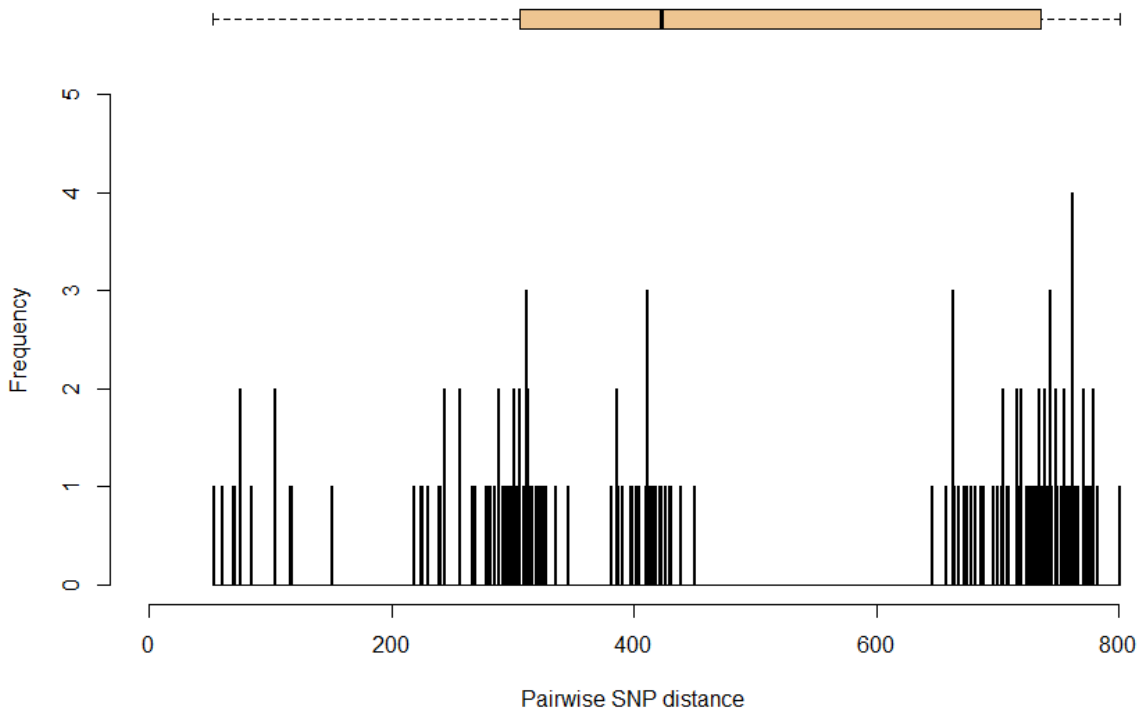


Figure 10 - Distribution of pairwise SNP distances of Angola strains. Lack of pairwise SNP distances below 25 SNPs indicates the strains in this data set do not share direct chains of transmission. Nevertheless, spoligotyping and lineage data indicates the existence of common recent ancestry.

Next, to provide a more accurate snapshot of the phylogenetic positioning of L4.3/LAM strains found in Angola (n=13) in a macroepidemiological context, we performed a phylogenetic tree reconstruction encompassing a global dataset of 1682 strains from thirty countries. This analysis was solely focused on L4.3 strains given its predominance in the Angolan sample set but also due to the epidemiological importance of L4.3/LAM strains across the entire Community of Portuguese Speaking Countries (CPLP). Herein, contrary to the previous distribution of pairwise SNP distances, the spectrum of pairwise SNP distances is not only more heterogeneous, with a median pairwise SNP distance of around 310 SNPs and most frequent SNP distances being found between 200 and 400 SNPs, but also denoting pairwise SNP distances in the recent/epidemiological link range (<5/12 SNPs) (**Figure 11**). In fact, although not at the same frequency as higher SNP distances, the frequency of SNP distances below 5/12 allows to extrapolate the possibility of direct chains of transmission between strains in this dataset.

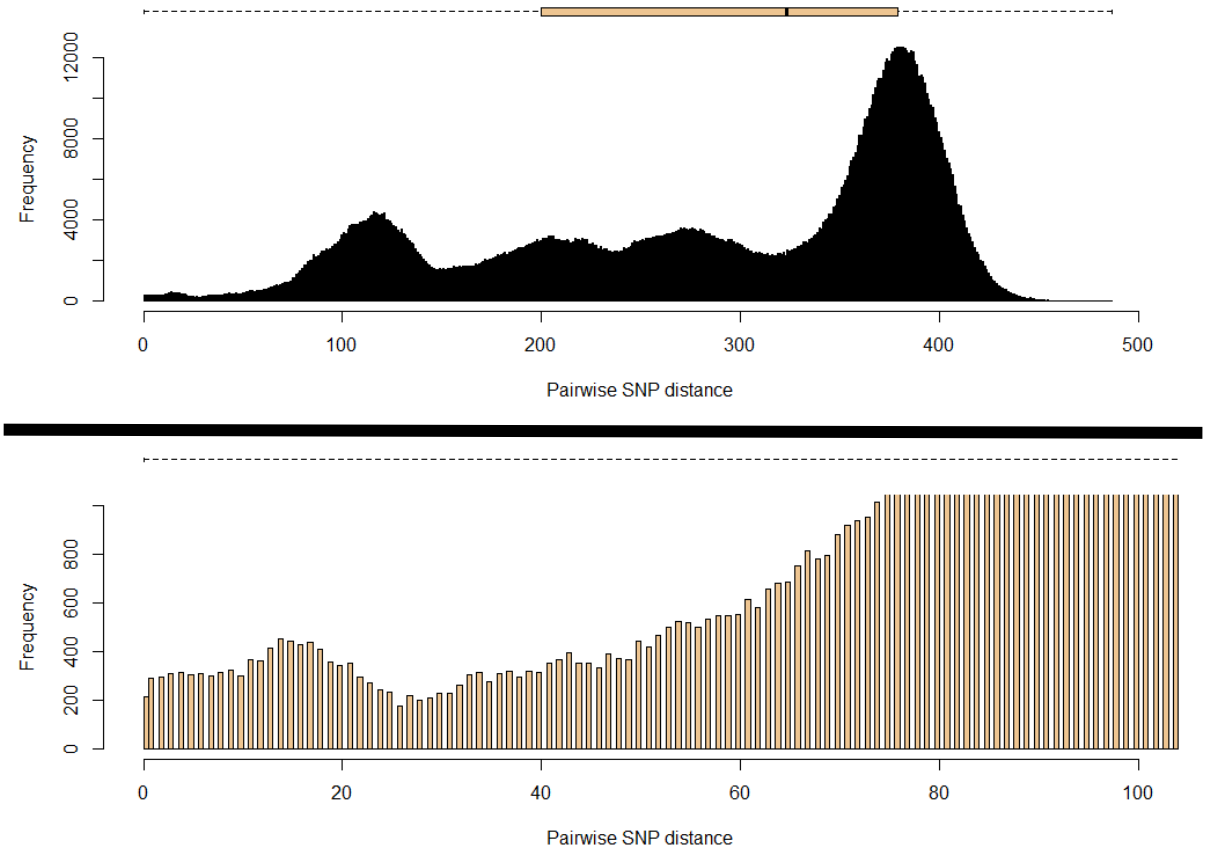


Figure 11 - Complete distribution of pairwise SNP distances in the global data set (above) and SNP distances between 0 and 100 SNPs (below). Although the majority of strains present SNP distances above 300 SNPs, this data set displays a non-insignificant amount of strains at distances below 100 SNPs, which indicates the possibility of clustering between strain within this data set.

Nonetheless, clustering of Angola strains only occurs using a cut-off of 50 SNP within two groups, henceforth denominated Phylogroup 1 (PG1) and Phylogroup 2 (PG2). PG2 groups AO000138 with strains isolated in Brazil, while PG1 groups AO000131, AO000111, AO000113 and AO000110 with strains isolated in Brazil, Canada, Malawi, United Kingdom and Portugal. The known dates of isolation for isolated strains in PG1 encompass an interval between 1996 and 2015.

Angola strains which did not group with any strains in the dataset were studied for phylogenetic common ancestry and phylogenetic trees were created for each, with information regarding country of origin and year of isolation (supplementary data, figures S1-S7). The trend of phylogenetic association with strains isolated in Portugal, Brazil, Malawi, United Kingdom and other English-speaking countries (Canada, Australia, Bangladesh) already found for PG1 continued to be found for the non-PG1/PG2 strains.

Although only thirteen strains were used, the data obtained may suggest that Malawi and Angola may serve as a bridge between MTB strain migration between the CPLP and the anglosphere, in regards to sharing of a common ancestor. However, it should not be ignored that Mozambique, another country of the CPLP, is geographically closer to Malawi than Angola. This phylogenetic commonality may be a combination of a limited sample size of Angola strains coupled with shared phylogenetic backgrounds of Mozambique and Malawi due to migratory flows. Nevertheless, assessment of the overall genetic distribution of all strains obtained by Perdigão et al. (110), from which these 13 strains originate, as well as assessment of the phylogenetic distribution of Malawi strains as detailed by Chihota et al. (139), suggests that Mozambique and Malawi have markedly distinct phylogenetic backgrounds for MTB, as the EAI spoligotype clade is dominant in Mozambique, whereas the LAM clade is dominant in both Malawi and Angola.

Figure 12 includes the complete phylogenetic tree of this data set, indicating phylogroups of interest, as well as PG1 and PG2 subtrees. It should be noted that, due to it's phylogenetic proximity, one strain which did not cluster with PG1, ERR1034863, was kept in the phylogenetic tree of PG1 as a close outgroup.

To investigate the association of the herein identified Angolan associated phylogroups we characterized the diversity and frequency of drug resistance associated mutations in both groups. According to mutation-based resistance prediction software, resistance to isoniazid is the most frequent in both PG1 (30% of strains) and PG2 (50% of strains) More than half the strains on either phylogroup do not present mutational profiles compatible with MDR-TB: over half the strains included in PG1 are susceptible to all antibiotics, whereas exactly half are susceptible in PG2 (**Table 7**).

