Ana Rita Frazão Macedo
Mestre em Microbiologia Clínica

Tuberculosis: new era for diagnosis and surveillance using whole-genome sequencing-based approaches

Dissertação para obtenção do Grau de Doutor em Biologia

Orientador: Doutora Isabel Portugal, Professora Auxiliar, iMed.ULisboa - Instituto de Investigação do Medicamento Faculdade de Farmácia, Universidade de Lisboa
Co-orientador: Doutor João Paulo dos Santos Gomes, Investigador Auxiliar com Habilitação, Instituto Nacional de Saúde Doutor Ricardo Jorge
Co-orientador: Doutor Jaime Mota, Professor Auxiliar, Faculdade de Ciências e Tecnologia, Universidade NOVA de Lisboa

Júri:
Presidente: Prof. Doutor Pedro Miguel Ribeiro Viana Baptista
Arguentes: Prof. Doutora Raquel Sá Leão
Prof. Doutora Laura Maria Brum da Cruz Martins
Vogais: João Paulo dos Santos Gomes
João Ruben Lucas Perdigão

FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE NOVA DE LISBOA

julho, 2019

Ana Rita Frazão Macedo
Mestre em Microbiologia Clínica

Tuberculosis: new era for diagnosis and surveillance using whole-genome sequencing-based approaches

Acknowledgments

Ao Doutor João Paulo Gomes. Penso que já lhe disse tudo e não há nada que possa escrever que vá acrescentar algo ao que ele já sabe. Agradeço a amizade, o carinho, a PACIÊNCIA (!!), o incentivo e motivação, principalmente nas alturas em que desesperava e achava que não iria ser capaz. Obrigada João, sabes que não teria conseguido sem ti.

À Doutora Isabel Portugal. Que me tem acompanhado em todas estas viagens de pré e pós-graduação. O percurso tem sido longo e cheio de "caminhos desviados", mas temo-nos sempre encontrado nas alturas certas.

Ao Doutor Jaime Mota por ter aceite o papel de meu orientador e pela simpatia e disponibilidade demonstradas ao longo deste caminho.

Aos restantes membros da Comissão de Acompanhamento de Tese (CAT), Doutora Isabel Couto, pela discussão científica e por todas as sugestões concedidas.

À Doutora Raquel Duarte que sempre me disse o que precisava de ouvir em todos os momentos. Pela amizade, pela disponibilidade e toda a contribuição e dedicação a este trabalho.

À minha equipa no laboratório, Inês João, Irene Rodrigues, Sónia Silva e Cristina Matos. Todas sabem que nunca teria conseguido sem elas; a disponibilidade que me deram ao assumir todas as tarefas extra e todos os incentivos que não me deixaram desistir. Obrigada é muito pouco. Prometo chocolates todas as semanas!

À Inês João. Minha "guru" e motivadora em todas as horas. Não sei como vou "sobreviver" sem ti.

Às minhas companheiras de almoço, e não só, que me aturaram sempre, principalmente nos meus piores momentos. Um agradecimento especial à Joana e Leonor, vocês sabem e não preciso escrevê-lo aqui. Por todos os "after works" e deixarem a vossa casa para estarem na minha.

À Andrea. Por tudo.

À Doutora Maria João Simões. Que me adotou desde que "nasci" no INSA.

Aos colegas da bioinformática que foram incansáveis e que tanto se cansaram em tentar satisfazer todos os meus caprichos e dúvidas (algumas ainda ficaram, mas a aluna não é das melhores...). Joana, Alexandra, Miguel e Vítor tem sido um prazer e uma oportunidade poder trabalhar convosco (e dar-vos trabalho!).

A toda a equipa da Unidade de Tecnologia e Inovação do INSA. Em especial, ao seu responsável, Doutor Luís Vieira, pelo seu empenho na implementação e otimização das metodologias de Sequenciação de Nova Geração, as quais foram essenciais para a execução deste trabalho.

Ao coordenador, Doutor Jorge Machado, e restantes responsáveis do DDI por me terem possibilitado o alcance deste objetivo.

À Faculdade de Ciências e Tecnologia da Universidade NOVA de Lisboa (FCT/UNL), em particular à Professora Isabel Sá Nogueira, Coordenadora do Programa Doutoral de Biologia.

A todos os meus amigos (alguns já aqui mencionados). Que sempre me apoiaram e me deram a força quando dela precisava.

À minha família. Mãe, Pai, Irmã, Obrigada por tudo. Embora algumas vezes com demonstração de alguma ingratidão (da minha parte, claro), só sou hoje quem sou graças a vocês. Desculpem o stress e os desabafos maldispostos durante o percurso. Prometo finalizar todas estas etapas e deixar de ser uma preocupação.

Para os meus filhos. Para saberem que mesmo nas piores circunstâncias, basta um bocadinho de força e alguém que acredite em nós. Desculpem a irritação, o cansaço e, às vezes, a falta de disponibilidade. Isto é tudo para vocês.

Resumo

Desde 1993 que a OMS declarou a Tuberculose (TB) como uma emergência de saúde pública global. É, atualmente, responsável por quase 2 milhões de mortes por ano, sendo a nona principal causa de morte em todo o mundo. O principal obstáculo para o controlo efetivo da TB é a resistência aos antibacilares, havendo assim a necessidade de implementação de novas tecnologias de diagnóstico rápido que possam traduzir-se no início precoce do tratamento e bloqueio das cadeias de transmissão.

Considerando os constrangimentos para o isolamento e tempo de crescimento das estirpes de *M. tuberculosis*, o principal objetivo desta dissertação consistiu em avaliar o potencial do uso de metodologias baseadas na Sequenciação Total do Genoma (WGS) para o diagnóstico de rotina e vigilância epidemiológica. Procedeu-se à avaliação de várias plataformas bioinformáticas para previsão *in silico* dos perfis de resistência aos antibacilares, bem como ao desenvolvimento de "pipelines" bioinformáticas para vigilância epidemiológica. Estas abordagens revelaram uma elevada sensibilidade quando comparadas com as metodologias tradicionais, tendo sido já implementadas na rotina laboratorial do Laboratório Nacional de Referência (LNR). Adicionalmente, demonstrámos a possibilidade de usar essas mesmas metodologias diretamente em amostras clínicas, diminuindo o tempo de resposta para cinco a oito dias. Além disso, e de acordo com as novas recomendações para o tratamento da TB, iniciámos estudos para identificação de mutações associadas à resistência aos fármacos recentemente adotados, de forma a enriquecer as bases de dados e, consequentemente, a performance do diagnóstico genotípico.

Concluindo, acreditamos ter contribuído para a validação das metodologias baseadas em WGS como ferramentas para ultrapassar as dificuldades do diagnóstico e vigilância fenotípica da TB em particular, na sua capacidade de fornecer informações muito mais rápidas sobre previsão de resistência e transmissão. Por último, este trabalho esteve na base da transição tecnológica iniciada no LNR para a vigilância da tuberculose.

Palavras-chave

Tuberculose, Multirresistência, Whole-genome sequencing, Vigilância, Mutações associadas a resistência, Epidemiologia

Abstract

Tuberculosis (TB) has been declared as a global public health emergency by the WHO since 1993. It still accounts for almost 2 million deaths each year, making it the ninth leading cause of death worldwide. The major obstacle for an effective TB control is antimicrobial resistance, thus, to be successful, new strategies must be addressed, for instance, the implementation of new rapid TB diagnostic technologies that could translate into early treatment initiation and blocking of transmission chains.

Considering the major constraints regarding the isolation and time of growth of *M. tuberculosis* strains, the main goal of this PhD dissertation was to acknowledge the potential of the use of WGS-based methodologies for routine diagnostic and epidemiological surveillance. We evaluated several software for *in silico* prediction of antibiotic resistance and developed bioinformatics pipelines for surveillance purposes, in particular for the identification of transmission chains. As they revealed high sensitivity, these approaches are already implemented in the routine of the Portuguese National Reference Laboratory (NRL). We also recognised the possibility to use these same approaches directly to samples collected from TB patients, lowering the time-to-results, for a complete drug resistance pattern and phylogeny analysis, for five to eight days. The validation of this methodology is ongoing and will be implemented in a near future. Additionally, and according to the new recommendations for TB treatment, we have initiated studies to identify new mutations associated with resistance to the recently adopted drugs, in order to enrich the available databases and improve the performance of the genotypic diagnostics pipelines.

This PhD dissertation highlights WGS-based methodologies as powerful tools to surpass the difficulties of phenotypic TB diagnosis and surveillance and to provide a much more rapid information regarding resistance prediction and eventual transmission chains. It also supported the technological transition performed at the NRL for TB surveillance.

Keywords

Tuberculosis, Multidrug resistance, Whole-genome sequencing, Surveillance, Resistance-associated Mutations, Epidemiology

Table of contents

Figure Index

Table index

List of Abbreviations

AD - allele differences

AFB - acid-fast bacilli

AG - arabinogalactan

AMK - amikacin

BDQ - bedaquiline

CAP - capreomycin

CFU - colony forming units

CICLO - cicloserine

CIP - ciprofloxacin

DLM - delamanid

DST - drug susceptibility testing

ECDC - European Center for Disease Control and Prevention

EMB - ethambutol

ETH - ethionamide

FQ - fluoroquinolone

GAT - gatifloxacin

GHD - General Health Directorate

HIV - Human-Immunodeficiency Virus

IGV - Integrative Genomics Viewer

INH - isoniazid

KAN - kanamycin

LAM - lipoarabinomannan

LTV - Lisbon and Tagus Valey

LSP - large sequence polymorphism

LVX - levofloxacin

LNZ - linezolid

MDR - multidrug resistant

MIC - Minimum Inhibitory Concentration

MIRU - Mycobacterial Interspersed Repetitive Units

MTBC - *M. tuberculosis* complex

MST - Minimum Spanning Tree

MXF - moxifloxacin

NAAT - Nucleic Acid Amplification Tests

NCCLS - National Committee for Clinical Laboratory Standards

NIH - National Institute of Health

NRL - National Reference Laboratory

NPV - Negative Predicted Value

NTM - Non-Tuberculous Mycobacteria

NTP - National TB program

OFX - ofloxacin

PAS - para-aminosalicylic acid

POA - pyrazinoic acid

PPV - Positive Predicted Value

PZA - pyrazinamide

Pzase - pyrazinamidase

RFLP - Restriction Fragment Length Polymorphism

RIF - rifampicin

RRDR - RIF Resistance Determining Region

RMP - rifampicin

SLIDs - Second-Line Injectable Drugs

SNP - Single Nucleotide Polymorphism

SNV - Single Nucleotide Variant

SRA - Sequence Read Archive

STR - streptomycin

TB - tuberculosis

VNTR - Variable Number of Tandem Repeat

WGS - Whole-Genome Sequencing

WHO - World Health Organization

XDR - extensively-drug resistant

ZN - Ziehl-Neelsen

Notes of the author: thesis organization, format and outline

This PhD dissertation is composed of seven chapters, including an Introduction, several research studies (either published or ongoing), and a final discussion. Its core is based on three manuscripts (listed below) that are presented as individual chapters and two additional chapters that include a proof of concept of methodologies to be soon implemented and submitted to a journal and an undergoing study with the major breakthroughs at the moment. The manuscripts have already been published in peer reviewed international journals, and the corresponding chapters essentially represent what was published. The chapters were organized so that they follow a rational order taking into account the objectives delineated for this PhD work. Each manuscript-based chapter is preceded by a title page describing the reference of the publication, the specific contributions of the author of the present PhD thesis, and, when applicable, the alterations that were performed regarding what is published (referred as "minor changes"). In brief, each chapter includes the following contents:

Chapter I. This chapter consists of a general introduction that intends to provide the reader with the state of the art in the subjects addressed in this doctoral dissertation around Tuberculosis, the first cause of death from an infectious disease. It includes a global overview of the major aspects of *M. tuberculosis* taxonomy, biology, molecular epidemiology and impact on human health, followed by insights into the genetic diversity and some already established genotype/phenotype associations. It ends with the description of the main objectives of this PhD project, which includes the specific research questions that drove the investigations carried out on behalf of each chapter.

Chapter II. Genetic prediction of antibiotic resistance

This chapter corresponds to the following published manuscript: “Rita Macedo, Alexandra Nunes, Isabel Portugal, Sílvia Duarte, Luís Vieira, João Paulo Gomes. Dissecting whole-genome sequencing-based online tools for predicting resistance in *Mycobacterium tuberculosis*: can we use them for clinical decision guidance? 2018. Tuberculosis. 110: 44–51”. It evaluates the use of several bioinformatics pipelines for *in silico* prediction of antibiotic resistance in *M. tuberculosis* aiming at overcoming the time-consuming laboratory procedure underlying the antibiotic susceptibility tests. The ultimate goal of this study was to implement one of the evaluated bioinformatics-based approaches in the routine practice of the National Reference Tuberculosis Laboratory at INSA.

Chapter III. Epidemiology of multidrug resistant tuberculosis in Portugal

This chapter corresponds to the following manuscript: "Rita Macedo, Raquel Duarte. Trends of MDR-TB clustering in Portugal. 2019. ERJ Open Research 5: 00151-2018". In general, this study focus on understanding the dynamics of MDR-TB emergence and transmission, to establish the rate of recent transmissions against newly developed resistant strains, to pinpoint the emergence of new cases in the population and to identify associated risk factors.

Chapter IV. Development of genomic-based surveillance methodologies

This chapter corresponds to the following manuscript: "Rita Macedo, Miguel Pinto, Vítor Borges, Alexandra Nunes, Olena Oliveira, Isabel Portugal, Raquel Duarte, João Paulo Gomes. Evaluation of a gene-by-gene approach for prospective whole-genome sequencing-based surveillance of multidrug resistant *Mycobacterium tuberculosis*. 2019. Tuberculosis 115: 81-88". This study ultimately aimed to implement a whole-genome-sequencing-based approach for surveillance purposes at the National Reference Tuberculosis Laboratory at INSA, following the recommendations of the international health authorities.

Chapter V. Whole-genome-sequencing of *Mycobacterium tuberculosis* directly from clinical samples

This chapter corresponds to an ongoing study enrolling the PhD student and the team of the Bioinformatics Unit of the Department of Infectious Diseases at INSA. It aims at developing and optimizing a laboratory procedure to capture *M. tuberculosis* genomes directly from the clinical samples without the need for culture propagation of the infecting strains. If this approach turns out to be a feasible method, it could be applied upstream of the use of the bioinformatics pipeline to predict antibiotic resistance (described in Chapter II). Using these two approaches in a tandem fashion would tremendously decrease (i.e., in several weeks) the time necessary to determine the suitable antibiotic therapy.

Chapter VI. Disclosing the genetic basis of antibiotic resistance

This chapter corresponds to a long-term ongoing study enrolling the same teams as the previous study. By using *in vitro* selective pressure scenarios (i.e., antibiotic pressure using sub-MIC) during *M. tuberculosis* propagation, this study intends to identify the mutations responsible for the emergence of clones with decreased susceptibility to a specific antibiotic, and thus enrich

the available genetic databases. Ultimately, it may contribute to the accuracy improvement of the software platforms used for *in silico* prediction of antibiotic resistance (described in chapter II).

Chapter VII. This chapter provides a global overview of the subjects addressed throughout the chapters, highlighting the main results and conclusions achieved in this PhD dissertation. Considering that each chapter focusing a research study contains its own “Discussion”, only the most relevant results are discussed in this final chapter in order to avoid redundancy. New research questions raised with this work that can be addressed in the future follow-up of these investigations are also presented.

Considering the dissimilar layouts and in-text reference styles adopted by different journals where the manuscripts were published, all chapters were formatted in a unique style, with all references being cited by sequential numbers (in parentheses) and listed in the "References" section according to the order in which they appear in the text. In this regard, a single section of "References" is presented.
Similarly, all the supplemental material is presented at the final of this PhD thesis, enumerated accordingly with the chapter they concern to.

Chapter I

General Introduction

1. General Introduction

1.1 The genus *Mycobacteria*

Lehmann and Neumann first introduced the genus *Mycobacterium* to the scientific community in 1896 (1). The subsequent history of the genus has been profoundly influenced by the fact that only very few of the almost 200 currently recognized species (http://www.bacterio.net/mycobacterium.html) have been a cause of human disease, above all, *M. tuberculosis*. Thus, studies of microbial physiology, structure, genetics and diagnostic tools have mainly focused on *M. tuberculosis*.

Mycobacterium is the only genus in the family of the Mycobacteriaceae, as it is defined in Bergey's Manual of Systematic Bacteriology (2), but it is considered to be closely related to other mycolic acid-containing genera: Caseobacter, Corynebacterium, Nocardia and Rhodococcus (3). All mycobacteria are aerobic (though some species are able to grow under a reduced oxygen atmosphere), nonspore-forming, nonmotile, slightly curved or straight rods (0.2–0.6 × 1.0–10 µm). Many species form whitish or creamcolored colonies, but especially among the rapid growers, there are also many bright yellow or orange species containing carotenoid pigments (4). In some cases, the pigments are only formed in response to light (photochromogenic species), but most pigmented species also form these pigments in the dark (scotochromogenic species). The classification of Runyon separates the genus *Mycobacterium* into four groups (photochromogens, scotochromogens, nonphotochromogens, and rapid growers) and was introduced in the late 1950s as a systematic base for the description of mycobacteria (5). This division, based on pigmentation and growth rate, is still of use to the clinical mycobacteriologist and the separation of the genus into two major groups on the basis of the growth rate of the individual species forms the basis of the mycobacterial taxonomy. The most prominent feature of mycobacteria that is uniformly present and distinctive of the genus is the lipid-rich cell envelope (1). Indeed, it is the complex cell envelope of mycobacteria that confers these bacteria the property of 'acidfastness' (i.e. resistance to decolourization when stained with carbolfuchsin and decolorized with dilute hydrochloric acid). Uniformly, they do not stain well with Gram stain and should be considered gramneutral. Mycobacteria possess a cell wall polysaccharide that resembles that of gram-positive bacteria; however, the mycobacterial peptidoglycan contains lipids in place of proteins and polysaccharides (1). Furthermore, the mycobacterial envelope contains a plasma membrane that is quite similar in structure and function to the plasma membrane of other bacteria, except for the presence of lipoarabinomannan (LAM), lipomannan and phosphatidylinositol mannosides. As a whole, the cell wall component of the envelope

confers size, shape, protection against osmotic pressure and probably protects the plasma membrane from deleterious molecules present in the environment of the cell. In summary, the peptidoglycan confers cell shape while the next layer of the envelope, arabinogalactan esterified to the mycolic acids, provides a hydrophobic permeability barrier. Other important fatty acids are waxes, phospholipids and mycoserosic and phthienoic acids, and tuberculostearic acid (10-R-methyl-octadecanoic), a unique cell component within the *Actinomycetales*, including the mycobacteria (1).

In 1947, Middlebrook first described growth of tubercle bacilli in the shape of serpentine cords ('cording'). For many years, cording was correlated with virulence and considered a distinctive feature of *M. tuberculosis*. However, it is now known that several mycobacterial species display cording and the correlation with virulence, if any, is unclear (1).

1.1.1 Taxonomy

The number of *Mycobacterium* species has increased from about 40 in 1980 (6) to over 180 in 2018 (http://www.bacterio.net/mycobacterium.html). The description of novel species is paralleled by the development of molecular methods and by the increased recognition that slow growing mycobacteria are clinically important and fast-growing mycobacteria are ecologically important. By the end of 1983, there were 52 described species, only six new species were added between 1984 and 1991, about four new species *per* year between 1992 and 2003 and most of the non-tuberculous mycobacteria (NTM) species were identified in recent years.

Currently the genus is broadly divided into "slow growers" and "rapid growers". Rapid growers are those species that under optimal solid culture conditions grow visible colonies within seven days. The slow growers exceed this time, in some cases, in several additional weeks. The most notable members of the slow growers belong to the *M. tuberculosis* complex (MTBC), which cause tuberculosis (TB) in both humans and animals. Another slow-grower is *M. ulcerans*, which is the cause of the Buruli Ulcer, a neglected tropical disease with its highest incidence in sub-Saharan Africa (7). Also of note is *M. avium* subsp. paratuberculosis, which causes Johnes disease in cattle and has long been suspected (but not yet proven) to be a contributor to Crohn's disease in humans (8). *M. leprae* causes leprosy, a disabling disease which is still endemic in isolated pockets of the world (9). All of the known rapid growing Mycobacteria are primarily environmental, with some having the ability to become opportunistic pathogens. The most

virulent and clinically relevant of these is *M. abscessus*, which can cause both wound and respiratory infections (10).

1.2 The *Mycobacterium tuberculosis* complex

The species *M. tuberculosis* belongs to the "*M. tuberculosis* complex" (MTBC), which is a group of closely related species that can cause TB disease in animals. Although currently defined as different species, they fall short of the minimum standard to be considered true species (i.e., more than 5% nucleotide divergence). Nevertheless, there are clear phenotypic and epidemiological differences between the members of the complex: *M. tuberculosis* is strictly a human pathogen; *M. bovis* and *M. caprae* can infect a wide range of animals, but of primary concern is its burden in cattle; *M. africanum* is mostly found in humans but it seems to be restricted geographically to West Africa (11); *M. canettii* is the most divergent species, differing from *M. tuberculosis* by at least 2% at the nucleotide level. It has unusual smooth colony morphology and a lower virulence (12). More recently, additional species of the MTBC were identified. One of those has emerged in banded mongooses (*Mungos mungo*) in Botswana and was named mongoose bacillus, or *M. mungi* sp. nov. This pathogen causes high mortality rates among banded mongooses that live in close association with humans because these animals live in human-made structures and scavenge human waste, including feces (13). Another new species is *M. suricattae*, isolated from free-living meerkats (*Suricata suricatta*) from the Kalahari Desert, South Africa, and was first reported in 2002 (14). Finally, Oryx bacilli, *M. oryx*, have been isolated from members of the *Bovidae* family, i.e., oryxes, gazelles, deer, antelope, and waterbucks, although their exact host range remains unsettled (15). However, for these novel subspecies, no cases of human disease have yet been reported to date.

The completion of the first *M. tuberculosis* reference genome (16) provided the opportunity to detect large sequence polymorphisms (LSPs). These LSPs were used as markers to reflect the deep evolutionary relationships between the members of the complex. Remarkably, they provided evidence that refuted the commonly proposed idea that human TB evolved from a bovine progenitor, as *M. bovis* was found to have diverged more recently than the other human strains (17).

Our knowledge of the MTBC was increased by sequence-based analyses of genes (18), and, more recently, by whole-genome sequencing analysis (19). This revealed the presence of seven human lineages, and one animal lineage, which includes *M. bovis* (Figure 1.1). *M. africanum* is split into two distinct lineages, West African 1 and 2. The other lineages are comprised of geographically

structured *M. tuberculosis* strains; Lineage 4, the Euro-American lineage, is the most widespread and commonly isolated (20) and Lineage 2, the East-Asian lineage, is split into Beijing and non-Beijing strains. The Beijing clone is of particular concern as it is typically highly drug resistant and it is supposed to have spread from East-Asia into Eastern-Europe (21,22).

In addition, these studies indicated that the MTBC had a highly clonal population structure (18) and that there was an absence of inter-genomic recombination occurring within the complex. MTBC is devoid of horizontal gene transfer, thereby exhibiting a closed genome, coupled with a low mutation rate (23). Consequently, is recognized as monomorphic bacteria, but, still, a successful pathogen that has subsisted as such since the dawn of humankind (19,24). *M. canettii* is an exception, as there is some evidence of recombination both within this species and with other members (12). The absence of recombination in the rest of the complex is currently unexplained, but could possibly be due to a loss of the required molecular mechanisms or a lack of opportunity due to its facultative intracellular lifestyle.

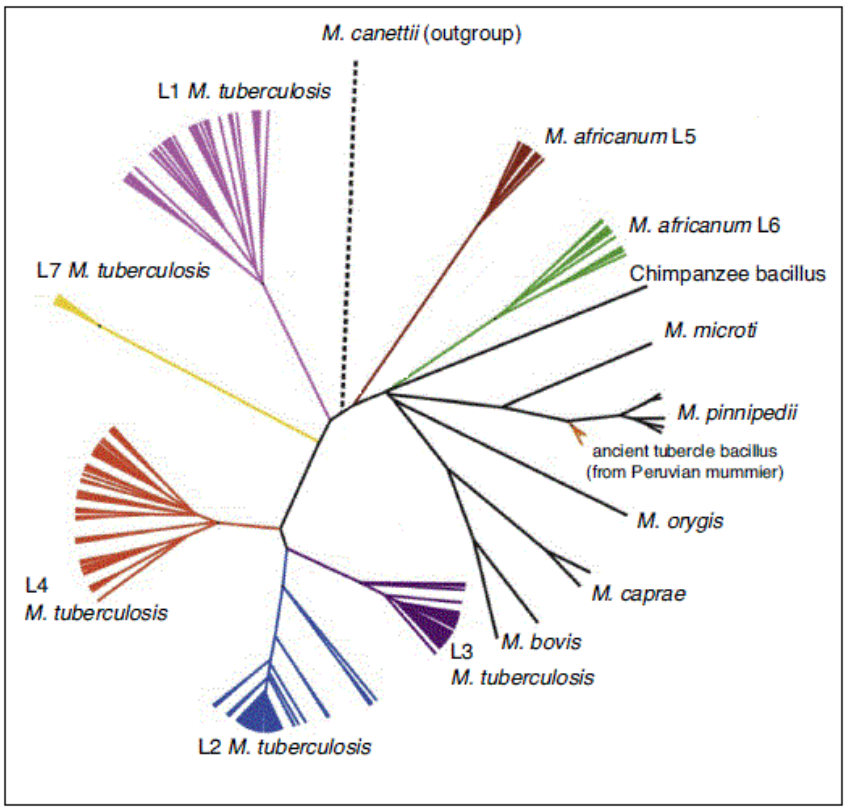


Figure 1.1. Phylogenetic relationships of the different MTBC lineages, based on SNPs of 261 mycobacterial genomes, adapted from the work from (25) - L1 correspond to East-African-Indian (EAI) strains, L2 to Beijing strains, L3 to Dehli/Central Asian strains (Dehli/CAS), L4 to Euro-American strains (T, Haarlem, LAM, S, X), L5 to *M. africanum* 1 and L6 to *M. africanum* 2 strains. L7 strains are restricted to Ethiopia and the Horn of Africa region (26).

1.3 History of Tuberculosis

Tuberculosis is considered an ancient disease, and evidence for TB-like disease, confirmed by both morphological and molecular methods, has been found in skeletons dating to the Neolithic era, approximately 9,000 years ago, in the Eastern Mediterranean (27). However, some estimates place the origin of the disease much earlier – 70,000 years old – when humans first started emerging from Africa (19).

TB is thought to have killed more people than any other microbial disease throughout history (28). Its significant impact on human society is reflected by the multiple designations that were attributed to this disease throughout the centuries. Hippocrates first described it as "consumption" (or Phthisis in Greek), probably relating to the "wasting away" and weight loss experienced by the patients (29). The term "White Plague" was used during the epidemics that spread throughout Europe during the 17th and 18th centuries (30), and presumably referred to the pale complexion given by the disease. TB incidence is thought to have reached its peak in the 19th century when it is estimated that a quarter of Europeans have died from the disease (29). It is against this catastrophic scenario and prognosis that Robert Koch made his famous presentation to the Physiological society of Berlin in 1882, where he demonstrated that the tubercle was the causative agent of TB. Not only was this one of the first pathogenic bacteria to be described, but he also established the "Koch's postulates", which set the standard of infectious diseases' etiology, still of relevance nowadays (28).

With the advent of antibiotics and improved public health measures, many in the western world have considered TB a disease of the past. Incidence declined gradually during the early and mid-19th century almost until the present day, although the exact reasons for this phenomenon remain unclear (28). Despite this, a third of the population is thought to be infected, and today, TB remains a disease of poverty in high and low/middle income countries, with the global burden mostly centralized in Africa, Asia and South America, where the majority of the infected people can be found (31,32).

1.4 Tuberculosis: pathogenesis, clinical features and diagnosis

1.4.1 Pathogenesis

Tuberculosis can develop through progression of recently acquired infection (primary disease), reactivation of latent infection, or exogenous reinfection (33). In immunocompetent individuals,

about 90% of those with TB infection never develop the disease; approximately 3-10% will develop the disease in the first 1-2 years after infection (34) and another 5% during their lifetime. The risk depends on the age of acquiring infection, being lowest in the age range of 5 to 9 years (35). Exogenous reinfection is thought to be uncommon in immunocompetent people, but life-style-related factors and chronic diseases, such as active or passive smoking (36,37), nutritional status (38) and diabetes mellitus (39) may significantly affect the risk of reactivation. In the setting of Human-Immunodeficiency Virus, HIV-1, infection, the risk of progressing rapidly to disease, once infected with *M. tuberculosis*, the risk of reactivation, and the risk of exogenous reinfection are all increased compared to seronegative persons (40–42).