Resistance to aminoglycosides, rifampicin and isoniazid are generally found limited to similar mutations is *rrs* (n.1401A>G), *rpoB* (p.Ser450Leu) and *katG* (p.Ser315Thr), respectively, with both INH-resistant strains in PG2 sharing this same mutation with 12 INH-resistant mutants of PG1. Resistance to pyrazinamide and fluoroquinolones, while limited to *pncA* and *gyrA/gyrB* respectively, present the biggest variations in mutation site: only two of six strains share the same genetic mutation for PZA resistance (*pncA* p.Pro69Leu), while none of the four strains share their mutation for FQ resistance with other mutants.

Analysis of distribution patterns of drug resistance of PG1 allow the differentiation of nine isolated groups wherein drug resistance mutations are generated independently: ERR275198; AO000111; ERR2653218 and ERR2653186; SRR5007183; ERR275214 and ERR1034907; ERR1034863; ERR1034815; ERR1035257 and AO000110; ERR2653193.

Noteworthy, the group encompassing ERR2653218 and ERR2653186 as well as the group containing ERR1035257 and AO000110, appear to reflect a pattern of accumulation of drug resistance-conferring mutations.

Apart from the strain of unknown spoligotype family, all strains in both phylogroups belong to spoligotype families as expected to belong to lineage 4. All four strains of PG2 belong to LAM9, three of which, including an Angola strain, belong to SIT42. The remaining strain belongs to SIT-388, suggesting a possible descent from SIT42 strains by loss of the fourth spacer region. As for PG1, 34 strains, including all Angola strains in the phylogroup, belong to SIT20-LAM1 and represent the biggest fraction containing drug-resistant strains. The pattern of diversity of drug resistance-conferring genetic mutations for SIT20-LAM1 verified in PG1 is one that has been noted in previous studies (140–143), as well as in Portugal, specifically the endemic cluster Lisboa3 belonging to this same lineage and representing one of the main drivers of drug resistance in the region (144). However, unlike the studied Angola strains within PG1 and PG2, the Lisboa3-SIT20-LAM1 strains belong to the 4.3.4.2 sub-lineage, sharing ancestry with a SIT42-LAM9-RDRio profile (144). Interestingly, four Angola strains (AO000115, AO000132 AO000134 and AO000142) share this same lineage. Analysis of their phylogenetic trees within the global context (**supplementary figures S1**, **S3** and **S4**) show proximity of three strains to phylogenetically adjacent Portuguese strains near the same mean SNP distance (62.9) to the Lisboa3 clade as the SIT42-LAM9-RDrio in the original study: AO000132 is at SNP distances of 66 and 87, AO000134 is at distances of 65 and 84 SNPs. In particular, AO000142 is at an SNP distance of 63 of the Portuguese strain it shares phylogenetic proximity with, showing the closest phylogenetic proximity of all Angola strains to the Lisboa3 strains included in this data set (data not shown). A further noteworthy aspect of these Angola strains is that all of them display pansusceptibility to anti-MTB drugs as per TB-Profiler data, which is another point of commonality with the SIT42-LAM9-RDrio strains of the referenced study (144). Further studies into the genetic make-up of tuberculosis in Angola would help elucidate any evolutionary links between MTB strains in Angola and Portugal.

The relative uniformity of spoligotyping data is within predicted outcomes, as all strains belonging to the worldwide dataset belong to sub-lineage 4.3. Furthermore, the positioning of all Angola strains within each phylogroup attests to the robustness of the SNP clustering performed as it confirms the results of spoligotyping and lineage data, and further contributes to the hypothesis of these strains sharing a common recent ancestor.

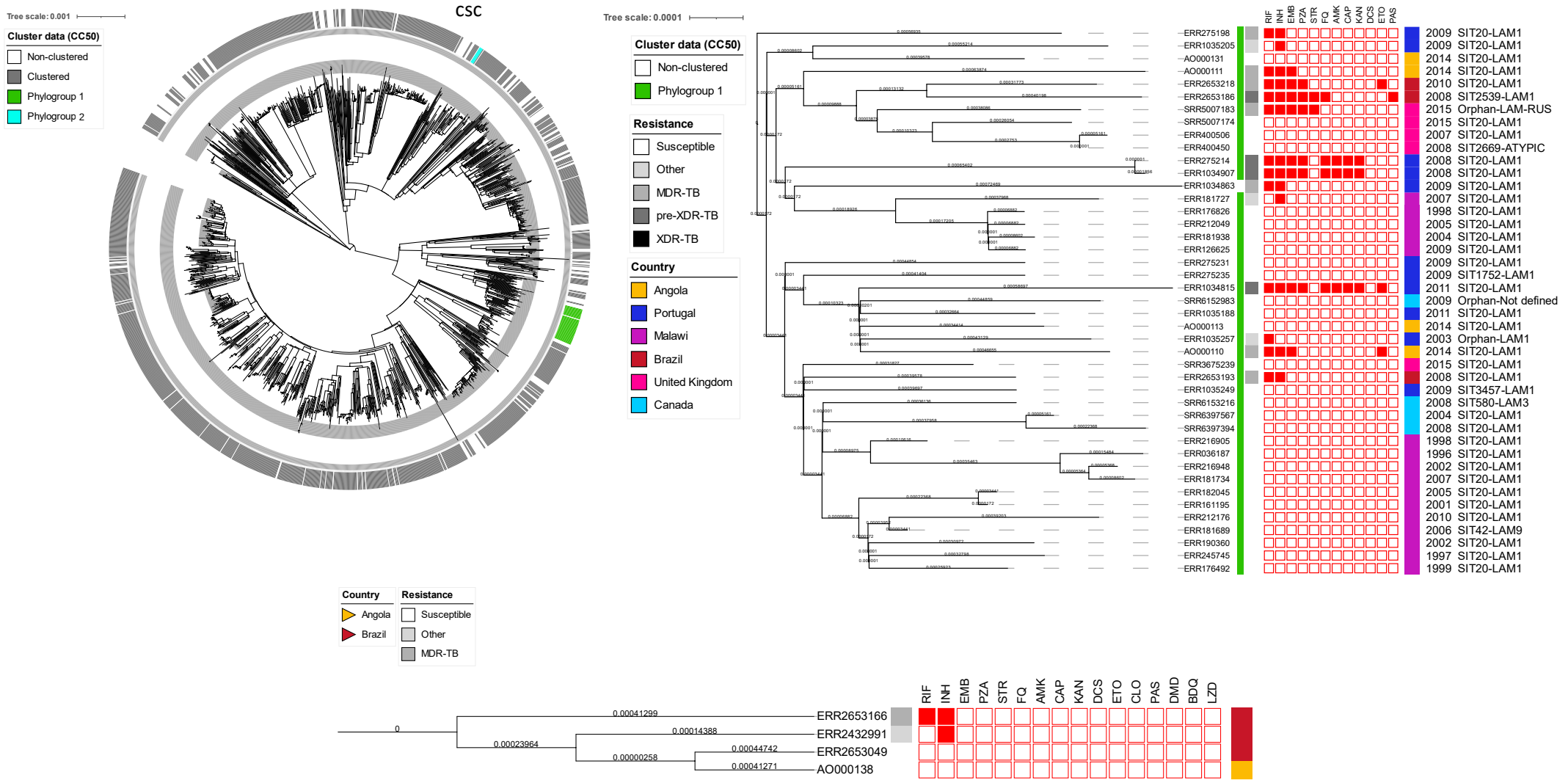


Figure 12 - Phylogenetic trees of global context; complete (upper left), Phylogtoup 1 (upper right), and Phylogroup 2 (bottom). Phylogroup 1 contains information of year of isolation as well as spoligotyping on the second to last and last columns, respectively.

Table 7 - Summary of phylonegetic and drug resistance information for PG1 and PG2.
* - indicates inclusion of a phylogenetically close strain that did not cluster with the phylogroup.

						Drug resistance type					
PG		**Total strains**	**Studied strains**	**Geographic distribution**	**Spoligotype**	**Susceptible**	**Other**	**MDR-TB**	**pre-XDR-TB**	**Drug resistance**	**Mutated genes**
1		43*	AO000110 AO000111 AO000113 AO000131	Angola (n=4) Brazil (n=3) Canada (n=4) Malawi (n=16), Portugal (n=11)* United Kingdom (n=5)	LAM1;LAM2 (n=38) LAM1 (n=3) Unknown (n=1)	29	3	7*	4	RIF (n=12)* INH (n=13) EMB (n=8) PZA (n=6) FQ (n=4) AMK (n=3) CAP (n=3) KAN (n=3) ETO (n=3) PAS (n=1)	RIF - *rpoB* p.Ser450Leu (n=8) - *rpoB* p.His445Tyr (n=1) - *rpoB* p.His445Arg (n=1)* - *rpoB* p.Ala286Val (n=1) - *rpoB* p.His445Leu (n=1) INH - *katG* p.Ser315Thr (n=13)* - *fabG1* c.-15C>T EMB - *embB* p.Met306Ile (n=4) - *embB* p.Met306Val (n=3) - *embB* p.Gly406Ala (n=1) PZA - *pncA* p.Pro69Leu (n=2) - *pncA* p.Pro62Leu (n=1) - *pncA* p.Tyr103* (n=1) - *pncA* p.Met175Ile (n=1) - *pncA* p.Thr135Pro (n=1) FQ - *gyrA* p.His70Arg (n=1) - *gyrA* p.Asp94Tyr (n=1) - *gyrB* p.Ala504Val (n=1) - *gyrB* p.Asp461His (n=1) AMK - *rrs* n.1401A>G (n=3) CAP - *rrs* n.1401A>G (n=3) KAN - *rrs* n.1401A>G (n=3) ETO - *ethA* c.490_491dupCC (n=1) - *ethA* c.735 741del (n=1) - *fabG1* c.-15C>T (n=1) PAS - *folC* p.Ser150Gly (n=1)

2		4	AO000138	Angola(n=1) Brazil(n=3)	LAM;T (n=3) LAM9 (n=1)	2	1	1	0	RIF(n=1) INH(n=2)	RIF - *rpoB* p.Ser450Leu (n=1) INH - *katG* p.Ser315Thr (n=2)

4. Conclusions and future perspectives

The spreading patterns of *M. tuberculosis* strains are a reflection of an increasingly interconnected world. While in the past the spreading of disease through commercial routes was a reality, with the spread of *Yersinia pestis* along the silk road as the most illustrative example (145), travel speeds and migration networks were not as advanced as today's, translating in the possibility of quick expansion of new diseases and variants across the globe, as the recent SARS-COV2 pandemic has shown (146). *M. tuberculosis* is no different. Although there is a distinction between low-burden and high-burden TB countries, strains originary of either can be found in the other country. The work developed in this study demonstrates that notion.