The "life cycle of TB" starts with the inhalation of infectious droplets that reach the alveoli (Figure 1.2). They are phagocytised by the alveolar macrophages and, at this point, the immune system either manages to confine the mycobacteria, leading to a latent asymptomatic infection and the formation of granulomas, which happens for the majority of the cases, or failure can lead to an active infection (43,44). In order to control the infection, the macrophages induce production of proteolytic enzymes and cytokines that attract T lymphocytes to the site. This initial control phase can last between two to 12 weeks (45). If this is successful, a granuloma will eventually be formed, which is a nodular type lesion formed of T lymphocytes and macrophages intended to confine the mycobacteria. This primary pulmonary granuloma and associated draining lymph nodes are known as the "Ghon complex" and can be detected radiologically (46). This environment is characterised by low oxygen and pH, in which the mycobacteria are able to survive in a dormant state. The lesion can then undergo calcification and fibrosis in order to keep the infection confined. Approximately 90% of those infected with *M. tuberculosis* maintain the infection in this dormant state for the rest of their lives (47). Of the remaining 10%, the granuloma fails to contain the bacilli allowing them to spread to a bronchus or nearby blood vessel (45). This allows the infection to spread throughout the respiratory system where progressive lung damage occurs through the formation of cavities. In some cases it spreads to other organs such as the lymphatic system, bones and meninges and a reactivated TB can affect almost any anatomical site (48). However, only pulmonary TB is transmissible and ensures the evolutionary success and adaptation of the bacteria.

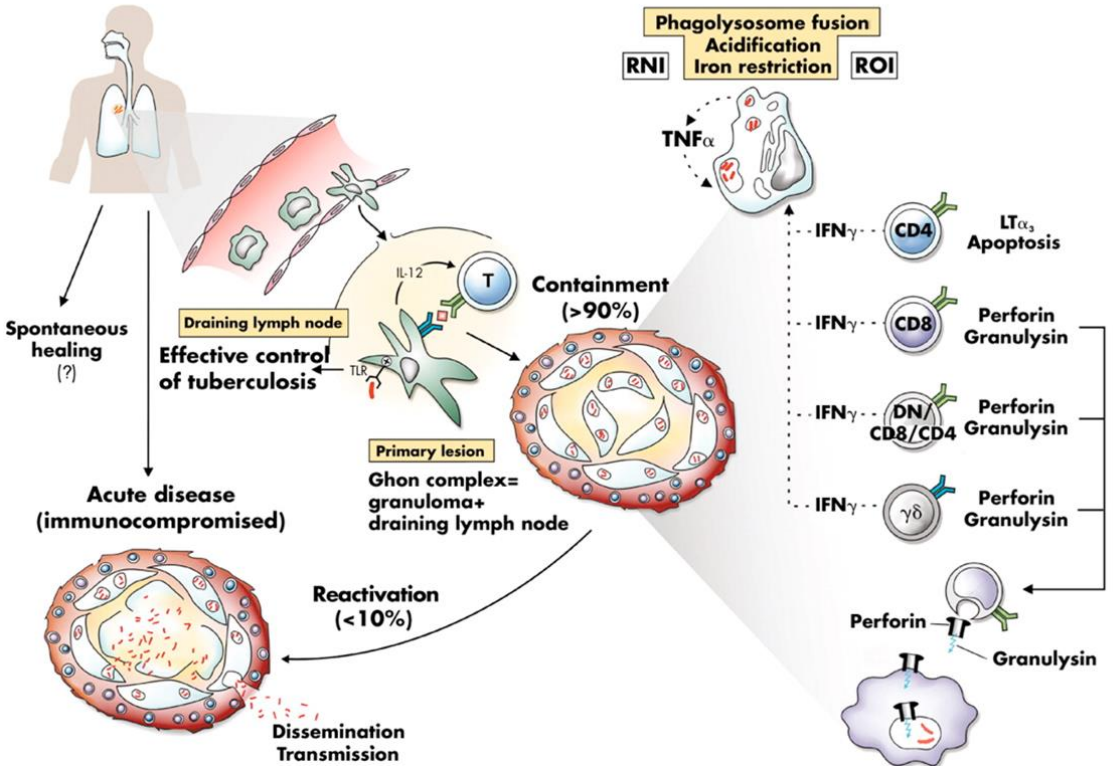


Figure 1.2. Phases of human tuberculosis. After inhalation of the bacteria, there is a blood-borne stage where the immune system attempts to control the infection. In 5-10% of individuals, this will lead to active or cavitary tuberculosis, which can allow *M. tuberculosis* ongoing transmission through aerosols production. Adapted from (44).

1.4.2 Clinical features

Most TB cases occur as pulmonary disease with only about 17% occurring at an extrapulmonary site. However, about 70% of HIV-1 infected patients will have evidence of extrapulmonary disease or mycobacteremia and these co-infected patients are more likely to present atypically, potentially delaying TB diagnosis (1).

The classical symptoms of early and progressive active TB disease can be unspecific but the most common are fevers, night sweats, fatigue, weight loss and a chronic cough (45,49). Additionally, the disease reveals localized symptoms according to the form it takes. If untreated, the estimated fatality rate of smear positive and negative pulmonary TB is about 70% and 20%, respectively (50). Some extrapulmonary manifestations, such as milliary TB or meningitis are universally fatal in the absence of treatment (49,51). Patients with pulmonary disease, of whom, those with cavities that function as open reservoirs of large numbers of bacilli, are the most infectious and more prone to transmit the disease (52). This happens when droplets are coughed

up from the bronchus, aerosolised, and remain airborne for minutes to hours allowing spread to other persons. These droplet nuclei are tiny ranging from 2–5 μm in diameter and containing as few as 1–3 cells (53). The timing of the development of active TB can vary greatly from weeks to decades after infection, is most often caused by a compromised immune system, and can later become latent, and then be reactivated multiple times throughout life.

1.4.3 Diagnosis

Although TB shares many clinical and radiological features with other respiratory diseases, symptoms remain an important tool to facilitate passive case-finding (54). Radiology (X-ray) is generally used in a first evaluation for TB diagnosis. In patients with progressive primary or postprimary TB, computed tomography scanning is often performed, in addition to chest radiography. Magnetic resonance imaging may also be used to evaluate complications of thoracic disease, such as the extent of thoracic wall involvement, but is of limited value in the evaluation of patients with pulmonary TB. These chest abnormalities are merely suggestive and can be used to rule out the disease but not for confirmative diagnostic purposes (55,56).

Patients suspected of having pulmonary TB should have at least two sputum specimens examined for microscopic evidence of acid-fast bacilli (AFB) (57). Fluorescence microscopy is faster to visualize and shows a higher sensitivity when compared to conventional Ziehl-Neelsen (ZN) smear microscopy (58). However, it has lower specificity than ZN smear microscopy for diagnosis, thereby suggesting a need for appropriate training, quality management, monitoring of performance and confirmation testing with ZN staining (59). In addition, specimens should be cultured in order to identify the specific mycobacterial species and for drug susceptibility testing. Mycobacteria of the MTBC are slow growers, dividing every 15 to 20 hours (60), and thus it can take over 3 weeks to see visible growth on standard culture media (61). This has significant clinical implications because therapy must be initiated before TB diagnosis is confirmed. In addition, drug susceptibility testing (DST) results are rarely available at the initiation of treatment. On the other hand, clinically significant disease can be present even in the absence of a positive culture, due to poor specimen collection, contamination and/or low bacillary count (54).

Nucleic acid amplification tests (NAAT) allow the rapid identification of MTBC (61). Several NAAT are commercially available for the laboratory-based diagnosis of TB and some of them also allow the detection of resistance to rifampicin and some other first- and second-line drugs (54). Using culture as the reference procedure, most of them reveal similar high sensitivity and specificity

for the detection of *M. tuberculosis* among sputum smear-positive patients (62,63). However, unlike culture isolation, detection by NAAT does not necessarily imply viability of the detected bacteria. Nevertheless, the WHO recommends that Xpert MTB/RIF should be used rather than conventional microscopy, culture and DST as the initial diagnostic test in adults and children suspected of having MDR-TB or HIV-associated TB (64).

1.4.4 Drug susceptibility testing

Drug susceptibility testing (DST) of *M. tuberculosis* should be performed on an initial isolate from all patients with TB (61). If the patient's culture remains positive after 3 months of therapy, a new DST should be performed. Testing is done using a standard methodology such as the recommended by the National Committee for Clinical Laboratory Standards (NCCLS) (65).

Drug resistance can be detected by a variety of *in vitro* methods that are usually contingent on demonstrating growth of the organism in the presence of a "critical" concentration of an antituberculosis drug. The two most commonly used qualitative methods are the proportion method and the BACTEC method. The agar proportion method has been proposed as the reference method for all antituberculosis drugs except pyrazinamide for which BACTEC is the reference method (65). With the proportion method, plates of drug-free agar and agar containing critical concentrations of antituberculosis drugs are inoculated with the isolate. For most drugs, resistance is determined by comparing the number of colony forming units (CFU) on drug-containing *versus* drug-free media, with clinically significant resistance being defined as greater than 1% growth on drug-containing media relative to drug-free media (65). Rapid broth-based methods (e.g., BACTEC, MIGIT, etc.) are recommended for initial susceptibility testing of first-line agents. The BACTEC method allows a more rapidly determination of minimal inhibitory concentrations (MICs), as the growth is facilitated by the addition of enhancing growth supplements.

The short turnover time of genotypic methods offers an attractive potential for rapid detection and characterization of drug resistance. However, genome sequencing of *M. tuberculosis* isolates revealed a large number of new genes, intergenic regions and nonsynonymous single nucleotide polymorphisms (SNP) showing consistent associations with drug resistance. This indicates that the genetic basis of drug resistance is more complex than previously anticipated (66) and that the available commercially and routinely used NAAT do not target all the necessary genes.

1.5 Tuberculosis treatment

Although TB is an ancient disease, effective drugs were not available for centuries. The pre-antibiotic therapy initially consisted of isolation of the patients in sanatoria to reduce the transmission to healthy contacts, with rest, adequate nutrition, and sunlight exposure (67–69).
The first evidence of a potential anti-TB drug was made in 1940, when a dapsone-derivative compound, known as promin, was administered to a sample of infected guinea pigs. However, that compound was never subjected to human clinical trials (70–72). In 1944, Schatz and Waksman stated that streptomycin (STR), a natural substance isolated from *Streptomyces griseus*, had bactericidal activity and thus could be prescribed for TB treatment. However, only few years later, the first resistant cases arose, compromising the use of a streptomycin-based monotherapy (73). Four years later, a new synthetic drug, called para-aminosalicylic acid (PAS), was presented as an alternative drug. Following the poor results of the monotherapy, in 1952, the first regimen based on the combination of STR, PAS, and isoniazid (INH) was proposed (71,72,74).
In 1954, pyrazinamide (PZA) was discovered and ethambutol (EMB) and rifampicin (RIF) were introduced in 1961 and 1963, respectively. At this time, the duration of therapies could last two years. In 1970, trials on RIF-including regimens showed good results with a therapy of 9 months, and in 1974, the inclusion of RIF and PZA at lower dosages demonstrated the efficacy of a 6-month treatment (71,72,75,76).
The choice of the antituberculosis drugs in the different phases is not random, but it is based on the epidemiology and on the specificity of action of the drugs, which are molecules with two different mechanisms of action—bactericidal and sterilizing effect (77). The first group is crucial in the intensive phase and allows a relevant reduction of the bacterial load; the indirect consequence of this activity is the reduction of the probability of selecting drug-resistant strains. The most important drugs prescribed for that aim are INH, PZA, RIF, and STR. The sterilizing activity is performed mostly in the continuation phase because it is oriented to kill mycobacteria in a dormancy state. Antituberculosis drugs of this group are PZA and RIF (77). These general principles are accepted worldwide and the WHO defined standardized regimens (78). As such, new cases of TB in patients that were never exposed to drugs have to be treated for 6 months. The intensive phase lasts two months and should be administered a combined regimen that includes EMB, INH, PZA, and RIF. The four-months continuation phase only includes INH and RIF (77,78). Previously treated cases require a different management and a rapid and conventional DST is required before the initiation of therapy, to ensure the most appropriate regimen to choose (77).

It is obvious that multidrug resistant cases (MDR, i.e., *in vitro* resistant to at least isoniazid and rifampicin) could represent a challenge because of the poorest therapeutic options. The so-called second- and third-line antituberculosis drugs are less efficacious, more toxic, and more expensive than the first-line drugs. However, and because of the lesser effectiveness, to obtain a clinical and a microbiological cure it is mandatory to treat individuals with MDR-TB for longer periods. The WHO suggests the prescription of at least four active drugs during the intensive phase and should include PZA, one of the injectable second-line drugs (amikacin - AMK, capreomycin - CAP, or kanamycin - KAN), a new-generation fluoroquinolone (FQ), ethionamide - ETH (or prothionamide), and cycloserine (or PAS) (78,79). The duration of the first phase of the treatment should depend on the culture conversion, but it should last at least eight months, whereas the duration of the second phase should be longer than 20 months (77–79).

With the emergence of more resistant cases, first described in 2007 in patients from Africa, new therapeutic options have been proposed (80). Extensively-drug resistant (XDR) TB (defined as MDR-TB with resistance to a FQ plus a second-line injectable drug) is a considerable threat, resulting in extremely poor treatment outcomes. In a recent study of XDR-TB in South Africa, 46% of patients died after a two-year follow-up (81); the same outcome would be expected without treatment at all. As such, there is a desperately need for new or improved drugs to prevent further resistance (eventually leading to strains resistant to all drugs) (82). For this reason, several drugs approved for infectious diseases other than TB were screened and showed *in vitro* and *in vivo* antimycobacterial activity; among them, imipenem-cilastatin, linezolid, and meropenem-clavulanate have had a relevant role in individuals with drug-resistant TB in the last few years. The new molecules recently approved or in the last clinical trial phases are bedaquiline (a new diarylquinoline, previously called TMC 207), delamanid (previously called OPC-67683), sutezolid (PNU 100480), and PA-824 (79,83).

1.6 Resistance to antituberculosis drugs

In *M. tuberculosis*, drug resistance occurs through chromosomal mutations that confer resistance to individual antituberculosis drugs. These mutations occur spontaneously and at predictable rates (84). For example, mutations conferring resistance to INH and RIF occur with an estimated frequency of approximately 3×10^{-8} and 2×10^{-10} mutations *per* bacterium *per* generation, respectively (84). All populations of *M. tuberculosis* will therefore have a certain number of naturally occurring drug-resistant mutants and this probability will be influenced by

the size of the bacterial population and the replication rate. The probability that simultaneous resistance to INH and RIF will develop in nature is then extremely small, as it is the mathematical product of each of the separate probabilities (85). The process of development of resistance in *M. tuberculosis* is basically through the selection of *de novo* mutations, either SNPs or indels, at *loci* usually termed as resistance associated genes. There are basically four mechanisms responsible for this: i) drug target modification, as a result of non-synonymous mutations; ii) unsuccessful prodrug activation, due to mutations that prevent the prodrug of reaching its active form; iii) target overexpression, usually from mutations at the promoter region that controls the expression of the drug target; and, iv) overexpression of drug modifying enzymes, rendering the drug inactive (86).