Phylogenetic analysis of the strains isolated in Angola does not suggest a strict relation with strains isolated in Portugal nor the rest of the world, although the existence of recent common ancestry may be inferred. Furthermore, the existence of Angolan SIT42-LAM9-RDrio strains within SNP distances of Portuguese strains suggests a possible evolutionary link between Angolan tuberculosis strains and Portuguese strains. Strains clustering between SNP distances above 25 and below 50 have dates of isolation too recent to have their distances explained solely by time, as MTB's rate of change is estimated at 0.5 SNP per genome per year, defining direct chains of transmission at a maximum of 5 SNP in a 3 year window (117). It should be noted that the boundaries defining direct chains of transmission through SNP clustering are a topic of contention. While Clark et al have defined SNP distances of 50 and below as a cluster (147), other studies have defined a distance between 5 and 12 SNPs as of greater epidemiological relevance (148,149). A study by Cancino-Muñoz et al, currently under review, claims SNP clustering as situational, depending on factors such as previous outbreaks and disease burden. Similarly, Stimson et al have developed a clustering framework that uses similar factors to improve accuracy, but has prerequisites such as prior knowledge of transmission dynamics (150). Nevertheless, current consensus defines the boundary of transmission at 5-12 SNPs.

Concerning this work, analysis of common ancestry indicated a relationship between these strains and strains found in Portugal, Brazil, Malawi, Australia, Canada and the United Kingdom. Strains from the global dataset which did not cluster showed similar patterns of association in regards to country of origin and appear to mostly cluster between 50 and 100 SNPs (suppleentary data).

With Malawi and Angola as the two African countries in this array, and considering their respective history and cultural positioning within the anglosphere and the community of portuguese-speaking countries respectively, it is possible to infer that their geographic proximity has led to a past dissemination of *M. tuberculosis* strains between both countries, which then spread along socio-culturally adjacent countries. This Angola-Malawi axis could be inferred as a bridge of connection between the CPLP and the anglosphere in regards to *M. tuberculosis* spread. Although this study shows a distinction in phylogenetic backgrounds between Malawi and Mozambique, a CPLP country that, unlike Angola, borders Malawi, it should be noted that Malawi has had a Portuguese presence in the recent past. Hence, there is also the possibility that these strains have their origin in Europe, spread to Africa through Portuguese colonization and developed into local strains over time. Nevertheless, the hypothesis that the Malawi strains herein studied may have first migrated from Mozambique into Malawi must not be discarded. Further studies to characterize the phylogenetic make-up of the region would allow elucidation of the chains of transmission of *M. tuberculosis* in the region. Greater sample sizes would allow for higher accuracy and better contextualization of the *M. tuberculosis* spread at a more macroepidemiological context.

Drug resistance screening has shown a reduced, but not insignificant number of inconsistencies between *in silico* predictive models and phenotypic, *in vitro* assays. Within a sample size of 17 strains, 13 strains showed consistent results between phenotypic drug resistance assays and genotypic drug prediction models.

In order to increase the accuracy of *in silico* drug resistance screening, resistance mutation libraries must be subject to constant updating and validation. It may also be possible that mutations with no apparent effect on mechanisms directly or indirectly involved with drug sensitivity or resistance may produce synergistic effects, leading to unexpected resistance to antibacterial drugs. Of particular interest in this study is the phenotypic resistance to INH by strain AO000132. This pattern of resistance may be explainable by overexpression of efflux pumps, decreasing the intracellular concentration of INH, thus compromising its anti-MTB effect. This study shows that, while *in silico* drug screening assays may be a powerful, time-saving tool, they are still not on the level of the gold standard in effectively predicting drug resistance.

The updating of drug resistance-conferring mutation libraries, the basis of drug resistance prediction software tools such as the ones used in this work, may be compromised by the genome sequencing analysis method most commonly used, of read mapping. Not only does read mapping have its number of implicit limitations concerning read lengths and quality control, the reference genome for mapping of new *M. tuberculosis* strains, H37Rv, presents its

own set of issues: it is a laboratory-derived strain (151), shows disease presentation inconsistencies compared to other clinically-obtained isolates (152), reveals inter-laboratory variability of genomic data (153), which may be caused by prolonged culture (154) and absence of genes commonly found in clinical isolates (155). These factors may introduce errors in mapped genomes and, taken together, may translate into H37Rv having a "hard ceiling" in regards to both mapping-based genome sequencing strategies as well as concerning the advancement and curation of drug resistance-conferring genetic libraries. As bioinformatics research involving *M. tuberculosis* matures, so may the state of the art reach a cross-roads which may dictate the course of the subject matter for years to come.

5. References

1. Smith I. Mycobacterium tuberculosis pathogenesis and molecular determinants of virulence. Vol. 16, Clinical Microbiology Reviews. 2003. p. 463–96.

2. Organization WH. Global Tuberculosis Report 2020 [Internet]. World Health Organization 2020 p. 232. Available from: http://publications.lib.chalmers.se/records/fulltext/245180/245180.pdf%0Ahttps://hdl.handle.net/20.500.12380/245180%0Ahttp://dx.doi.org/10.1016/j.jsames.2011.03.003%0Ahttps://doi.org/10.1016/j.gr.2017.08.001%0Ahttp://dx.doi.org/10.1016/j.precamres.2014.12

3. Pai M, Behr MA, Dowdy D, Dheda K, Divangahi M, Boehme CC, et al. Tuberculosis. Vol. 2, Nature Reviews Disease Primers. 2016.

4. Churchyard G, Kim P, Shah NS, Rustomjee R, Gandhi N, Mathema B, et al. What We Know about Tuberculosis Transmission: An Overview. Journal of Infectious Diseases. 2017.

5. Diego VS, Gabriela RR, Vanessa ST. Extrapulmonary tuberculosis. Vol. 5, Revista Bionatura. 2020. p. 1066–71.

6. Riojas MA, McGough KJ, Rider-Riojas CJ, Rastogi N, Hazbón MH. Phylogenomic analysis of the species of the mycobacterium tuberculosis complex demonstrates that mycobacterium africanum, mycobacterium bovis, mycobacterium caprae, mycobacterium microti and mycobacterium pinnipedii are later heterotypic synonyms of mycob. Int J Syst Evol Microbiol. 2018;68(1):324–32.

7. Senghore M, Diarra B, Gehre F, Otu J, Worwui A, Muhammad AK, et al. Evolution of Mycobacterium tuberculosis complex lineages and their role in an emerging threat of multidrug resistant tuberculosis in Bamako, Mali. Sci Rep. 2020;10(1).

8. Gagneux S. Ecology and evolution of Mycobacterium tuberculosis. Vol. 16, Nature Reviews Microbiology. 2018. p. 202–13.

9. Gerstmans H, Rodríguez-Rubio L, Lavigne R, Briers Y. From endolysins to Artilysin®s: Novel enzyme-based approaches to kill drug-resistant bacteria. Biochem Soc Trans. 2016;44:123–8.

10. Adigun R, Singh R. Tuberculosis [Internet]. StatPearls. 2020. Available from: http://www.ncbi.nlm.nih.gov/pubmed/28722945

11. Chiner-Oms, Sánchez-Busó L, Corander J, Gagneux S, Harris SR, Young D, et al. Genomic determinants of speciation and spread of the Mycobacterium tuberculosis complex. Sci Adv. 2019;

12. Sreevatsan S, Pan X, Stockbauer KE, Connell ND, Kreiswirth BN, Whittam TS, et al. Restricted structural gene polymorphism in the Mycobacterium tuberculosis complex indicates evolutionarily recent global dissemination. Proc Natl Acad Sci U S A. 1997;

13. Coscolla M, Gagneux S, Menardo F, Loiseau C, Ruiz-Rodriguez P, Borrell S, et al. Phylogenomics of mycobacterium africanum reveals a new lineage and a complex evolutionary history. Microb Genomics. 2021;

14. O'Neill MB, Shockey A, Zarley A, Aylward W, Eldholm V, Kitchen A, et al. Lineage specific histories of Mycobacterium tuberculosis dispersal in Africa and Eurasia. Mol Ecol. 2019;

15. Gehre F, Otu J, DeRiemer K, de Sessions PF, Hibberd ML, Mulders W, et al. Deciphering the Growth Behaviour of Mycobacterium africanum. PLoS Negl Trop Dis. 2013;

16. Isea-Peña MC, Brezmes-Valdivieso MF, González-Velasco MC, Lezcano-Carrera MA, López-Urrutia-Lorente L, Martín-Casabona N, et al. Mycobacterium africanum, an emerging disease in high-income countries? Int J Tuberc Lung Dis. 2012;

17. Blouin Y, Hauck Y, Soler C, Fabre M, Vong R, Dehan C, et al. Significance of the Identification in the Horn of Africa of an Exceptionally Deep Branching Mycobacterium tuberculosis Clade. PLoS One. 2012;

18. Semuto Ngabonziza JC, Loiseau C, Marceau M, Jouet A, Menardo F, Tzfadia O, et al. A sister lineage of the Mycobacterium tuberculosis complex discovered in the African Great Lakes region. bioRxiv. 2020;

19. Collins TFB. The history of southern Africa's first tuberculosis epidemic. South African Med J. 1982;

20. Comas I, Hailu E, Kiros T, Bekele S, Mekonnen W, Gumi B, et al. Population Genomics of Mycobacterium tuberculosis in Ethiopia Contradicts the Virgin Soil Hypothesis for Human Tuberculosis in Sub-Saharan Africa. Curr Biol. 2015;

21. Taye H, Alemu K, Mihret A, Ayalew S, Hailu E, Wood JLN, et al. Epidemiology of Mycobacterium tuberculosis lineages and strain clustering within urban and peri-urban settings in Ethiopia. PLoS One. 2021;

22. de Jong BC, Antonio M, Gagneux S. Mycobacterium africanum-review of an important cause of human tuberculosis in West Africa. PLoS Neglected Tropical Diseases. 2010.