STR interferes with protein synthesis by inhibiting genetic translation (87). Minimum Inhibitory Concentration (MIC) range between 1.0-2.0 mg/L (1.0 mg/L for *M. tuberculosis* H37Rv) and this drug has a moderate bactericidal activity against susceptible isolates (88). Comparing with the two other aminoglycosides used in TB treatment, KAN and AMK, STR is the least toxic (89). Resistance is usually due to mutations in *rpsL* and are associated with high-level resistance, in particular, the K43R mutation (90). Another mechanism of STR resistance occurs through *rrs* gene mutations and usually yields a lower resistance level (90,91). *gidB* mutations have also been detected in clinical isolates, although its role in resistance is not yet fully understood. They appear in resistant and susceptible isolates (92,93), suggesting they could be phylogeny-related (93). This is the case for the endemic MDR/XDR-TB Q1 clade in Portugal that was defined based on the A80P mutation on *gidB*, which is simultaneously associated with an intermediate-level resistance to STR (94,95).

INH is a synthetic prodrug that requires activation by the bacterial catalase peroxidase encoded by the *katG* gene and enters the cell by passive diffusion (96,97). Its efficacy is in part due to its low MIC: 0.02 mg/L for *M. tuberculosis* H37Rv and 0.02-0.05 mg/L in susceptible clinical isolates (88). INH has a bactericidal activity against rapidly growing mycobacteria and is bacteriostatic against slow-growers, although bactericidal activity is also observed in *M. tuberculosis* (98). Resistance usually develops as a consequence of *katG* mutations that decrease the ability of the catalase-peroxidase to convert INH to its active form. The most common mutation is a serine to threonine substitution at codon 315 (S315T), found in up to 93% of INH resistant isolates and is associated to high-level resistance (99–101). Another important mechanism of INH resistance, is the acquisition of mutations in the promoter region of the *mabA(fabG1)-inhA* operon (102–104). These mutations usually lead to INH low-level resistance, and an unusual high-prevalence

of *inhA* promoter mutations (up to 91%) have been detected in Lisbon, Portugal, and in strains from *M. africanum* West-Africa 1 lineage (105,106).

RIF is a semi-synthetic drug derived from rifamycin (107) and binds to the *β*-subunit of the DNA-dependent RNA polymerase (encoded by *rpoB*) (108), physically blocking transcription (109). It has a bactericidal activity (110) against metabolically active bacteria, but also possesses some sterilizing activity against latent bacilli (111). RIF MIC ranges between 0.2-0.4 mg/L for susceptible clinical isolates (0.4 mg/L for *M. tuberculosis* H37Rv) (88) and the acquisition of resistance is usually the result of aminoacid substitutions in an 81-bp region of *rpoB*, named RIF resistance determining region (RRDR) (112,113). Besides aminoacid substitutions, deletions or insertions in *rpoB* have been reported in some studies (114). The most common substitutions occur in codons 450 (prevalence of 31.0-76.9% in RIF-resistant isolates), 445 (7.7-43.0%) and 435 (3.4-28.6%), according to *M. tuberculosis* RpoB numbering (92,112,113,115–119). RIF resistance level depends on the mutation, *e.g.*, S450L and H445D result in high-level resistance whereas D435V mostly results in an intermediate-level resistance (91,120,121). It is of most importance to note that resistance to RIF rarely emerges before resistance to other drugs, especially INH, and, for this reason, is considered a marker for prediction of MDR-TB cases (122).

EMB is an antimycobacterial drug synthesized from ethylenediamine and targets the cell wall biosynthesis by inhibiting the arabinosylation of the cell wall arabinogalactan (AG) and lipoarabinomannan (LAM) (123). It has a bacteriostatic activity against metabolically active bacilli (124,125) and MIC range of 0.5-2 mg/L for susceptible isolates (0.5 mg/L for *M. tuberculosis* H37Rv) (88). EMB resistance is mainly associated with mutations in *embB* (126) and the most common ones occur in codon 306, usually involving the substitution of a methionine by a valine, leucine or isoleucine (103,127–129). Although strains bearing *embB*306 mutations have been associated with a higher level of resistance (130), the molecular basis of EMB resistance is not fully determined, as mutations at this site have been described to appear in both resistant and sensitive strains (91,103,129).

PZA is a nicotinamide synthetic prodrug that enters the cell by passive diffusion where it is converted by the bacterial pyrazinamidase (PZase)/nicotinamidase into pyrazinoic acid (POA) in an acidic pH environment (131). However, there is recent evidence showing that PZA can also act on neutral pH conditions (132). Concerning resistance in clinical isolates, the main mechanism supporting PZA resistance is the acquisition of mutations in *pncA*, which have been identified by numerous other authors in about 72.0-99.9% of the PZA resistant isolates studied (92,131,133,134).

The second-line injectable drugs (SLIDs) for TB treatment are KAN, AMK and CAP. Although KAN and AMK are, such as STR, aminoglycosides and CAP is a macrocyclic peptide, these drugs share the same action mechanism and cross-resistance between the three has been well documented (135,136). KAN, AMK and CAP have shown *in vitro* bactericidal activities but CAP has also a bactericidal effect against latent *M. tuberculosis* bacilli (137). These three drugs act through the inhibition of the genetic translation due to ribosomal binding. Mutations in the 16S rRNA-encoding *rrs* gene can mediate cross-resistance between CAP, KAN and AMK, being the most common mutation the A1401G, present in 49.3-88.6% of the resistant isolates (91,103,136). On the other hand, the C1402T mutation, mediates resistance to KAN, CAP, but not to AMK whereas the G1484T mutation also mediates resistance to the three drugs (KAN, AMK and CAP) (136,138). KAN resistance has also been linked with *eis* overexpression, which is responsible only for low-level KAN resistance (139). CAP resistance can also be mediated by *tlyA* mutations (140). Nevertheless, *tlyA* mutations in CAP resistant isolates are rare and it is more likely that CAP resistance develops because of KAN/AMK cross-resistance (140).

FQs are quinolones fluorinated at the central ring system. These are broad-spectrum antibacterial drugs that target the bacterial DNA gyrase and topoisomerase IV, therefore inhibiting DNA replication (141). FQ resistance has been reported to be increasing as a result of previous exposure to FQ prior to TB diagnosis and treatment as it is commonly used for treatment to other bacterial infections (79). Gatifloxacin (GAT) and moxifloxacin (MXF) appear to induce FQ resistant mutants at a lower rate than ciprofloxacin (CIP) and levofloxacin (LVX) (142). Furthermore GAT and MXF are more effective (lower MIC) than CIP and ofloxacin (OFX) and can be used to treat FQ low-level resistant isolates (143). Von Groll *et al* has also observed almost complete cross-resistance between OFX, MXF and GAT (144). The molecular basis of FQ resistance has been associated with *gyrA* and *gyrB* mutations (145). The most common mutations associated with FQ resistance occur in *gyrA* in codons 94 and 90, in 46.2-71.9% and 4.0-43.0% isolates, respectively (103,143,144,146). Mutations occurring in *gyrB* have also been described, although at a lesser frequency and some with questionable role in FQ resistance (142,146).

The molecular mechanisms responsible for the resistance to the remaining second and third-line anti-TB drugs, e.g. linezolide, PAS and cycloserine, are still poorer investigated and, as such, the correlation between genotypic and phenotypic DST has low sensitivity.

1.7 Epidemiology of tuberculosis

The WHO declared TB as a global public health emergency in 1993 and, at the present, almost three decades after, TB still accounts for almost 2 million deaths each year, making it the ninth leading cause of death worldwide (147,148).

The major obstacle for an effective TB control is, undoubtedly, the antimicrobial resistance (149). The new WHO's End TB framework is aiming towards TB elimination by 2035 but, to be successful, this new strategy must effectively address increasingly different challenges (148). One of them regards the use and the development of new anti-TB drugs and efficient vaccines. The other refers to the implementation of TB diagnostic technologies that could translate into proper care, early treatment initiation and blocking of the transmission chains.

In 2017, 10 million people fell ill with TB, and 1.6 million died from the disease (including 0.3 million among people with HIV) (31). WHO estimates that there were 558 000 new RIF-resistant cases, of which 82% were also MDR-TB. According to this, about 500 000 new MDR-TB cases have occurred in 2017 (31), which, accordingly to the latest estimates of 2014, will likely result in the death of approximately 78 400 of these patients (about 16%). XDR-TB associated mortality and treatment success are even lower than those for MDR-TB: 28% and 30%, respectively (148). Across the WHO European region alone, 15 363 (17.7%) new MDR-TB cases were reported in 2017 (150). However, only approximately 40% of all RIF-resistant and MDR-TB cases are currently being detected in Europe, due to the lack of universal DST coverage or rapid testing, and, as such, the increase of primary MDR-TB transmission is being potentiated (150). Although it is estimated that 54 million lives were saved through TB diagnosis and treatment between 2000 and 2017, the detection rate is still very far from the WHO established target of 85%, making the deployment of adequate molecular testing an urgent matter and the knowledge of the resistance targets a pressing need (31).

In Portugal, TB incidence has been steadily decreasing in the last years, with an average of about 5% per year and the proportion of patients with MDR-TB remained stationary with 1% of the cases (151). In 2017, the report of the Portuguese National TB program (NTP) reported 1607 new cases of pulmonary TB with 12 MDR-TB cases (151).

Considering that M/XDR-TB are more difficult and expensive to treat compared to drug sensitive TB, survival rates are poorer, and knowing that primary transmission is the major cause for this epidemic, controlling these cases is the key for successful TB control programs and for achieving the targets of the "End-TB Strategy" (148).

1.8 Molecular typing of *M. tuberculosis* strains

The development of molecular techniques to differentiate strains within a species has been useful both as a public health and as a research tool. Over the last decades, several techniques have been developed taking advantage of the most variable *loci* in the *M. tuberculosis* genome. The first typing method developed was based on restriction fragment length polymorphism (RFLP) analysis using the insertion sequence element IS*6110* as a probe (152). Another commonly used technique, spoligotyping, targets specific repeat sequences found in multiple copies at a single locus in the *M. tuberculosis* genome (the direct repeat *locus*) using a DNA probe (153). Variable number of tandem repeat (VNTR) typing is the most recently developed method, based on the presence and number of mycobacterial interspersed repeat *loci* (MIRU) and is currently recognised as the gold standard (154–156). The three methods vary in their reliability, resolution and time-on-hands, but there is no clear winner with different laboratories across the world preferring either one method or employing all of them at once.

Despite the undeniable usefulness of these genotyping techniques, it is equally undeniable that they have their limitations. By design, these *loci* are at the extremes of variation so are unrepresentative of the genome as a whole. As such, they can only provide us with a basic idea of relatedness, and are difficult to resolve with temporal information. Their major issue for public health applications is that they can also lack discriminatory power as isolates with identical DNA fingerprints may not always be epidemiologically linked (157).

1.9 Whole genome sequencing

1.9.1. Illumina Sequencing technology

In the last decade there have been major advances in the so called “next generation” sequencing technologies in order to allow sequencing to become more high throughput and affordable. There several technologies currently available, but the Illumina platforms currently dominate the high-throughput sequencing market, and have been used for all the sequencing analysis carried out for this thesis, so will be discussed in more detail.

Illumina sequencing is similar to the Sanger method in that it is based on a sequencing-by-synthesis approach, where a polymerase is used to synthesise a complementary strand to the single stranded target DNA with terminator nucleotides used to halt the synthesis. However, the Illumina technology utilises reversible terminators so that the chain termination process is not permanent and synthesis can continue after each base is detected. Fluorescently tagged nucleotides are used to determine which base is being incorporated as the synthesis proceeds one base at a time. In order to achieve this, "libraries" of the target DNA need to be prepared. First, the genomic DNA is fragmented with an aim to produce lots of overlapping fragments within a specific size range (for bacterial genomes this is most often 250-500bp). Adaptors are attached to the fragments, which serve four functions: ligation to the flowcell, as primers for PCR amplification, sequencing primer-binding sites and as index tags to allow multiplexing of multiple libraries in a single run. After adaptor ligation, a PCR step is then typically used to enrich for DNA fragments with the adaptors in the correct orientation. The DNA is then denatured to produce single strands, which are then ligated to a flowcell, where each fragment is amplified to form clusters of clonal DNA, which will increase the intensity of the fluorescent signal. The sequencing reaction is carried out with modified versions of the four nucleotides (dATP, dGTP, dCTP, dTTP) with a different fluorescent dye and blocking group. When a base is incorporated as complementary to the template strand, the fluorescent dye is photographed and then removed. This allows the sequence of the millions of DNA fragments to be determined at once, one base at a time (158).

1.9.2. Application of whole genome sequencing (WGS) to *M. tuberculosis*

It took 13 years and 3.8 billion dollars (159) for the completion of the Human Genome Project in 2003 (160). Since then, advances in sequencing technology have made it possible to sequence an entire human genome in a few days and costing a few thousand dollars. As impressive as this is, bacterial genomes are megabases long as opposed to gigabases and WGS is starting to become a clinical reality for infection diagnosis. First, its higher resolution compared with the traditional genotyping technologies allows academic researchers to better understand bacterial population structure, mutation rates and evolutionary processes. Using genome wide SNP as the basis for these kinds of studies means we can start to understand temporal parameters, as these units of variation are likely to be more clock-like associated than those studied using traditional genotyping techniques. The second major advantage of WGS is that it can provide information on variants other than SNP. Both mapping and *de novo* assembly approaches can

be used to detect deletions, insertions and the acquisition of horizontally transferred elements. This allows us to understand the evolutionary dynamics of these elements and how they affect pathogenicity and antibiotic resistance. Finally WGS is becoming cheaper, meaning that irrespective of all the other advantages it provides, it can equally be, if not more, economically viable than current techniques.

The first WGS study of a *M. tuberculosis* outbreak was published in 2011 (157), and described a putative outbreak involving 32 people in British Columbia, Canada. All the isolates from the outbreak were identical using MIRU-VNTR, but could be differentiated using the genome-wide data. The resultant phylogeny was combined with detailed epidemiological contact tracing, which together suggested that what was thought to be a single outbreak consisted of two different outbreaks. Similar studies on other tuberculosis outbreaks (161) or recent transmission (162,163) have provided further evidence to support the advantages of genome-analysis in determining transmission patterns over previous techniques.

WGS has also enabled significant progress in our understanding of antibiotic resistance. Work by Sebastian Gagneux and colleagues provided the first convincing evidence for the existence of compensatory mutations in TB (164). Furthermore our knowledge of possible drug resistance causing mutations have also been expanded either through the identification of convergent mutations (165), or through more complex methods that identify evidence of diversifying selection in both genic and intergenic regions associated with drug resistant strains (66).

More recently, advances have been made to achieve WGS results directly from clinical samples (166,167) making it possible to analyse *M. tuberculosis* strains' epidemiology and genetic resistance markers without the constraints of the isolation in culture. This may provide faster results and the improvement of public health measures and proper therapeutic schemas.

1.10 Aims and general research plan

With this PhD dissertation, we intended to demonstrate the usefulness of WGS-based methodologies in the diagnosis and monitoring of TB through the genomic analyses of *M. tuberculosis* strains isolated from patients with MDR-TB. The major question that drove this main objective was to address if we could use, or replace, some of the phenotypic testing with procedures based on WGS-sequencing that could allow more discriminatory and faster results to be sent to the clinician as well as to the public health authorities. As such, several specific aims were pursued in the course of this PhD work:

i) to evaluate the use of specific and tailored platforms for the *in silico* prediction of known mutations associated with the resistance to anti-TB drugs using WGS-based methodologies after *M.tuberculosis* isolation in culture (Chapter II);

ii) to evaluate the clustering rates and association with the epidemiological links using the traditional genotyping method, MIRU-VNTR, throughout the last five years, and after the establishment of specific reference centers for the managing of all the MDR-TB cases diagnosed in Portugal (Chapter III);

iii) to establish a WGS-based molecular surveillance methodology, in order to infer a more reliable association between the strains to be signalized as possible transmission cases to the health authorities (Chapter IV);

iv) to surpass the time and constraints regarding the isolation of *M. tuberculosis* in culture, by applying the above defined WGS-based methodologies directly to sputum samples isolated from the TB patients. This would speed up even more the diagnostic/DST to be sent to the clinician and public health authorities (Chapter V);

v) to contribute to a better knowledge regarding the mutations associated with the acquisition of phenotypic resistance, by using antibiotic selective pressure on fully susceptible *M. tuberculosis* strains to monitor the development of resistance. This would enrich the publicly available databases thus contributing to the improvement of the bioinformatics platforms evaluated in Chapter II (Chapter VI).

Chapter II

Genetic prediction of antibiotic resistance

Dissecting whole-genome sequencing-based online tools for predicting resistance in *Mycobacterium tuberculosis*: can we use them for clinical decision guidance?

Manuscript published in
2018, Tuberculosis. 110: 44–51
DOI: 10.1016/j.tube.2018.03.009
Rita Macedo, Alexandra Nunes, Isabel Portugal, Sílvia Duarte, Luís Vieira, João Paulo Gomes. Dissecting whole-genome sequencing-based online tools for predicting resistance in *Mycobacterium tuberculosis*: can we use them for clinical decision guidance?

RM contributed to the design of the study, performed most of the experimental work and part of the bioinformatics analyses, interpreted data and wrote the manuscript.

2. Dissecting whole-genome sequencing-based online tools for predicting resistance in *Mycobacterium tuberculosis*: can we use them for clinical decision guidance?

Abstract

Whole-genome sequencing (WGS)-based bioinformatics platforms for the rapid prediction of resistance will soon be implemented in the Tuberculosis (TB) laboratory, but their accuracy assessment still needs to be strengthened. Here, we fully-sequenced a total of 54 multidrug-resistant (MDR) and five susceptible TB strains and performed, for the first time, a simultaneous evaluation of the major four free online platforms (TB Profiler, PhyResSE, Mykrobe Predictor and TGS-TB).

Overall, the sensitivity of resistance prediction ranged from 84.3% using Mykrobe predictor to 95.2% using TB profiler, while specificity was higher and homogeneous among platforms. TB profiler revealed the best performance robustness (sensitivity, specificity, PPV and NPV above 95%), followed by TGS-TB (all parameters above 90%). We also observed a few discrepancies between phenotype and genotype, where, in some cases, it was possible to pin-point some "candidate" mutations (e.g., in the rpsL promoter region) highlighting the need for their confirmation through mutagenesis assays and potential review of the anti-TB genetic databases. The rampant development of the bioinformatics algorithms and the tremendously reduced time-frame until the clinician may decide for a definitive and most effective treatment will certainly trigger the technological transition where WGS-based bioinformatics platforms could replace phenotypic drug susceptibility testing for TB.

Abbreviations

AMK, Amikacin; CAP, Capreomycin; CICLO, Cicloserine; DST, Drug Susceptibility Testing; EMB, Etambutol; ETH, Ethionamide; FLQ, Fluoroquinolones; IGV, Integrative Genomics Viewer; INH, Isoniazid; KAN, Kanamycin; LNZ, Linezolid; MDR-TB, Multidrug-resistant Tuberculosis; MOX, Moxifloxacin; MTBC, Mycobacterium tuberculosis complex; NPV, Negative Predicted Value; OFX, Ofloxacin; PAS, Para-aminosalicylic acid; PPV, Positive Predicted Value; PZA, Pyrazinamide; RMP, Rifampicin; SNP, Single Nucleotide Polymorphisms; SRA, Sequence Read Archive; STR, Streptomycine; TB, Tuberculosis; WGS, Whole genome sequencing; WHO, World Health Organization; XDR-TB, Extensively drug-resistant Tuberculosis.

Keywords: Multidrug-resistant tuberculosis; Whole-Genome Sequencing; TB profiler; TGS-TB; PhyResSE; Mykrobe predictor.

2.1 Introduction

Tuberculosis (TB) remains one of the leading causes of morbidity and mortality worldwide, with estimates of 10.4 million new cases (312 000 in Europe) and 1.7 million deaths in 2016 (148). Despite millions of people are successfully treated for TB each year, resistance to commonly used antituberculous drugs is increasing (168).

Nowadays, multidrug-resistant TB (MDR-TB; resistant to at least the most powerful first-line anti-TB drugs, rifampicin and isoniazid) and extensively drug-resistant TB (XDR-TB; MDR-TB with additional resistance to any fluoroquinolone and one or more among amikacin, kanamycin or capreomycin) represent a major threat for global TB control. According to World Health Organization (WHO) estimates, there were 490 000 cases of MDR-TB in 2016, of which ~6.2% were also XDR-TB (148). Underlying this continuous multidrug resistance scenario is the mismanagement of TB treatment and person-to-person transmission. Since MDR- and XDR-TB are more difficult and expensive to treat and survival rates are poorer when compared to drug sensitive TB (estimates show that only 54% of the patients with MDR-TB and 30% with XDR-TB survive), the disease needs continuous vigilance (148). As most TB deaths could be prevented with early diagnosis and appropriate treatment, in 2015, the WHO proposed the expansion of rapid laboratory diagnosis as one of the five high priority actions to fight the global drug resistant TB crisis (169). Phenotypic testing of samples collected from possible TB patients, using conventional culture methods, can take up to eight weeks for the isolation of the strains and two to four additional weeks until the final report with the resistance profile is issued (170). The introduction of molecular-based diagnostic tools for the detection of drug resistance enables earlier diagnosis of MDR-TB, preventing the spread of these resistant cases and allowing effective therapeutic options (171). Nowadays, these PCR-based targeted molecular tests, in particular the WHO endorsed Cepheid Xpert MTB/RIF assay (Cepheid, Sunnyvale, CA, USA) and the Hain line-probe assays (Hain Lifescience, Nehren, 64 Germany) (169,171), are generally being used, but only examine a limited number of both *M. tuberculosis* genomic regions and anti-TB drugs, being unable to provide a full "genetic drug resistance" result (170,172). Thus, the fight against M/XDR-TB still lacks more accurate, complete and comprehensive laboratory tests to ensure effective diagnosis with the correct treatment initiation and outcome and, consequently, reduction of TB transmission.

In recent years, whole genome sequencing (WGS) has emerged as a very powerful tool for the surveillance, outbreak investigation and drug resistance monitoring of human pathogens, including *M. tuberculosis*. WGS can detect transmission events missed by epidemiological investigation, discriminate relapse TB from re-infection and it also allows the simultaneous

characterisation of the complete antimicrobial resistance profile of a given isolate (168,173–177), which may be tremendously important for TB patients. Fortunately, this issue is particularly facilitated for *M. tuberculosis* as antibiotic resistance in this bacterial pathogen is known to be provided essentially by single nucleotide polymorphisms (SNPs) and sometimes small insertions or deletions (indels) rather than by the lateral acquisition of resistance genes (178–181). A potential barrier for the implementation of WGS methodologies in a routine diagnostic setting could be the lack of bioinformatics expertise among clinical microbiologists. In fact, the data generated is of great complexity and may hamper its use and full potential, so there is a need for user-friendly comprehensive and validated workflows, before the putative adoption of WGS-based laboratory diagnosis. However, major progress has already been achieved (182,183) as there are several tailor-made automated bioinformatics platforms for rapid prediction of antimicrobial resistance that can be used and implemented in a routine diagnostic setting (184,185,186). The most commonly used bioinformatics platforms are TB Profiler (184), PhyResSE (185), Mykrobe predictor (187) and TGS-TB (186), which are freely available online, user-friendly, can be run in a common desk/laptop computer, and accept raw data directly from the sequencer (FASTQ files). While the existing few studies have demonstrated a high sensitivity and specificity by using such platforms for the first line drugs isoniazid and rifampicin, on the other hand, there seems to exist a substantial variation with the remaining first and second line drugs (173–175,182,188–190). This may be of considerable importance when using WGS to guide clinical decisions so there is an urgent need for standardization of the laboratory methodologies and data interpretation, as previously suggested (191). Although it is now clear that WGS has the potential to revolutionize the detection of drug resistance in MTB strains, it is crucial to demonstrate that its use in the routine diagnostic setting will have an impact on patient outcomes. However, there is a lack of comprehensive and exhaustive studies on the use of these WGS-based analyses in order to guide clinical decision-making. In fact, the available published studies only used one or two of these platforms with a limited number of strains and potential resistance hits (183–185,190,192). Here, in order to contribute to the global picture regarding the usefulness of bioinformatics platforms to predict antimicrobial resistance in TB, we performed a comparative study of the most used freely available online platforms (TB Profiler (184), PhyResSE (185), Mykrobe predictor (187) and TGS-TB (186). We used the largest set of MDR-TB strains to date (enrolling about 300 phenotypic hits) that was simultaneously analysed by multiple platforms.

2.2 Materials and Methods

2.2.1 Samples

For the study purposes, 54 M/XDR- and five susceptible-TB strains were analysed. All MDR-TB strains isolated in Portugal are regularly sent to the TB National Reference Laboratory from the Portuguese National Institute of Health for drug susceptibility testing and genotyping. Drug susceptibility testing (DST) for first- (isoniazid-INH, rifampicin-RMP, etambutol-EMB, streptomycine-STR and pyrazinamide-PZA) and second-line (amikacin-AMK, kanamycin-KAN, capreomycin-CAP, ofloxacin-OFX, moxifloxacin-MOX, ethionamide-ETH, linezolid-LNZ, cicloserine-CICLO and para-aminosalicylic acid-PAS) antibiotics was performed using MGIT960 system (Becton Dickinson), according to manufacturer´s instructions, or the proportion method using solid media (CICLO and PAS). The five susceptible-TB strains were retrieved from Chatterjee et al. 2017 (190), for which susceptibility testing for first- and second-line anti-TB drugs were performed using the same methodology as the one described in the present study.

2.2.2 Whole genome sequencing (WGS)

For each strain, WGS was performed as previously described (193). Total DNA was extracted from solid cultures using commercial extraction kits (QIAmp, Quiagen) after an initial step of cell inactivation where samples were subjected to 95ºC for 1h followed by enzyme digestion for 3h with proteinase K. Quantification and quality assessment of the purified DNA was performed using Qubit Fluorometer with hsDNA Assay Kit (Thermo Fisher Scientific) and agarose gel electrophoresis (0,8%), respectively. High-quality DNA samples were then used to prepare dual-indexed Nextera XT Illumina libraries using the KAPA HiFi HotStart ReadyMix PCR Kit (KAPA Biosystems) in the indexing step to improve amplification of the GC-rich genome regions. Libraries were subsequently subjected to cluster generation and paired-end sequencing (2×250bp) on a MiSeq Illumina platform (Illumina Inc.), according to the manufacturer's instructions.

2.2.3 *In silico* prediction of drug-resistance using online tools

For each strain, genotypic drug resistance was predicted using the freely available software platforms tailor-made for *M. tuberculosis*, namely: TB Profiler webserver v1 (184), PhyResSE v1.0

(185), Mykrobe Predictor v0.1.3 (187) and TGS-TB v2 (186). All of these online platforms analyse WGS data directly from uploaded raw sequence files (in FASTQ format). Basically, they map raw or trimmed (after pre-processing raw data) reads against a modified version of the H37Rv reference genome (GenBank accession number: NC_000962) that contains drug resistance and lineage specific mutations. After SNPs/indels calling using specific algorithms (contrarily to the remaining platforms, Mykrobe predictor algorithm only calls SNPs), the identified mutations are compared to a curated list/database to determine known and putative polymorphisms. Sensitivity, specificity, Positive Predicted Value (PPV) and Negative Predicted Value (NPV) were estimated for each platform in order to access their drug resistance predictability for all anti-TB drugs.

2.2.4 Confirmation of genotypic drug prediction

In parallel, for confirmation purposes we used bioinformatics pipelines established in our lab, in order to access both the accuracy of SNPs/indels calling performed by the above cited platforms as well as all genotypic/phenotypic discrepancies. This was evaluated by a reference-based mapping strategy using Snippy v3.1 software (https://github.com/tseemann/snippy), which included BWA, SAMtools, Freebayes, vcftools and snpEff. FastQC v0.11.3 tool (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) was applied to evaluate reads' quality before and after quality improvement measures carried out by Trimmomatic v0.36 software (http://www.usadellab.org/cms/?page=trimmomatic). Furthermore, as two of the four online tools (TB Profiler and Mykrobe Predictor) used for prediction of TB drug-markers do not pre-process the raw data prior to the analysis, Snippy confirmatory analyses were additionally performed using both raw and trimmed read sequences. For each isolate, after mapping the reads (both raw and trimmed) against the H37Rv reference genome, variants were called in sites that filled the following criteria: i) minimum mapping quality of 20; ii) minimum number of reads covering the variant position >10; and iii) minimum proportion of reads differing from the reference of 90%. The only exception occurred for three variant positions, where a minimum of five reads was accepted to support polymorphisms since they did not occur at sites with unusual low coverage. All putative SNPs/indels were carefully inspected and confirmed using the Integrative Genomics Viewer (IGV) v2.4.2 (http://software.broadinstitute.org/software/igv/).

2.2.5 Data availability

Raw sequence reads of all 54 Portuguese TB strains subjected to WGS were deposited in the Sequence Read Archive (SRA) (http://www.ncbi.nlm.nih.gov/sra) under the study accession number SRP131205 (detailed in Supplementary Table S1). The previously published sequence reads of the five susceptible-TB strains (190) used in the present study for comparative purposes were also downloaded from SRA (Run accession numbers: SRR5341263, SRR5341215, SRR5341214, SRR5341212 and SRR5341211).