23. Yimer SA, Norheim G, Namouchi A, Zegeye ED, Kinander W, Tønjum T, et al. Mycobacterium tuberculosis lineage 7 strains are associated with prolonged patient delay in seeking treatment for pulmonary tuberculosis in Amhara region, Ethiopia. J Clin Microbiol. 2015;

24. Negrete-Paz AM, Vázquez-Marrufo G, Vázquez-Garcidueñas MS. Whole-genome comparative analysis at the lineage/sublineage level discloses relationships between Mycobacterium tuberculosis genotype and clinical phenotype. PeerJ. 2021;

25. Varghese B, Enani M, Alrajhi A, Al Johani S, Albarak A, Althawadi S, et al. Impact of Mycobacterium tuberculosis complex lineages as a determinant of disease phenotypes from an immigrant rich moderate tuberculosis burden country. Respir Res. 2018;

26. Grosset J. Mycobacterium tuberculosis in the extracellular compartment: An underestimated adversary. Antimicrobial Agents and Chemotherapy. 2003.

27. Armstrong JA, Hart D. Response of cultured macrophages to Mycobacterium Tuberculosis, with observations on fusion of lysosomes with phagosomes. J Exp Med. 1971;

28. Jayachandran R, Sundaramurthy V, Combaluzier B, Mueller P, Korf H, Huygen K, et al. Survival of Mycobacteria in Macrophages Is Mediated by Coronin 1-Dependent Activation of Calcineurin. Cell. 2007;

29. Repasy T, Lee J, Marino S, Martinez N, Kirschner DE, Hendricks G, et al. Intracellular Bacillary Burden Reflects a Burst Size for Mycobacterium tuberculosis In Vivo. PLoS Pathog. 2013;

30. Bermudez LE, Sangari FJ, Kolonoski P, Petrofsky M, Goodman J. The efficiency of the translocation of Mycobacterium tuberculosis across a bilayer of epithelial and endothelial cells as a model of the alveolar wall is a consequence of transport within mononuclear phagocytes and invasion of alveolar epithelial cells. Infect Immun. 2002;

31. Orme IM, Robinson RT, Cooper AM. The balance between protective and pathogenic immune responses in the TB-infected lung. Nature Immunology. 2015.

32. World Health Organization (WHO). Latent tuberculosis infection : updated and consolidated guidelines for programmatic management. Vol. 25, World Health Organization. 2018. 1–78 p.

33. Organization WH. WHO | Guidelines for treatment of tuberculosis. Who [Internet]. 2010;147. Available from: http://apps.who.int/iris/bitstream/10665/44165/1/9789241547833_eng.pdf?ua=1&ua=1

34. Meeting report of the WHO expert consultation on the definition of extensively drug-resistant tuberculosis, 27-29 October 2020. Geneva: World Health Organization; 2021. CC BY-NC-SA 3.0 IGO. 2021.

35. World Health Organization (WHO). WHO consolidated guidelines on tuberculosis. Module 4: Treatment -Drug resistant tuberculosis treatment. Who. 2020.

36. Timmins GS, Master S, Rusnak F, Deretic V. Nitric oxide generated from isoniazid activation by KatG: Source of nitric oxide and activity against Mycobacterium tuberculosis. Antimicrob Agents Chemother. 2004;

37. Takayama K, Wang L, David HL. Effect of isoniazid on the in vivo mycolic acid synthesis, cell growth, and viability of Mycobacterium tuberculosis. Antimicrob Agents Chemother. 1972;

38. Vilchèze C, Wang F, Arai M, Hazbón MH, Colangeli R, Kremer L, et al. Transfer of a point mutation in Mycobacterium tuberculosis inhA resolves the target of isoniazid. Nat Med. 2006;

39. Larsen MH, Vilchèze C, Kremer L, Besra GS, Parsons L, Salfinger M, et al. Overexpression of inhA, but not kasA, confers resistance to isoniazid and ethionamide in mycobacterium smegmatis, M. bovis BCG and M. tuberculosis. Mol Microbiol. 2002;

40. Ito K, Yamamoto K, Kawanishi S. Manganese-Mediated Oxidative Damage of Cellular and Isolated DNA by Isoniazid and Related Hydrazines: Non-Fenton-Type Hydroxyl Radical Formation. Biochemistry. 1992;

41. TSUKAMURA M, TSUKAMURA S. ISOTOPIC STUDIES ON THE EFFECT OF ISONIAZID ON PROTEIN SYNTHESIS OF MYCOBACTERIA. Jpn J Tuberc Chest Dis. 1963;

42. Murthy PS, Sirsi M, Ramakrishnan T. Effect of isoniazid on the carbohydrate metabolism of isoniazid-susceptible and isoniazid-resistant Mycobacterium tuberculosis H37Rv. Am Rev Respir Dis. 1973;

43. MIDDLEBROOK G. Sterilization of tubercle bacilli by isonicotinic acid hydrazide and the incidence of variants resistant to the drug in vitro. Am Rev Tuberc. 1952;

44. Zhang Y, Garbe T, Young D. Transformation with katG restores isoniazid-sensitivity in Mycobacterium tuberculosis isolates resistant to a range of drug concentrations. Mol Microbiol. 1993;

45. Ramaswamy S, Musser JM. Molecular genetic basis of antimicrobial agent resistance in Mycobacterium tuberculosis: 1998 update. Tuber Lung Dis. 1998;

46. Banerjee A, Dubnau E, Quemard A, Balasubramanian V, Um KS, Wilson T, et al. inhA, a gene encoding a target for isoniazid and ethionamide in Mycobacterium tuberculosis. Science (80-). 1994;

47. Tuberculosis M, Mache A, Mengistu Y, Hoffner SE, Anderson DA. Analysis of Compensatory ahpC Promoter Region Mutations and Relative Fitness of Inh-Resistant. Int J Infect Dis. 2008;

48. Seifert M, Catanzaro D, Catanzaro A, Rodwell TC. Genetic mutations associated with isoniazid resistance in Mycobacterium tuberculosis: A systematic review. PLoS One. 2015;

49. Vilchèze C, Weisbrod TR, Chen B, Kremer L, Hazbón MH, Wang F, et al. Altered NADH/NAD+ ratio mediates coresistance to isoniazid and ethionamide in mycobacteria. Antimicrob Agents Chemother. 2005;

50. Unissa AN, Subbian S, Hanna LE, Selvakumar N. Overview on mechanisms of isoniazid action and resistance in Mycobacterium tuberculosis. Infection, Genetics and Evolution. 2016.

51. Piddock LJV, Williams KJ, Ricci V. Accumulation of rifampicin by Mycobacterium aurum, Mycobacterium smegmatis and Mycobacterium tuberculosis. J Antimicrob Chemother. 2000;

52. Van Deun A, Aung KJM, Hossain MA, De Rijk P, Gumusboga M, Rigouts L, et al. Disputed rpoB mutations can frequently cause important rifampicin resistance among new tuberculosis patients. Int J Tuberc Lung Dis. 2015;

53. Van Deun A, Aung KJM, Bola V, Lebeke R, Hossain MA, De Rijk WB, et al. Rifampin drug resistance tests for tuberculosis: Challenging the gold standard. J Clin Microbiol. 2013;

54. Rigouts L, Gumusboga M, De Rijk WB, Nduwamahoro E, Uwizeye C, De Jong B, et al. Rifampin resistance missed in automated liquid culture system for Mycobacterium tuberculosis isolates with specific rpoB mutations. J Clin Microbiol. 2013;

55. Malenfant JH, Brewer TF. Rifampicin Mono-Resistant Tuberculosis - A Review of an Uncommon but Growing Challenge for Global Tuberculosis Control. Open Forum Infectious Diseases. 2021.

56. Zhang Y, Shi W, Zhang W, Mitchison D. Mechanisms of Pyrazinamide Action and Resistance. Microbiol Spectr. 2014;

57. Shi W, Zhang X, Jiang X, Ruan H, Barry CE, Wang H, et al. Pyrazinamide inhibits trans-translation in Mycobacterium tuberculosis: a potential mechanism for shortening the duration of tuberculosis chemotherapy. Science. 2011;

58. Gopal P, Sarathy JP, Yee M, Ragunathan P, Shin J, Bhushan S, et al. Pyrazinamide triggers degradation of its target aspartate decarboxylase. Nat Commun. 2020;

59. Maeda N, Nigou J, Herrmann JL, Jackson M, Amara A, Lagrange PH, et al. The cell surface receptor DC-SIGN discriminates between Mycobacterium species through selective recognition of the mannose caps on lipoarabinomannan. J Biol Chem. 2003;

60. Escuyer VE, Lety MA, Torrelles JB, Khoo KH, Tang JB, Rithner CD, et al. The Role of the embA and embB Gene Products in the Biosynthesis of the Terminal Hexaarabinofuranosyl Motif of Mycobacterium smegmatis Arabinogalactan. J Biol Chem. 2001;

61. Zhang N, Torrelles JB, McNeil MR, Escuyer VE, Khoo KH, Brennan PJ, et al. The Emb proteins of mycobacteria direct arabinosylation of lipoarabinomannan and arabinogalactan via an N-terminal recognition region and a C-terminal synthetic region. Mol Microbiol. 2003;

62. Zhang L, Zhao Y, Gao Y, Wu L, Gao R, Zhang Q, et al. Structures of cell wall arabinosyltransferases with the anti-tuberculosis drug ethambutol. Science (80-). 2020;

63. Telenti A, Philipp WJ, Sreevatsan S, Bernasconi C, Stockbauer KE, Wieles B, et al. The emb operon, a gene cluster of Mycobacterium tuberculosis involved in resistance to ethambutol. Nat Med. 1997;

64. He L, Wang X, Cui P, Jin J, Chen J, Zhang W, et al. UbiA (Rv3806c) encoding DPPR synthase involved in cell wall synthesis is associated with ethambutol resistance in Mycobacterium tuberculosis. Tuberculosis. 2015;

65. Sharma K, Gupta M, Krupa A, Srinivasan N, Singh Y. EmbR, a regulatory protein with ATPase activity, is a substrate of multiple serine/threonine kinases and phosphatase in Mycobacterium tuberculosis. FEBS J. 2006;

66. Xiang X, Gong Z, Deng W, Sun Q, Xie J. Mycobacterial ethambutol responsive genes and implications in antibiotics resistance. Journal of Drug Targeting. 2021.