2.3 Results

In the present study, we have performed a comparative evaluation of four free most used WGS-based online tools to automatically predict drug resistance for a total of 12 anti-TB drugs (streptomycin, isoniazid, rifampicin, ethambutol, pyrazinamide, amikacin, capreomycin, kanamycin, quinolones, ethionamide, linezolide and para-aminosalicilic acid) for 54 MDR- plus five susceptible-TB strains (the latter were used in order to strength analysis of platforms' specificity). Phenotypic DST was performed throughout the study period as soon as the liquid culture was positive for further analysis. Of the 54 MDR-TB strains, 8 (15%) were also XDR. Globally, a total of 293 phenotypic hits were achieved, which were subjected to genotypic prediction.

Considering that the input of all online tools is the output of WGS apparatus, we started by evaluating the quality of the data used for genotype predictions. This was accessed by using both raw and trimmed reads since two of the four analyzed platforms do not pre-process the raw data prior to the analysis. Despite the variation on the number of paired reads across samples (median 2.1 and 1.8 million of raw and trimmed reads, respectively), the percentage of chromosome covered was homogeneous. On average, 98.8% of raw and 98.6% of trimmed reads mapped to the H37Rv reference genome (ranging across samples from 96.7% to 99.7% for raw reads and from 95.5% to 99.7% for trimmed reads). The median depth of coverage was 88.6-fold using raw reads and 68.6-fold using trimmed reads. The majority of the genome (~96% with raw reads and ~94% with trimmed reads) was covered to at least tenfold depth.

2.3.1 Genotypic/Phenotypic correlation

Figure 2.1. summarizes both the phenotypic and genotypic results obtained per anti-TB drug and platform.

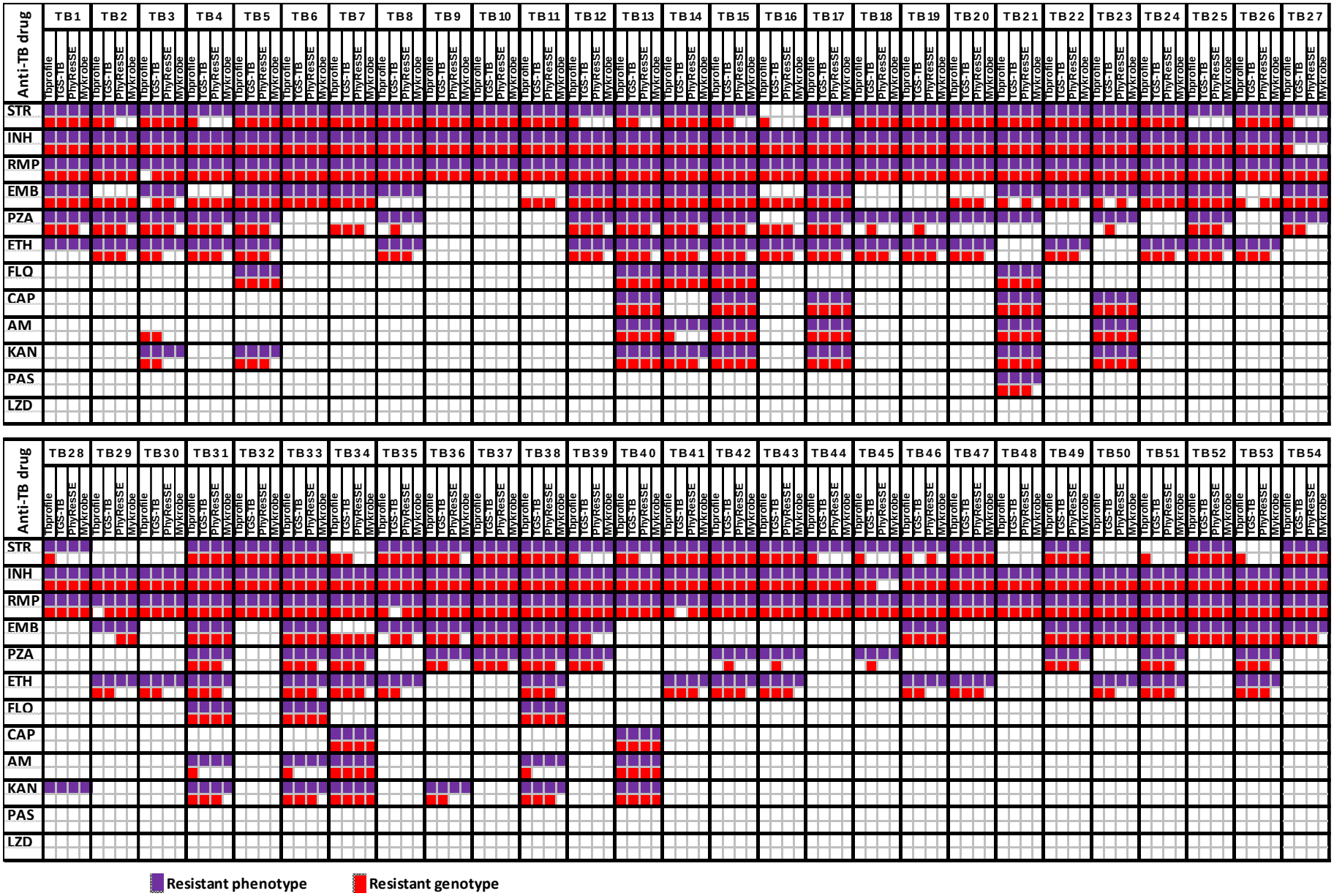


Figure 2.1. Overview of the agreement between phenotypes and genotypes predicted by the four platforms under evaluation. The figure shows the phenotypic results observed for each of the 54 MDR-TB strains enrolled in the present study, as well as the respective genotypic profile given per platform (TB profiler, TGS-TB, PhyResSE and Mykrobe predictor) and per anti-TB drug tested. MDR-TB strains are designated as TB1 to TB54 in the upper row. The anti-TB drugs were: Streptomycin (STR), Isoniazid (INH), Rifampicin (RMP), Ethambutol (EMB), Pyrazinamide (PZA), Ethionamide (ETH), Fluoroquinolones (FLQ), Capreomycin (CAP), Amikacin (AMK), Kanamycin (KAN), Para-aminosalicylic acid (PAS), and Linezolide (LZD). No phenotypic and genotypic hits were observed for LZD.

The outputs were compared and further analysed in order to determine which platform yielded the best correlation with the phenotypic DST. Regarding the five full susceptible TB strains, no false positive results were found with any of the four platforms. Comparative analysis of phenotypic and genotypic drug resistances confirmed the MDR profile of the 54 strains. These phenotypic/genotypic correlation varied per anti-TB drug and platform used for the analysis (Figures 2.1 and 2.2).

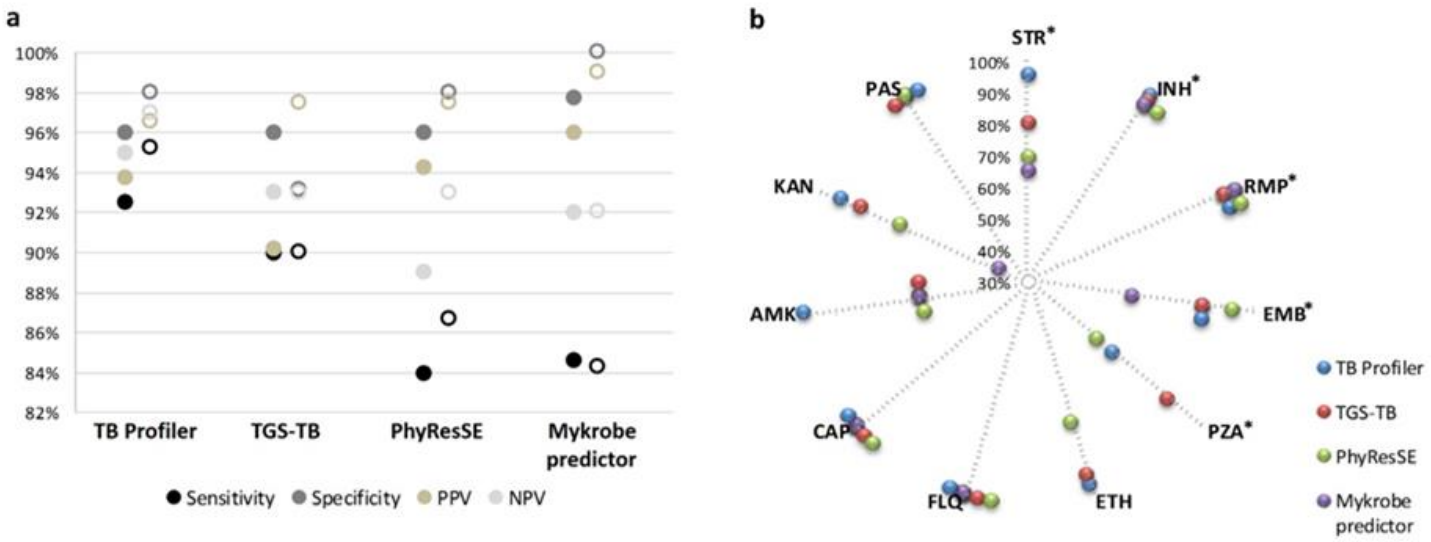


Figure 2.2. Performance values of the bioinformatics platforms for predicting antibiotic resistance. (A) Sensitivity, specificity, Positive Predictive Value (PPV) and Negative Predictive Value (NPV) estimates for each platform. The filled circles correspond to values considering DST as the gold standard, while the unfilled circles are corrected values after detailed analysis of false positives and negatives. All calculations were performed considering the anti-TB drugs that each platform can evaluate, since Mykrobe predictor does not analyze resistance to Ethionamide and Pyrazinamide. (B) Sensitivity of each online platform to predict resistance for each of the tested antibiotics: Streptomycin (STR), Isoniazid (INH), Rifampicin (RMP), Ethambutol (EMB), Pyrazinamide (PZA), Ethionamide (ETH), Fluoroquinolones (FLQ), Capreomycin (CAP), Amikacin (AMK), Kanamycin (KAN), and Para-aminosalicylic acid (PAS). First-line anti-TB drugs are marked with an asterisk (*).

For simplification purposes, we considered only a single genotype for each observed phenotype, even when more than one hit was given by a platform (i.e., when, for some strains, more than one SNP/indel was identified for the same phenotype), leading to a maximum of 293 possible genotype predictions (or 227 for Mykrobe predictor, since it does not analyse resistance to ETH and PZA (174–176,187).

In order to evaluate the implementation of these platforms in the laboratory setting, it is of utmost importance to measure some performance parameters (like sensitivity, specificity, NPV and PPV) that will allow to better infer the most robust platform to be used for clinical guidance. As a first approach, considering the phenotypic DST results as gold standard, TB profiler showed

the highest sensitivity (92.5%) and NPV (95.0%), whereas Mykrobe predictor showed the highest specificity (97.7%) and PPV (96.0%) (Figure 2.2A). However, as shown in Figure 1, some discrepancies between genotypes and phenotypes were due to the absence of a resistance phenotype. As will be explained below, some of these were considered as phenotypic failures, so the performance of each platform was recalculated by taking into account this correction. With this approach, the sensitivity of resistance prediction ranged from 84.3% using Mykrobe predictor to 95.2% using TB profiler, while specificity was higher and homogeneous among platforms (varying from 93.0% with TGS-TB to 100% with Mykrobe predictor) (Figure 2.2A). Globally, TB profiler revealed the best performance robustness for drug resistance prediction (with all four parameters under evaluation above 95%), followed by TGS-TB (all parameters above 90%), PhyResSE and Mykrobe predictor. All these calculations took into account the number of anti-TB drugs analysed with each platform (Figure 2.2A).

The analysis per anti-TB drug (Figure 2.2B) revealed that, with the exception of the major first line antibiotics (INH and RMP) and of the second-line quinolones, capreomycin, and para-aminosalicylic acid, heterogeneous accuracy values were observed among platforms for the remaining eight antibiotics. In particular, for isoniazid and rifampicin resistance, all platforms showed a prediction accuracy rate above 96% (Figure 2.2B), which is in agreement with the results obtained by other authors (182,184). TB profiler and TGS-TB failed to predict resistance to RMP in two cases and TGS-TB and PhyResSE/Mykrobe predictor failed the prediction of INH-resistance in one and two cases, respectively. For the remaining first line anti-TB drugs, the major differences occurred with the analysis of resistance to streptomycin, ethambutol and pyrazinamide, with accuracy rates ranging from 65-96%, 76-97% and 58-87%, respectively (Figure 2.2B). These high variations are in agreement with previous studies (182,190,192) showing overall accuracy rates on the range of 57-100%, 71-100% and 43-100% for STR, EMB and PZA, respectively. Such variations may be due to the lack of data on molecular mechanisms and mutations in genes that confer resistance to these drugs, as well as to the lack of standardization of phenotypic methodologies or drug concentrations for phenotypic DST, especially for PZA (182,188,189,194). Regarding the second-line anti-TB drugs, detection of 100% of the fluoroquinolone and capreomycin resistant profiles was observed independently of the platform used (Figure 2.2B). These were more promising outcomes than the ones previously described in some studies (182,184) where prediction of fluoroquinolones (FLQ) resistance did not exceed 89.2%. With respect of the remaining aminoglycosides, we observed sensitivities ranging from 40-93% and 64-100% for kanamycin and amikacin, respectively, which is in line with previous studies, where sensitivity for prediction of resistance varied from 60-80% (182–184,192). Globally, TB profiler showed up among the best ranked platforms to predict resistance

for almost all drugs, presenting, for instance, a considerably higher sensitivity (100%) to predict resistance to AMK than the other platforms (64%). A notable exception occurred for PZA for which TGS-TB was more sensitive (87%) than the other platforms (<65%).

2.3.2 Analysis of discrepancies

We found a total of 20 discrepancies (enrolling six phenotypes with no genotype and 14 genotypes with no phenotype) involving resistance to three first-line (STR, EMB and PZA) and three second-line (ETH, AMK and KAN) anti-TB drugs (Figure 2.3A).