67. Krause KM, Serio AW, Kane TR, Connolly LE. Aminoglycosides: An overview. Cold Spring Harb Perspect Med. 2016;

68. Stanley RE, Blaha G, Grodzicki RL, Strickler MD, Steitz TA. The structures of the anti-tuberculosis antibiotics viomycin and capreomycin bound to the 70S ribosome. Nat Struct Mol Biol. 2010;

69. Cuevas-Córdoba B, Cuellar-Sánchez A, Pasissi-Crivelli A, Santana-álvarez CA, Hernández-Illezcas J, Zenteno-Cuevas R. Rrs and rpsL mutations in streptomycin-resistant isolates of Mycobacterium tuberculosis from Mexico. J Microbiol Immunol Infect. 2013;

70. Maus CE, Plikaytis BB, Shinnick TM. Mutation of tlyA confers capreomycin resistance in Mycobacterium tuberculosis. Antimicrob Agents Chemother. 2005;

71. Zaunbrecher MA, Sikes RD, Metchock B, Shinnick TM, Posey JE. Overexpression of the chromosomally encoded aminoglycoside acetyltransferase eis confers kanamycin resistance in Mycobacterium tuberculosis. Proc Natl Acad Sci U S A. 2009;

72. Hooper DC. Bacterial topoisomerases, anti-topoisomerases, and anti-topoisomerase resistance. In: Clinical Infectious Diseases. 1998.

73. Avalos E, Catanzaro D, Catanzaro A, Ganiats T, Brodine S, Alcaraz J, et al. Frequency and geographic distribution of gyrA and gyrB mutations associated with fluoroquinolone resistance in clinical Mycobacterium tuberculosis isolates: A systematic review. PLoS One. 2015;

74. Mayer C, Takiff H. The Molecular Genetics of Fluoroquinolone Resistance in Mycobacterium tuberculosis . Microbiol Spectr. 2014;

75. Wang F, Langley R, Gulten G, Dover LG, Besra GS, Jacobs WR, et al. Mechanism of thioamide drug action against tuberculosis and leprosy. J Exp Med. 2007;

76. Morlock GP, Metchock B, Sikes D, Crawford JT, Cooksey RC. ethA, inhA, and katG Loci of Ethionamide-Resistant Clinical Mycobacterium tuberculosis Isolates. Antimicrob Agents Chemother. 2003;

77. Baulard AR, Betts JC, Engohang-Ndong J, Quan S, McAdam RA, Brennan PJ, et al. Activation of the pro-drug ethionamide is regulated in mycobacteria. J Biol Chem. 2000;

78. Vilchèze C, Av-Gay Y, Barnes SW, Larsen MH, Walker JR, Glynne RJ, et al. Coresistance to isoniazid and ethionamide maps to mycothiol biosynthetic genes in Mycobacterium bovis. Antimicrob Agents Chemother. 2011;

79. Wang X De, Gu J, Wang T, Bi LJ, Zhang ZP, Cui ZQ, et al. Comparative analysis of mycobacterial NADH pyrophosphatase isoforms reveals a novel mechanism for isoniazid and ethionamide inactivation. Mol Microbiol. 2011;

80. Beckert P, Hillemann D, Kohl TA, Kalinowski J, Richter E, Niemann S, et al. rplC T460C identified as a dominant mutation in linezolid-resistant Mycobacterium tuberculosis strains. Antimicrob Agents Chemother. 2012;

81. Zhang S, Chen J, Cui P, Shi W, Shi X, Niu H, et al. Mycobacterium tuberculosis mutations associated with reduced susceptibility to Linezolid. Antimicrob Agents Chemother. 2016;

82. Locke JB, Hilgers M, Shaw KJ. Novel ribosomal mutations in Staphylococcus aureus strains identified through selection with the oxazolidinones linezolid and torezolid (TR-700). Antimicrob Agents Chemother. 2009;

83. Andries K, Verhasselt P, Guillemont J, Göhlmann HWH, Neefs JM, Winkler H, et al. A diarylquinoline drug active on the ATP synthase of Mycobacterium tuberculosis. Science (80-). 2005;

84. Segala E, Sougakoff W, Nevejans-Chauffour A, Jarlier V, Petrella S. New mutations in the mycobacterial ATP synthase: New insights into the binding of the diarylquinoline TMC207 to the ATP synthase C-Ring structure. Antimicrob Agents Chemother. 2012;

85. Andries K, Villellas C, Coeck N, Thys K, Gevers T, Vranckx L, et al. Acquired resistance of Mycobacterium tuberculosis to bedaquiline. PLoS One. 2014;

86. Almeida D, Ioerger T, Tyagi S, Li SY, Mdluli K, Andries K, et al. Mutations in pepQ confer low-level resistance to bedaquiline and clofazimine in Mycobacterium tuberculosis. Antimicrob Agents Chemother. 2016;

87. Organization WH. Guidance for the surveillance of drug resistance in tuberculosis [Internet]. 2020. vi, 97 p. Available from: http://apps.who.int/bookorders.

88. Schön T, Miotto P, Köser CU, Viveiros M, Böttger E, Cambau E. Mycobacterium tuberculosis drug-resistance testing: challenges, recent developments and perspectives. Clinical Microbiology and Infection. 2017.

89. Salman HS, Rüsch-Gerdes S. MGIT Procedure Manual. Mycobact Growth Indic Tube Cult Drug Susceptibility Demonstr Proj. 2006;

90. World Health Organization (Organization). Technical manual for drug susceptibility testing of medicines used in the treatment of tuberculosis. Who. 2018.

91. Mitchison DA. Drug resistance in tuberculosis. Eur Respir J. 2005;

92. Schön T, Werngren J, Machado D, Borroni E, Wijkander M, Lina G, et al. Antimicrobial susceptibility testing of Mycobacterium tuberculosis complex isolates – the EUCAST broth microdilution reference method for MIC determination. Clinical Microbiology and Infection. 2020.

93. Nguyen TNA, Berre VA Le, Bañuls AL, Nguyen TVA. Molecular diagnosis of drug-resistant tuberculosis; A literature review. Front Microbiol. 2019;

94. WHO. Module 3: Diagnosis - Rapid diagnostics for tuberculosis detection. WHO Consol Guidel Tuberc. 2020;

95. Probe-based qPCR Kits [Internet]. [cited 2021 Nov 19]. Available from: https://www.takarabio.com/learning-centers/real-time-pcr/overview/probe-based-qpcr-kits

96. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A. 1977;

97. Prober JM, Trainor GL, Dam RJ, Hobbs FW, Robertson CW, Zagursky RJ, et al. A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides. Science (80-). 1987;

98. Swerdlow H, Gesteland R. Capillary gel electrophoresis for rapid, high resolution DNA sequencing. Nucleic Acids Res. 1990;

99. Saiki RK, Gelfand DH, Stoffel S, Scharf SJ, Higuchi R, Horn GT, et al. Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. Science (80-). 1988;

100. Loman NJ, Pallen MJ. Twenty years of bacterial genome sequencing. Nature Reviews Microbiology. 2015.

101. Loman NJ, Constantinidou C, Chan JZM, Halachev M, Sergeant M, Penn CW, et al. High-throughput bacterial genome sequencing: An embarrassment of choice, a world of opportunity. Nat Rev Microbiol. 2012;

102. Prediction of Susceptibility to First-Line Tuberculosis Drugs by DNA Sequencing. N Engl J Med. 2018;

103. Walker TM, Miotto P, Köser CU, Fowler PW, Knaggs J, Iqbal Z, et al. The 2021 WHO Catalogue of *Mycobacterium Tuberculosis* Complex Mutations Associated with Drug Resistance: A New Global Standard for Molecular Diagnostics. SSRN Electron J. 2021;

104. Moreno-Molina M, Comas I, Furió V. The Future of TB Resistance Diagnosis: The Essentials on Whole Genome Sequencing and Rapid Testing Methods. Arch Bronconeumol (English Ed. 2019;

105. Auwera GA Van der, Carneiro MO, Chris Hartl RP, Angel G del, Levy-Moonshine A, Jordan T, et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinforma. 2014;

106. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;

107. 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.

108. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;

109. Perdigão J, Silva C, Diniz J, Pereira C, Machado D, Ramos J, et al. Clonal expansion across the seas as seen through CPLP-TB database: A joint effort in cataloguing Mycobacterium tuberculosis genetic diversity in Portuguese-speaking countries. Infect Genet Evol. 2019;

110. Perdigão J, Clemente S, Ramos J, Masakidi P, Machado D, Silva C, et al. Genetic diversity, transmission dynamics and drug resistance of Mycobacterium tuberculosis in Angola. Sci Rep. 2017;

111. Hunt M, Bradley P, Lapierre SG, Heys S, Thomsit M, Hall MB, et al. Antibiotic resistance prediction for Mycobacterium tuberculosis from genome sequence data with mykrobe [version 1; peer review: 2 approved, 1 approved with reservations]. Wellcome Open Res. 2019;

112. Feuerriegel S, Schleusener V, Beckert P, Kohl TA, Miotto P, Cirillo DM, et al. PhyResSE: A web tool delineating Mycobacterium tuberculosis antibiotic resistance and lineage from whole-genome sequencing data. J Clin Microbiol. 2015;

113. 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;

114. 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. 2018;

115. 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;

116. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, Von Haeseler A, et al. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Mol Biol Evol. 2020;

117. Kalyaanamoorthy S, Minh BQ, Wong TKF, Von Haeseler A, Jermiin LS. ModelFinder: Fast model selection for accurate phylogenetic estimates. Nat Methods. 2017;

118. Walker TM, Ip CLC, Harrell RH, Evans JT, Kapatai G, Dedicoat MJ, et al. Whole-genome sequencing to delineate Mycobacterium tuberculosis outbreaks: A retrospective observational study. Lancet Infect Dis. 2013;

119. Tamura K, Stecher G, Kumar S. MEGA11: Molecular Evolutionary Genetics Analysis Version 11. Mol Biol Evol. 2021;

120. Letunic I, Bork P. Interactive tree of life (iTOL) v5: An online tool for phylogenetic tree display and annotation. Nucleic Acids Res. 2021;

121. Xia E, Teo YY, Ong RTH. SpoTyping: Fast and accurate in silico Mycobacterium spoligotyping from sequence reads. Genome Med. 2016;

122. 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;

123. Gibson AL, Huard RC, Gey Van Pittius NC, Lazzarini LCO, Driscoll J, Kurepina N, et al. Application of sensitive and specific molecular methods to uncover global dissemination of the major RDRio sublineage of the Latin American-Mediterranean Mycobacterium tuberculosis spoligotype family. J Clin Microbiol. 2008;

124. Lazzarini LCO, Spindola SM, Bang H, Gibson AL, Weisenberg S, Carvalho WDS, et al. RDRio Mycobacterium tuberculosis infection is associated with a higher frequency of cavitary pulmonary disease. J Clin Microbiol. 2008;

125. De Almeida IN, Vasconcellos SEG, De Assis Figueredo LJ, Dantas NGT, Augusto CJ, Hadaad JPA, et al. Frequency of the Mycobacterium tuberculosis RDRio genotype and its association with multidrug-resistant tuberculosis. BMC Infect Dis. 2019;

126. Bocanegra-García V, Cortez-de-la-Fuente LJ, Nakamura-López Y, González GM, Rivera G, Palma-Nicolás JP. RDRio Mycobacterium tuberculosis strains associated with isoniazid resistance in Northern Mexico. Enferm Infecc Microbiol Clin. 2021;

127. Verza M, Scheffer MC, Salvato RS, Schorner MA, Barazzetti FH, Machado H de M, et al. Genomic epidemiology of Mycobacterium tuberculosis in Santa Catarina, Southern Brazil. Sci Rep. 2020;

128. Kato-Maeda M, Gagneux S, Flores LL, Kim EY, Small PM, Desmond EP, et al. Strain classification of Mycobacterium tuberculosis: Congruence between large sequence polymorphisms and spoligotypes. Int J Tuberc Lung Dis. 2011;

129. Tulu B, Ameni G. Spoligotyping based genetic diversity of Mycobacterium tuberculosis in Ethiopia: A systematic review. BMC Infect Dis. 2018;

130. Wang Y, Li Q, Gao H, Zhang Z, Liu Y, Lu J, et al. The roles of rpsL, rrs, and gidB mutations in predicting streptomycin-resistant drugs used on clinical Mycobacterium tuberculosis isolates from Hebei Province, China. Int J Clin Exp Pathol. 2019;

131. Islam MM, Tan Y, Hameed HMA, Chhotaray C, Liu Z, Liu Y, et al. Phenotypic and Genotypic Characterization of Streptomycin-Resistant Multidrug-Resistant Mycobacterium tuberculosis Clinical Isolates in Southern China. Microb Drug Resist. 2020;

132. Spies FS, Ribeiro AW, Ramos DF, Ribeiro MO, Martin A, Palomino JC, et al. Streptomycin resistance and lineage-specific polymorphisms in Mycobacterium tuberculosis gidB gene. J Clin Microbiol. 2011;

133. Perdigão J, Macedo R, Machado D, Silva C, Jordão L, Couto I, et al. GidB mutation as a phylogenetic marker for Q1 cluster Mycobacterium tuberculosis isolates and intermediate-level streptomycin resistance determinant in Lisbon, Portugal. Clin Microbiol Infect. 2014;

134. Feuerriegel S, Oberhauser B, George AG, Dafae F, Richter E, Rüsch-Gerdes S, et al. Sequence analysis for detection of first-line drug resistance in Mycobacterium tuberculosis strains from a high-incidence setting. BMC Microbiol. 2012;

135. Machado D, Couto I, Perdigão J, Rodrigues L, Portugal I, Baptista P, et al. Contribution of efflux to the emergence of isoniazid and multidrug resistance in Mycobacterium tuberculosis. PLoS One. 2012;

136. Jiang X, Zhang W, Zhang Y, Gao F, Lu C, Zhang X, et al. Assessment of efflux pump gene expression in a clinical isolate Mycobacterium tuberculosis by real-time reverse transcription PCR. Microb Drug Resist. 2008;

137. Morris RP, Nguyen L, Gatfield J, Visconti K, Nguyen K, Schnappinger D, et al. Ancestral antibiotic resistance in Mycobacterium tuberculosis. Proc Natl Acad Sci U S A. 2005;

138. Louw GE, Warren RM, Gey Van Pittius NC, McEvoy CRE, Van Helden PD, Victor TC. A balancing act: Efflux/influx in mycobacterial drug resistance. Antimicrobial Agents and Chemotherapy. 2009.

139. Chihota VN, Niehaus A, Streicher EM, Wang X, Sampson SL, Mason P, et al. Geospatial distribution of Mycobacterium tuberculosis genotypes in Africa. PLoS One. 2018;

140. Millet J, Streit E, Berchel M, Bomer AG, Schuster F, Paasch D, et al. A systematic follow-up of mycobacterium tuberculosis drug-resistance and associated genotypic lineages in the French departments of the Americas over a seventeen-year period. Biomed Res Int. 2014;

141. Millet J, Berchel M, Prudenté F, Streit E, Bomer AG, Schuster F, et al. Résistance aux antituberculeux de première ligne et principales familles génotypiques de Mycobacterium tuberculosis en région Antilles-Guyane: profils, évolution et tendances de 1995 à 2011. Bull la Soc Pathol Exot. 2014;

142. Solo ES, Suzuki Y, Kaile T, Bwalya P, Lungu P, Chizimu JY, et al. Characterization of Mycobacterium tuberculosis genotypes and their correlation to multidrug resistance in Lusaka, Zambia. Int J Infect Dis. 2021;

143. Panwalkar N, Chauhan DS, Desikan P. Spoligotype defined lineages of Mycobacterium tuberculosis and drug resistance: Merely a casual correlation? Indian Journal of Medical Microbiology. 2017.

144. Perdigão J, Gomes P, Miranda A, Maltez F, Machado D, Silva C, et al. Using genomics to understand the origin and dispersion of multidrug and extensively drug resistant tuberculosis in Portugal. Sci Rep. 2020;

145. Morelli G, Song Y, Mazzoni CJ, Eppinger M, Roumagnac P, Wagner DM, et al. Phylogenetic diversity and historical patterns of pandemic spread of Yersinia pestis. Nat Genet. 2010;

146. Nadeau SA, Vaughan TG, Scire J, Huisman JS, Stadler T. The origin and early spread of SARS-CoV-2 in Europe. Proc Natl Acad Sci U S A. 2021;

147. Clark TG, Mallard K, Coll F, Preston M, Assefa S, Harris D, et al. Elucidating emergence and transmission of multidrug-resistant tuberculosis in treatment experienced patients by whole genome sequencing. PLoS One. 2013;

148. Fine PEM, Crampin AC, Houben RMGJ, 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;

149. Bryant JM, Harris SR, Parkhill J, Dawson R, Diacon AH, van Helden P, et al. Whole-genome sequencing to establish relapse or re-infection with Mycobacterium tuberculosis: A retrospective observational study. Lancet Respir Med. 2013;

150. Stimson J, Gardy J, Mathema B, Crudu V, Cohen T, Colijn C. Beyond the SNP Threshold: Identifying Outbreak Clusters Using Inferred Transmissions. Mol Biol Evol. 2019;

151. KUBICA GP, KIM TH, DUNBAR FP. Designation of Strain H37Rv as the Neotype of Mycobacterium tuberculosis. Int J Syst Bacteriol. 1972;

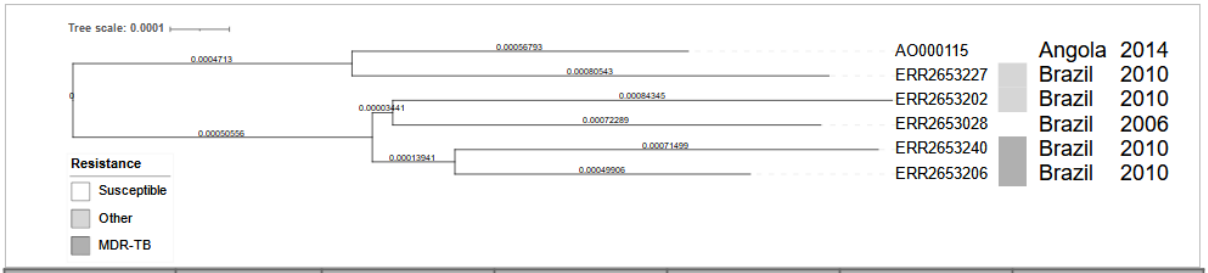
152. Al Shammari B, Shiomi T, Tezera L, Bielecka MK, Workman V, Sathyamoorthy T, et al. The Extracellular Matrix Regulates Granuloma Necrosis in Tuberculosis. In: Journal of Infectious Diseases. 2015.