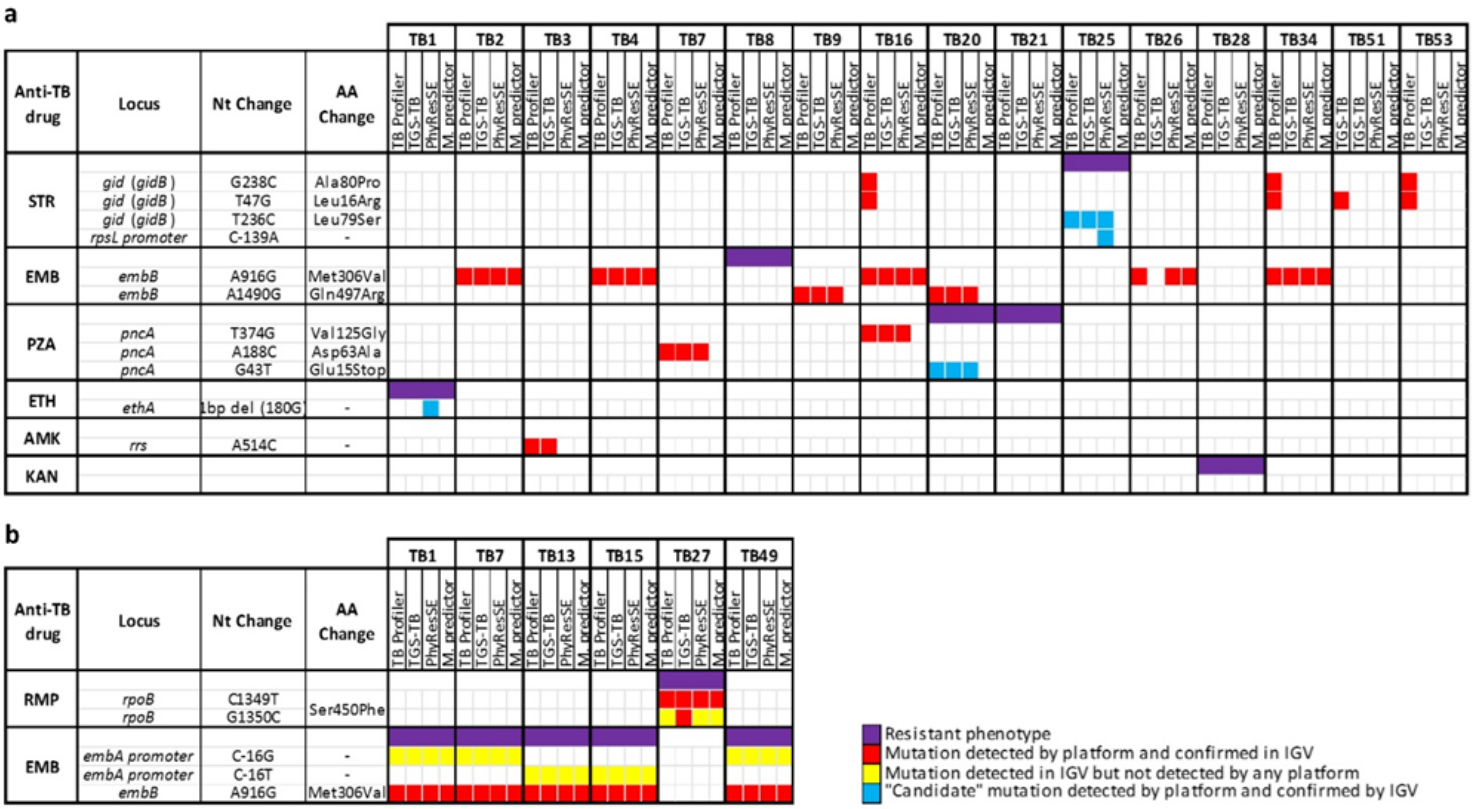


Figure 2.3. Detailed analysis of discrepancies between phenotypes and genotypes found for all 54 MDR-TB strains. (A) For each strain (upper row), only the discrepancies observed for a particular anti-TB drug are shown, as well as the respective locus and associate nucleotide mutations. The concordances among phenotypes and genotypes are not displayed in this figure (as they can be seen in Figure 2.1). For the cases where a resistance phenotype was observed but for which platforms did not point any genotype resistance hit, "candidate" mutations occurring in genes or promoter regions known to be associated with resistance are also shown. (B) Additional mutations (found solely by in-house bioinformatics) that occurred in parallel with other resistance mutations detected by online platforms.

These discrepancies were not due to mixed infections or heteroresistance, as no mixed SNP/indel calls were observed for any of the anti-TB drugs in analysis. Low-density inoculums or bad quality samples were also discarded, since almost all cases of failed correlation between genotype and phenotype occurred for a single drug. In parallel, considering that SNP/indel-calling analysis is influenced by coverage, the later was inspected for each drug-resistance target identified in order to understand if coverage deficiencies/oscillations may underlie the observed discrepancies. This inspection was performed using both raw and trimmed read sequences (since TB Profiler and Mykrobe Predictor do not pre-process the raw data prior to the analysis) covering the 12 genes coding for drug targets and four promoter regions that were affected by the SNPs and indels identified by the four online platforms for all 54 MDR-TB strains. All genes displayed a coverage with medians from 62.6 to 82.6 for raw data and 52.4 to 66.9 for trimmed data, clearly exceeding the tenfold depth cut-off for almost all strains, regardless the pre-processing of reads as well as the %GC content (Figure 2.4). The only exceptions (coverage depth of 8-10) occurred for three strains with lower genome coverage depth as result of a low WGS performance. Similar coverage depth results were observed for promoter regions (data not shown). According to our confirmatory analyses, all genotypic discrepancies involved nucleotide sites whose coverage exceeded the tenfold depth, except for three positions, where a minimum of five reads was accepted to support polymorphism as all of them contained the same mutation. Also, the pre-processing of reads prior to the analysis by some platforms did not significantly impact the coverage depth (Figure 2.4) neither the respective SNP/indel calling analysis.

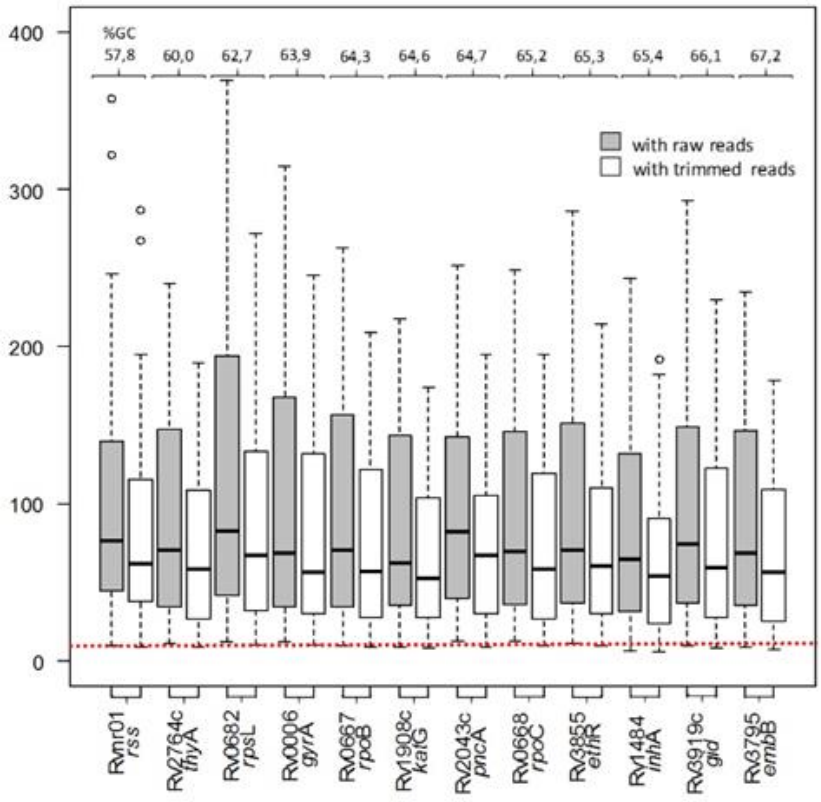


Figure 2.4. Evaluation of the WGS performance for the 12 loci associated with anti-TB resistance. The figure shows the coverage depth for the 12 drug-resistance genes affected by the SNPs and indels identified by the four online platforms for all 54 MDR-TB strains. Boxplot chart was generated using R statistical software v. 3.4.2, and consists of boxes (median and interquartile range) and whiskers that extend to the most extreme data points that were no more than 1.5 times the interquartile range from the box. Boxes represent the variability of coverage depth found for each gene among all strains using both raw and trimmed data. The horizontal black line within each box marks the median, while the lowest and the highest coverage values observed are represented by the extremes of the whisker below and above each box, respectively. Outliers are indicated by open circles. The red line represents the tenfold depth cut-off. The %GC of each gene is also shown above the respective boxplots.

In opposition to resistance to isoniazid and rifampicin, assessing the activity of the other first-line antituberculous drugs presents significant challenges (173–176). Regarding the resistance to streptomycin, four strains (TB16, TB34, TB51, TB53) were found to be exclusively classified as resistant by TB profiler (which highlighted mutations in *gidB* gene) although no resistance phenotype had been observed (Figure 2.3A). All the 37 STR phenotypically resistant strains (with the exception of TB27, TB28 and TB36) that had been correctly predicted as resistant at genotypic level, always exhibited mutations in *rpsL* independently of the existence of other mutations. For TB27 and TB 28 strains, the resistant phenotype was only confirmed by a *gidB* nonsynonymous mutation (Leu16Arg) (Supplementary Table S2.2). As phenotypic failures are known to occur for STR, these data are in agreement with what has been already described by

other authors (184,191), that mutations in *gidB* alone (in particular, Leu16Arg) are not sufficient to provide STR resistance in *M. tuberculosis*. These findings reinforce the assumption that these mutations in *gidB* are not resistance markers (195,196), but phylogenetic informative SNPs that have been historically misclassified. For TB36, besides the GidB Leu16Arg, we also found a C492T synonymous substitution in *rrs*, which has been described as a phylogenetic marker as well (184), pointing to a false phenotype. Since these platforms have also the parallel functionality to analyse the phylogenetic lineage, these *gidB* and *rrs* mutations may be miscalled from the shared databases. Interestingly, concerning TB25 strain, we found phenotypic resistance but, apparently, none of the platforms returned any hit mutation that could be responsible for this resistance. We have then verified the existence of "candidate" mutations in target genes already associated with drug resistance. TGS-TB and PhyResSE highlighted not only a "candidate" mutation in *gidB* gene that was different from any of the two hit mutations detected by TB profiler, but also a mutation in the promoter region of *rpsL*. This particular case constitutes an intriguing scenario, of whether the mutation in the promoter region of *rpsL* (Figure 2.3A) (for which mutations in the coding region have been associated with resistance to STR) is the responsible for the resistant phenotype or the *gidB* mutation (alone or in synergy with *rpsL* mutation) yields resistance to STR (180). On this regard, mutagenesis studies could be helpful to elucidate these hypotheses and further contribute to a revision of the current *M. tuberculosis* drug-resistance databases.

Seven phenotypically sensitive strains revealed mutations associated with resistance to EMB (Figure 2.3A). Such discrepancies are not surprising as this anti-TB drug has been reported by other authors to frequently yield phenotypic failures (103,197,198), so these seven genotypic predictions are likely true positives that were phenotypically missed by routine laboratory procedures. These findings strongly point to an urgent re-evaluation of phenotypic laboratories methodologies for this anti-TB drug (199), highlighting the increasing need to perform confirmatory genotypic DST in the near future for treatment determination. Indeed, EMB resistance is closely associated with M/XDR-TB cases and Ethambutol is used in the treatment regimens only if *M. tuberculosis* strains are susceptible by DST results. Similar to what was observed for STR, there was one strain (TB8) phenotypically resistant to EMB but no known resistance-associated SNP or indel was detected by any of the platforms used for the analysis, not even "candidate" mutations. As no allelic variation was seen on currently known EMB-resistance targets, we speculate that this may not be a true phenotype or it may be due to a not yet identified resistance marker.

With respect to pyrazinamide, we found four strains with no phenotypic/genotypic correlation (Figure 2.3A), which, as above, is not surprising. In fact, on one hand, the routine drug

susceptibility testing is complicated as the growth of bacilli is dependent of the acidic conditions required for optimal drug activity (200–202). On the other hand, our understanding of the mechanisms of resistance to pyrazinamide remains limited (184,188,189). Therefore, the re-evaluation of the laboratory technical procedures for the assessment of the resistance to pyrazinamide is needed. Additionally, DNA sequencing studies have revealed that mutations (SNPs and/or indels) are distributed throughout the entire length of the non-essential *pncA*, suggesting that sequencing the whole gene and respective promoter region would be essential to capture all possible mutations (134,189). Considering that *pncA* is responsible for the activation of the pro-drug pyrazinamide (133), any loss-of-function mutation can confer resistance and a wide variety of mutations were found clinically (191,199). We observed three putative loss-of-function scenarios, involving a 1bp frameshift insertion and large deletions affecting either the 5' or the 3´ of *pncA* (Supplementary Table S2.2). Although none of the four platforms had returned positive results for TB20, when we analysed "candidate" mutations, we found one mutation in *pncA* (Figure 2.3A), detected by three platforms. We hypothesize that this mutation could be responsible for the PZA resistant phenotype, but further studies should be performed in order to evaluate if this "candidate" mutation can be in fact classified as a resistance marker. For TB21, we did not find any mutation or "candidate" mutation that could be responsible for the observed resistance, suggesting a false positive phenotype.

For the second-line anti-TB drugs, we verified that all platforms failed to detect resistance to ethionamide in strain TB1. As a structural analogue of pro-drug INH (203), ETH also inhibits the mycolic acid biosynthesis (204), sharing common molecular targets (205). However, we did not find mutations in *inhA* or within the *inhA* promoter for this specific strain, suggesting that we are not facing a cross resistance INH-ETH scenario (206,207). This strain was resistant to INH but such phenotype is supported by the highly prevalent Ser315Thr alteration in KatG (identified by all platforms) (180,181), which is not associated with ETH resistance (203). Nevertheless, we found a "candidate" 1-bp deletion in *ethA* with PhyResSE, leading to a premature stop codon. Curiously, it has been shown that mutations in *ethA* are associated with ETH resistance (208,209). Indeed, together with the regulatory protein EthR, EthA enzyme has been demonstrated to activate ETH (209,210), so that genetic alterations leading to reduced or even loss of EthA activity would be expected to result in increased ETH resistance (211). Still, whether this deletion is exclusively responsible for the observed ETH-resistant phenotype needs further investigation. In TB3, both TB profiler and TGS-TB found the synonymous mutation A514C in *rrs* gene as the only alteration associated with resistance to amikacin although this strain was phenotypical susceptible. In a recent review, this same mutation was considered as unlikely associated with aminoglycosides, based on expert knowledge (191), but strictly with STR. In

fact, in our survey, this mutation was also identified by all platforms to confer STR resistance in TB3, so we may be facing a miscalling scenario for aminoglycosides as suggested in the literature. The resistant phenotype to kanamycin in TB28 was not genotypically confirmed by any platform (Figure 2.3A). Although a false phenotype cannot be discarded, KAN is known to yield phenotypic consistent data, so we have no reasonable explanation for the putative genotypic failure of all four platforms.

For two first-line anti-TB drugs (RMP and EMB), although a phenotypic/genotypic correlation was found for six strains, we also detected (IGV analysis) additional mutations that were already described as resistance markers in the current *M. tuberculosis* drug-resistance databases (Figure 2.3B). With exception of TGS-TB for TB27, all platforms failed in detecting these extra mutations. We have no reasonable explanation for that as no coverage depth deficiencies were found in these variant positions and surrounding regions. Although we do not have phenotypic evidence, we speculate that these additional mutations may confer higher levels of resistance since this phenomenon was already described for the major first-line anti-TB drug INH with *inh*A and *kat*G genes (212).