153. Ioerger TR, Feng Y, Ganesula K, Chen X, Dobos KM, Fortune S, et al. Variation among genome sequences of H37Rv strains of Mycobacterium tuberculosis from multiple laboratories. J Bacteriol. 2010;

154. Solans L, Aguiló N, Samper S, Pawlik A, Frigui W, Martín C, et al. A Specific polymorphism in Mycobacterium tuberculosis H37Rv causes differential ESAT-6 expression and identifies whiB6 as a novel ESX-1 component. Infect Immun. 2014;

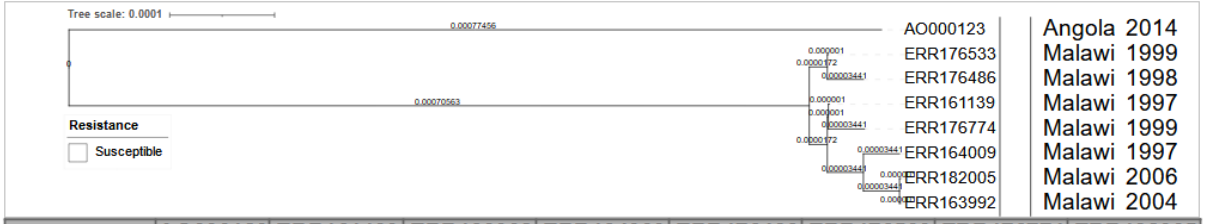
155. Fleischmann RD, Alland D, Eisen JA, Carpenter L, White O, Peterson J, et al. Whole-genome comparison of Mycobacterium tuberculosis clinical and laboratory strains. J Bacteriol. 2002;

6. Supplementary materials



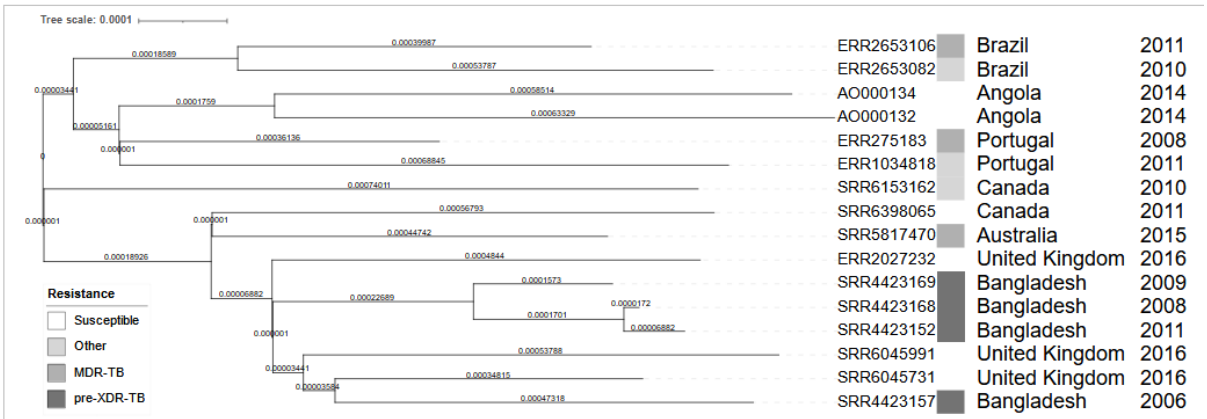
Strain	AO000115	ERR2653028	ERR2653202	ERR2653206	ERR2653227	ERR2653240
AO000115	0	133	140	126	79	138
ERR2653028	133	0	91	81	146	93
ERR2653202	140	91	0	88	153	100
ERR2653206	126	81	88	0	139	70
ERR2653227	79	146	153	139	0	151
ERR2653240	138	93	100	70	151	0

Supplementary figure S1 - Phylogenetic tree of non-clustered strain AO000115 in global dataset (above); SNP distances of each strain belonging to the phylogenetic tree (below).



Strain	AO000123	ERR161139	ERR163992	ERR164009	ERR176486	ERR176533	ERR176774	ERR182005
AO000123	0	82	90	91	89	86	84	91
ERR161139	82	0	4	4	4	2	2	4
ERR163992	90	4	0	4	8	6	6	0
ERR164009	91	4	4	0	8	6	6	4
ERR176486	89	4	8	8	0	2	6	8
ERR176533	86	2	6	6	2	0	4	6
ERR176774	84	2	6	6	6	4	0	6
ERR182005	91	4	0	4	8	6	6	0

Supplementary figure S2 - Phylogenetic tree of non-clustered strain AO000123 in global dataset (above); SNP distances of each strain belonging to the phylogenetic tree (below).



Strains (no.)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
AO000132 (1)	0	70	87	93	91	81	68	92	94	90	83	89	88	95	94	96
AO000134 (2)	70	0	84	91	89	80	65	89	91	87	80	85	84	92	92	93
ERR1034818 (3)	87	84	0	87	84	76	61	86	89	84	78	82	81	89	88	89
ERR2027232 (4)	93	91	87	0	86	77	68	55	59	52	49	58	52	61	84	65
ERR2653082 (5)	91	89	84	86	0	52	66	85	86	83	77	81	78	88	85	88
ERR2653106 (6)	81	80	76	77	52	0	57	76	79	74	68	72	71	79	78	79
ERR275183 (7)	68	65	61	68	66	57	0	67	70	65	59	63	62	70	69	70
SRR4423152 (8)	92	89	86	55	85	76	67	0	58	5	22	57	51	60	84	64
SRR4423157 (9)	94	91	89	59	86	79	70	58	0	55	52	59	46	59	86	67
SRR4423168 (10)	90	87	84	52	83	74	65	5	55	0	19	54	48	57	80	61
SRR4423169 (11)	83	80	78	49	77	68	59	22	52	19	0	49	45	54	75	57
SRR5817470 (12)	89	85	82	58	81	72	63	57	59	54	49	0	51	60	79	59
SRR6045731 (13)	88	84	81	52	78	71	62	51	46	48	45	51	0	53	79	59
SRR6045991 (14)	95	92	89	61	88	79	70	60	59	57	54	60	53	0	87	67
SRR6153162 (15)	94	92	88	84	85	78	69	84	86	80	75	79	79	87	0	86
SRR6398065 (16)	96	93	89	65	88	79	70	64	67	61	57	59	59	67	86	0

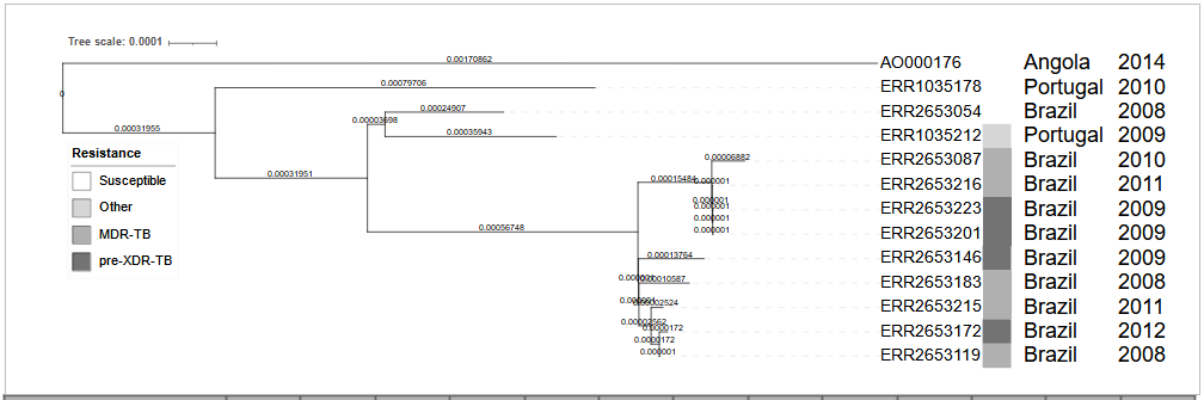
Supplementary figure S3 - Phylogenetic tree of non-clustered strains AO000132 and AO000134 in global dataset (above); SNP distances of each strain belonging to the phylogenetic tree (below). Due to space limitations, strains in columns are represented according to the assigned number in their respectibe rows.

Strain	AO000142	ERR1034864
AO000142	0	63
ERR1034864	63	0

Supplementary figure S4 - Phylogenetic tree of non-clustered strains AO000142 in global dataset (above); SNP distances of each strain belonging to the phylogenetic tree (left).

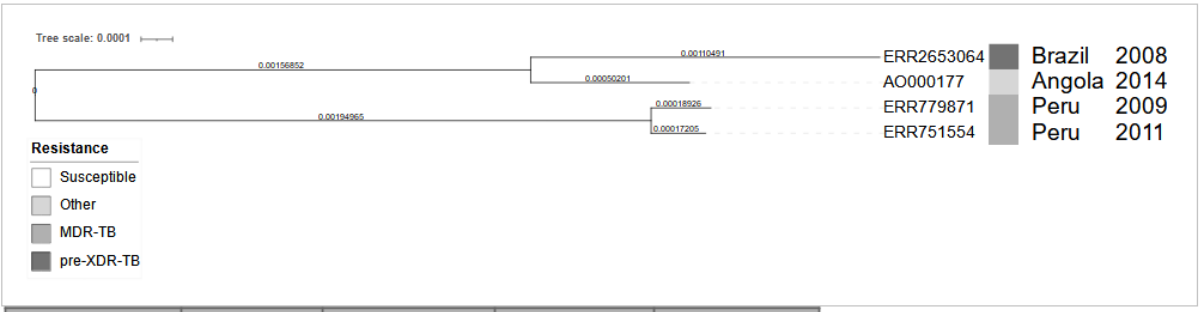
Strain	AO000149	ERR1035216	ERR1036275	ERR1036278	ERR2653030	ERR275190
AO000149	0	66	91	81	69	82
ERR1035216	66	0	84	74	62	75
ERR1036275	91	84	0	84	74	87
ERR1036278	81	74	84	0	66	77
ERR2653030	69	62	74	66	0	46
ERR275190	82	75	87	77	46	0

Supplementary figure S5 - Phylogenetic tree of non-clustered strain AO000149 in global dataset (above); SNP distances of each strain belonging to the phylogenetic tree (below).



Strain (no.)	1	2	3	4	5	6	7	8	9	10	11	12	13
AO000176 (1)	0	159	153	133	181	169	176	166	166	177	171	177	177
ERR1035178 (2)	159	0	83	70	105	93	99	91	92	101	95	101	101
ERR1035212 (3)	153	83	0	33	68	57	63	57	59	64	58	64	64
ERR2653054 (4)	133	70	33	0	61	50	56	50	52	57	51	57	57
ERR2653087 (5)	181	105	68	61	0	15	21	16	19	4	16	4	4
ERR2653119 (6)	169	93	57	50	15	0	10	1	8	11	2	11	11
ERR2653146 (7)	176	99	63	56	21	10	0	11	14	17	11	17	17
ERR2653172 (8)	166	91	57	50	16	1	11	0	9	12	3	12	12
ERR2653183 (9)	166	92	59	52	19	8	14	9	0	15	8	15	15
ERR2653201 (10)	177	101	64	57	4	11	17	12	15	0	12	0	0
ERR2653215 (11)	171	95	58	51	16	2	11	3	8	12	0	12	12
ERR2653216 (12)	177	101	64	57	4	11	17	12	15	0	12	0	0
ERR2653223 (13)	177	101	64	57	4	11	17	12	15	0	12	0	0

Supplementary figure S6 - Phylogenetic tree of non-clustered strains AO000176 in global dataset (above); SNP distances of each strain belonging to the phylogenetic tree (below). Due to space limitations, strains in columns are represented according to the assigned number in their respectibe rows.



Strain	AO000177	ERR2653064	ERR751554	ERR779871
AO000177	0	90	242	243
ERR2653064	90	0	268	269
ERR751554	242	268	0	21
ERR779871	243	269	21	0

Supplementary figure S7 - Phylogenetic tree of non-clustered strains AO000177 in global dataset (above); SNP distances of each strain belonging to the phylogenetic tree (left).

Supplementary table ST1 - Comparative table of results of different *in silico* drug resistance prediction software tools. Mutations assigned to each resistance in brackets.

Strain	Phyres	Mykrobe	TB-Profiler
AO000110	RIF(*rpoB* S450L), INH(*katG* S315T), EMB(*embB* M306I)	RIF(*rpoB* S450L), INH(*katG* S315T), EMB(r)(*embB* M306I)	RIF(*rpoB* S450L), INH(*katG* S315T), EMB(*embB* M306I), ETO(*ethA* c.735 741del)
AO000111	RIF(*rpoB* S450L), INH(*katG* S315T), EMB(*embB* M306V), STR(*gidB* V88A)	RIF(*rpoB* S450L), INH(*katG* S315T), EMB(*embB* M306V), STR(r)(*gid* T137W), PZA(*pncA* GAG430TAG-CTC2288810CTA)	INH(*katG* S315T), EMB(*embB* M306V)
AO000113	Sensitive	Sensitive	Sensitive
AO000115	Sensitive	Sensitive	Sensitive
AO000121	Sensitive	Sensitive	Sensitive
AO000123	Sensitive	Sensitive	Sensitive
AO000125	Sensitive	Sensitive	Sensitive
AO000127	Sensitive	Sensitive	Sensitive
AO000131	Sensitive	Sensitive	Sensitive
AO000132	Sensitive	Sensitive	Sensitive
AO000134	Sensitive	Sensitive	Sensitive
AO000138	Sensitive	Sensitive	Sensitive
AO000142	Sensitive	Sensitive	Sensitive

AO000149	Sensitive	Sensitive	Sensitive
AO000169	INH(*katG* S315T)	INH(*katG* S315T)	INH(*katG* S315T)
AO000176	Sensitive	Sensitive	Sensitive
AO000177	INH(*katG* S315T)	INH(*katG* S315T)	INH(*katG* S315T)

Supplementary table ST2 - Other mutations detected by TB-Profiler, by position in genome, coding DNA or protein, where applicable.

Strain	Mutation (change)
AO000110	7362(p.Glu21Gln), 7585(p.Ser95Thr), 9304(p.Gly668Asp), 764995(c.1626C>G), 781395(c.-165T>C), 1472753(r.908a>c), 1917972(c.33A>G), 2289141(p.Tyr34Cys), 3073868(p.Thr202Ala), 3086788(c.-32T>C), 3841652(c.-232A>T), 3841654(c.-234A>C), 3841662(c.-242A>G), 3841663(c.-243G>A), 4242643(c.2781C>T), 4407791(p.Ala138Pro), 4408156(p.Leu16Arg)
AO000125	7362(p.Glu21Gln), 781395(c.-165T>C), 1474001(r.344c>t), 1917972(c.33A>G), 2714787(p.Trp182*), 3841652(c.-232A>T), 3841654(c.-234A>C), 3841662(c.-242A>G), 3841663(c.-243G>A)
AO000111	7362(p.Glu21Gln), 7585(p.Ser95Thr), 9304(p.Gly668Asp), 764817(p.Val483Gly), 764995(c.1626C>G), 781395(c.-165T>C), 1917972(c.33A>G), 2288812(p.Glu144*), 3073868(p.Thr202Ala), 3086788(c.-32T>C), 3841652(c.-232A>T), 3841654(c.-234A>C), 4242643(c.2781C>T), 4407794(p.Arg137Trp), 4407940(p.Val88Ala), 4408156(p.Leu16Arg)
AO000113	7362(p.Glu21Gln), 7585(p.Ser95Thr), 9304(p.Gly668Asp), 764995(c.1626C>G), 781395(c.-165T>C), 1917972(c.33A>G), 3073868(p.Thr202Ala), 3086788(c.-32T>C), 3841652(c.-232A>T), 3841654(c.-234A>C), 3841662(c.-242A>G), 3841663(c.-243G>A), 4242643(c.2781C>T), 4408156(p.Leu16Arg)
AO000115	6140(p.Val301Leu), 7362(p.Glu21Gln), 7585(p.Ser95Thr), 9304(p.Gly668Asp), 764995(c.1626C>G), 781395(c.-165T>C), 1917972(c.33A>G), 3073868(p.Thr202Ala), 3086788(c.-32T>C), 3840719(c.702T>C), 4242643(c.2781C>T), 4408156(p.Leu16Arg)
AO000121	7362(p.Glu21Gln), 7585(p.Ser95Thr), 9304(p.Gly668Asp), 765150(p.Gly594Glu), 781395(c.-165T>C), 1476055(r.2398c>t), 1917972(c.33A>G), 2726210(c.18T>C), 3086788(c.-32T>C), 3641447(p.Thr302Met), 4240897(c.1035C>G), 4242643(c.2781C>T), 4242803(p.Val981Leu), 4249408(c.2895G>A)
AO000123	7362(p.Glu21Gln), 7565(c.264C>G), 7585(p.Ser95Thr), 8715(p.Pro472Ser), 9304(p.Gly668Asp), 759608(c.-199C>T), 764995(c.1626C>G), 781395(c.-165T>C), 1917972(c.33A>G), 2726323(p.Pro44Arg), 3073868(p.Thr202Ala), 3086788(c.-32T>C), 3640384(c.-159T>G), 3841006(p.Asp139His), 4242643(c.2781C>T), 4408156(p.Leu16Arg)
AO000127	7362(p.Glu21Gln), 7585(p.Ser95Thr), 9304(p.Gly668Asp), 765150(p.Gly594Glu), 781395(c.-165T>C), 1917972(c.33A>G), 3086788(c.-32T>C), 4242643(c.2781C>T), 4242803(p.Val981Leu), 4249408(c.2895G>A), 4327691(p.Asp48Gly)
AO000134	7362(p.Glu21Gln), 7585(p.Ser95Thr), 9304(p.Gly668Asp), 761707(p.Asp634Gly), 764995(c.1626C>G), 781395(c.-165T>C), 1917972(c.33A>G), 3073868(p.Thr202Ala), 3086788(c.-32T>C), 3087872(c.1053G>A), 4242643(c.2781C>T), 4408156(p.Leu16Arg)
AO000131	7362(p.Glu21Gln), 7585(p.Ser95Thr), 9304(p.Gly668Asp), 764995(c.1626C>G), 781395(c.-165T>C), 1917972(c.33A>G), 3086788(c.-32T>C), 3841652(c.-232A>T), 3841654(c.-234A>C), 3841662(c.-242A>G), 3841663(c.-243G>A), 4242643(c.2781C>T), 4408156(p.Leu16Arg)
AO000132	7362(p.Glu21Gln), 7585(p.Ser95Thr), 8555(c.1254G>T), 9304(p.Gly668Asp), 761707(p.Asp634Gly), 764995(c.1626C>G), 781395(c.-165T>C), 1917972(c.33A>G), 3073868(p.Thr202Ala), 3086788(c.-32T>C), 4242643(c.2781C>T), 4408156(p.Leu16Arg)
AO000138	7362(p.Glu21Gln), 7585(p.Ser95Thr), 8040(p.Gly247Ser), 9304(p.Gly668Asp), 764995(c.1626C>G), 781395(c.-165T>C), 1476056(r.2399g>a), 1674748(p.Gly183Ser), 1917972(c.33A>G), 2518919(p.Gly269Ser), 3073868(p.Thr202Ala), 3086788(c.-32T>C), 4242643(c.2781C>T), 4245083(c.1851A>G), 4408156(p.Leu16Arg)
AO000142	6817(c.1578G>A), 7362(p.Glu21Gln), 7585(p.Ser95Thr), 9304(p.Gly668Asp), 764995(c.1626C>G), 781395(c.-165T>C), 1917972(c.33A>G), 1918247(p.Arg103His), 3073868(p.Thr202Ala), 3086788(c.-32T>C), 4242643(c.2781C>T), 4408156(p.Leu16Arg)

AO000149	7362(p.Glu21Gln), 7585(p.Ser95Thr), 9304(p.Gly668Asp), 764995(c.1626C>G), 781395(c.-165T>C), 1917972(c.33A>G), 3073868(p.Thr202Ala), 3086788(c.-32T>C), 3841652(c.-232A>T), 3841654(c.-234A>C), 3841662(c.-242A>G), 3841663(c.-243G>A), 4242643(c.2781C>T), 4408156(p.Leu16Arg)
AO000169	7362(p.Glu21Gln), 7585(p.Ser95Thr), 9304(p.Gly668Asp), 765150(p.Gly594Glu), 781395(c.-165T>C), 1471833(c.-13G>A), 1917972(c.33A>G), 2726210(c.18T>C), 3086788(c.-32T>C), 3641447(p.Thr302Met), 4240897(c.1035C>G), 4242643(c.2781C>T), 4242803(p.Val981Leu), 4249408(c.2895G>A)
AO000176	7585(p.Ser95Thr), 8040(p.Gly247Ser), 9304(p.Gly668Asp), 764995(c.1626C>G), 781395(c.-165T>C), 800990(p.Lys61Thr), 1917972(c.33A>G), 2518919(p.Gly269Ser), 3073868(p.Thr202Ala), 3086788(c.-32T>C), 4242182(p.Ala774Ser), 4242643(c.2781C>T), 4408156(p.Leu16Arg)
AO000177	7362(p.Glu21Gln), 7585(p.Ser95Thr), 9304(p.Gly668Asp), 764995(c.1626C>G), 781395(c.-165T>C), 1472753(r.908a>c), 1917972(c.33A>G), 3073868(p.Thr202Ala), 3086788(c.-32T>C), 3841652(c.-232A>T), 3841654(c.-234A>C), 3841662(c.-242A>G), 3841663(c.-243G>A), 4242643(c.2781C>T), 4326587(p.Val296Gly), 4407947(p.Leu86Phe), 4408156(p.Leu16Arg)