2.4 Discussion

One of the major problems for TB control is the lack of accurate and rapid laboratory methodologies, hampering early diagnosis and the initiation of proper treatment regimen. In fact, the phenotypic methods for TB testing are expensive, technically complex and time-consuming (8-15 weeks). It is current practice to initially test only for first-line antibiotics and to subsequently perform second- and third-line susceptibility testing in case of drug resistance (176,213). Although bacterial culture is still a roadblock (7-30 days), the ongoing advances in WGS technology and bioinformatics have expanded opportunities for TB drug resistance monitoring, allowing a faster prediction of resistance to the majority of anti-TB drugs used for treatment (1 to 3 days for sequencing and computational analysis, depending on the sequencing platform and number of samples). This is of considerable importance as it can be used to guide clinical decisions and significantly impact patients' outcome. However, there is still a lack of exhaustive studies to assess their accuracy. In the present study, we simultaneously evaluated the performance of the four most used free online WGS-based platforms to predict resistance to first- and second-line anti-TB drugs, using a set of MDR-TB strains enrolling about 300 phenotypic hits. As SNP/indel calling should reflect the real mutations occurring in the clinical isolates, it is of crucial importance to use WGS platforms less prone to errors. Accordingly, we

have chosen the MiSeq Illumina platform rather than others known to have higher error rates, such as the MinION.

Overall, the sensitivity of resistance prediction ranged from 84.2% using Mykrobe predictor to 95.2% using TB profiler, while specificity was higher and homogeneous among platforms (varying from 94.0% with TGS-TB to 100.0% with Mykrobe predictor) (Figure 2.2A). Our results are in line with previous published surveys performed by other authors around the world (181–183,214). For some MDR-TB strains, a few lack of agreement between phenotype and genotype was observed (Figure 2.3A), but almost all of these cases involved anti-TB drugs known to cause frequent phenotypic failures (STR, EMB, PZA and ETH) (103,184,188,189,191,197,198), strongly pointing out to an urgent re-evaluation of phenotypic laboratories methodologies, and/or mutational hits that had been miscalled as resistance markers from the shared databases (184,191,195,196). Furthermore, we also detected some "candidate" mutations, in particular a SNP in the promoter region of *rpsL* that may be associated with resistance to STR as well as a 1bp-deletion in *ethA* that may confer ETH resistance, highlighting the need to perform mutagenesis-based assays for confirmation purposes.

TB profiler and TGS-TB showed-up as the best ranked platforms to predict resistance to almost all anti-TB drugs, with all performance parameters above 90%, appearing as highly promising tools to be implemented in the routine clinical practice of any microbiology laboratory. In fact, they provide an answer to the clinician at a reasonable cost and tremendously lower time frame when compared with phenotypic methods (183,213). It is rather complex to speculate about the relative importance of each of these performance values so we believe both sensitivity and specificity must be taken into account. Whereas a low sensitivity may discard the usefulness of a platform, a low specificity may impact the beginning of an inaccurate therapeutic regimen. Of worth note, although none of these platforms are licensed for clinical diagnostic use, its clinical impact in allowing the rapid detection of drug resistant TB led to the implementation of similar bioinformatics pipelines in England, in 2017, by the Public Health England Laboratory (215). Furthermore, all mutations described in our survey underlying a phenotype/genotype agreement had been described in other studies, as the patterns of drug resistance emergence seem conserved globally (180,181), which strengths the use of bioinformatics platforms for clinical decision guidance. Nevertheless, as also suggested by other authors (180,181,184,191,192,216,217) the genetic databases underlying *M. tuberculosis* resistance to anti-TB drugs should be reviewed, regularly updated, and likely reunited in a single public database.

2.5 Conclusion

Our findings reinforce the role of WGS to revolutionize the diagnosis of *M. tuberculosis* infection. Using freely available online platforms, we believe that it is now feasible to start implementing WGS analysis in the clinical practice of any microbiology laboratory, with the potential to detect resistance weeks before traditional phenotypic culture methods, which will impact treatment options and prevent resistance dissemination.

Chapter III

Epidemiology of multidrug resistant tuberculosis in Portugal

Trends of MDR-TB clustering in Portugal

Manuscript published in

2019, ERJ Open Research 2019; 5: 00151-2018

DOI: 10.1183/23120541.00151-2018

Rita Macedo, Raquel Duarte

Trends of MDR-TB clustering in Portugal.

RM contributed to the design of the study, performed the experimental work and the bioinformatics analyses, interpreted data and wrote the manuscript.

3. Trends of MDR-TB clustering in Portugal

Multidrug-resistant Tuberculosis (MDR-TB) represent a major threat for global TB control. In 2017, World Health Organization (WHO) estimated 460 000 cases of MDR-TB cases, of which, 8.5% were also extensively drug resistant cases (XDR-TB) (31). In Portugal, over the last decade, the decreasing tendency of TB cases is about 7% per year, and the proportion of MDR-TB cases remain steadily around 1% of the total TB cases. In 2017 (151), the preliminary report of the Portuguese National TB program reported 1607 new cases of pulmonary TB with 12 MDR-TB cases.

Since 2014, there are specific centres for the diagnosis, consultancy, monitoring, and treatment of the M/XDR-TB cases. Besides the clinical approach, these centres also aim to monitor these resistant cases linking the epidemiological survey performed within the community by Public Health Authorities (218) and systematic molecular genotyping performed by the National Reference Laboratory (NRL). Since the Portuguese NRL receives all the strains isolated from all the MDR-TB patients from Portugal (mandatory since 2007) (219), this approach can allow a very good correlation between the genetic and epidemiological information in order to detect both the resistance profiles as possible relations between strains due to occurring ongoing transmission (220,221).

With this study, the authors intend to analyse MDR-TB clustering rate in Portugal.

From a total of 78 M/XDR-TB strains identified and notified in the country during 2014-2017, 71 (91.0%) were available for molecular analysis. From these 78 strains, seven were not available for further analysis due to contamination of the culture or MDR diagnosis based only in molecular biology methodologies (GeneXpert or other line-probe assays). The drug susceptibility profiles are described in table 3.1.

For each strain, 24-loci MIRU-VNTR (Mycobacteria Insterspersed Repetitive Units – Variable Number of Tandem Repeats) genotyping was performed by standardized protocols using MIRU-VNTR typing kit (Genoscreen) according to manufacturer´s instructions. Dendrograms were constructed using the online free software MIRU-VNTRplus (https://miru-vntrplus.org/MIRU/miruinfo.faces). A molecular cluster was defined whenever different strains shared the exact MIRU-VNTR profile. All clusters identified were further analysed with the available epidemiological data.

The majority of the MDR-TB cases were male (75,6%) with a median age of 44.3 years old (minimum 15 and maximum 75 years old). Most of these cases were notified in Lisbon and Tagus Valley (LTV) (64.0%) and North region (23.1%). XDR-TB cases were identified in 15 cases (19.2%), of which 86.7% were from LTV region (table 3.1).

Table 3.1 Microbiological and demographic characteristics of the patients enrolled in the study.

Nr Lab	Diagnosis year		Gender	Age	Origin of isolation	TB-M/XDR	STR	INH	RMP	EMB	PZA1	AMI	CAP	ETI	MOX	OFL	LIN	KAN	CIC	PAS1	Cluster nr
S199228	2014		M	42	Lisbon and Tagus Valley	XDR	R	R	R	R	R	R	R	R	R	R	S	S	S	S	
	2014	no culture isolation	M	43	Lisbon and Tagus Valley	XDR	R	R	R	R	R	R	R	R	R	R	S	S	S	S	
	2014	only LPA	M	47	Lisbon and Tagus Valley	MR	NA	R	R	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	
TB25429	2014		M	41	Lisbon and Tagus Valley	MR	R	R	R	R	R	S	S	R	S	S	S	S	S	S	cluster6
P1595	2014		M	37	Lisbon and Tagus Valley	MR	R	R	R	S	R	S	S	R	S	S	S	S	S	S	cluster1
P1279	2014		F	31	Lisbon and Tagus Valley	MR	R	R	R	R	R	S	S	R	S	S	S	R	S	S	
S207797	2014		M	53	Lisbon and Tagus Valley	MR	R	R	R	S	R	S	S	R	S	S	S	S	S	S	cluster2
P1428	2014		M	48	Lisbon and Tagus Valley	MR	R	R	R	S	R	S	S	R	S	S	S	S	S	S	cluster1
P1378	2014		M	50	Lisbon and Tagus Valley	XDR	R	R	R	R	R	S	S	R	R	R	S	R	S	S	cluster3
P163	2014		M	53	Lisbon and Tagus Valley	XDR	R	R	R	R	R	R	S	R	R	R	S	R	S	S	cluster3
TB24818	2014		F	32	Lisbon and Tagus Valley	MR	R	R	R	R	S	S	S	S	S	S	S	S	S	S	cluster7
P291	2014		M	28	Lisbon and Tagus Valley	MR	R	R	R	R	S	S	S	S	S	S	S	S	S	S	
TB25274	2014		M	37	Lisbon and Tagus Valley	XDR	R	R	R	R	R	S	S	R	S	R	S	R	R	S	
S211891	2014		M	36	Lisbon and Tagus Valley	MR	R	R	R	R	S	S	S	R	S	S	S	S	S	S	cluster7
P1599	2014		M	43	Lisbon and Tagus Valley	MR	R	R	R	R	S	S	S	S	S	S	S	S	S	S	cluster6
P187	2014		M	58	Lisbon and Tagus Valley	MR	R	R	R	R	R	S	S	R	S	S	S	S	S	S	cluster2
P88	2014		M	63	North region	MR	R	R	R	S	S	S	S	S	S	S	S	S	S	S	cluster4
P423	2014		F	75	North region	MR	R	R	R	S	S	S	S	S	S	S	S	S	S	S	cluster4
P292	2014		F	41	North region	MR	R	R	R	S	S	S	S	S	S	S	S	S	S	S	cluster4
P1536	2014		M	23	North region	MR	R	R	R	S	S	S	S	S	S	S	S	S	S	S	cluster4
P729	2014		M	47	North region	MR	R	R	R	S	S	S	S	S	S	S	S	S	S	S	
P356	2014		M	42	North region	MR	S	R	R	S	S	S	S	S	S	S	S	S	S	S	
P92	2014		M	59	Centre region	MR	R	R	R	S	S	S	S	R	S	S	S	S	S	S	cluster2
P340	2014		M	48	Centre region	XDR	R	R	R	R	S	R	S	R	S	R	S	R	S	S	cluster3
	2015	only LPA	M	31	Lisbon and Tagus Valley	MR	NA	R	R	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	
P2624	2015		M	28	Lisbon and Tagus Valley	MR	R	R	R	R	S	S	S	R	S	S	S	S	S	S	cluster7
P1928	2015		M	61	Lisbon and Tagus Valley	MR	R	R	R	R	R	S	S	R	S	S	S	S	S	S	cluster1
P1229	2015		M	39	Lisbon and Tagus Valley	XDR	R	R	R	R	R	R	R	R	R	R	S	R	S	S	cluster1
P1876	2015		M	55	Lisbon and Tagus Valley	XDR	R	R	R	R	R	R	S	R	R	R	S	R	S	S	cluster3
P1926	2015		M	52	Lisbon and Tagus Valley	XDR	R	R	R	R	R	R	R	R	R	R	S	R	S	S	cluster1
P2829	2015		M	44	Lisbon and Tagus Valley	MR	S	R	R	S	S	S	S	R	S	S	S	S	S	S	cluster1
P2184	2015		M	75	Lisbon and Tagus Valley	MR	R	R	R	R	R	R	R	R	S	S	S	R	S	S	cluster1
P1994	2015		M	41	Lisbon and Tagus Valley	MR	R	R	R	S	R	S	S	R	S	S	S	S	S	S	cluster2
P2058	2015		M	57	Lisbon and Tagus Valley	MR	R	R	R	S	R	S	S	R	S	S	S	S	S	S	
TB24737	2015		M	44	Lisbon and Tagus Valley	XDR	R	R	R	R	R	R	R	S	R	R	S	R	S	R	
P1585	2015		F	42	Lisbon and Tagus Valley	MR	R	R	R	R	S	S	S	R	S	S	S	S	S	S	cluster3
P2354	2015		F	27	Lisbon and Tagus Valley	MR	R	R	R	R	R	R	R	S	S	S	S	R	S	S	
P2471	2015		M	34	Lisbon and Tagus Valley	MR	R	R	R	R	S	S	S	R	S	S	S	S	S	S	cluster3
	2015	only LPA	F	32	Lisbon and Tagus Valley	MR	NA	R	R	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	
P884	2015		M	61	Lisbon and Tagus Valley	MR	R	R	R	S	R	S	S	R	S	S	S	S	S	S	cluster2
P2579	2015		M	35	Lisbon and Tagus Valley	MR	R	R	R	R	R	S	S	R	S	S	S	S	S	S	
P2452	2015		M	57	North region	MR	R	R	R	S	S	S	S	R	S	S	S	S	S	S	cluster3
P982	2015		M	37	North region	MR	R	R	R	R	R	S	S	S	S	S	S	S	S	S	
P1880	2015		M	58	North region	MR	R	R	R	S	S	S	S	S	S	S	S	R	S	S	
P2353	2015		M	21	Island of Madeira	MR	S	R	R	R	S	S	S	R	S	S	S	S	S	S	cluster5
P27	2015		M	44	Algarve	MR	S	R	R	S	S	S	S	R	S	S	S	S	S	S	
	2015	only LPA	M	39	Alentejo region	MR	NA	R	R	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	
S308144	2016		M	41	Lisbon and Tagus Valley	XDR	R	R	R	R	R	R	S	R	R	R	S	R	S	S	cluster3
S309968	2016		F	41	North region	MR	R	R	R	S	S	S	S	S	S	S	S	S	S	S	cluster4
S310368	2016		M	42	North region	XDR	R	R	R	R	R	R	S	R	R	R	S	R	S	S	cluster3
S312205	2016		F	63	Lisbon and Tagus Valley	MR	S	R	R	S	R	R	R	R	S	S	S	R	S	S	cluster1
S314371	2016		M	15	Lisbon and Tagus Valley	MR	R	R	R	R	S	S	S	R	S	S	S	S	S	S	cluster6
S324134	2016		M	61	Centre region	MR	R	R	R	R	R	S	S	S	S	S	S	R	S	S	
	2016	no culture isolation	F	45	Lisbon and Tagus Valley	MR	R	R	R	R	R	NA	NA	NA	NA	NA	NA	NA	NA	NA	
S326551	2016		M	40	Lisbon and Tagus Valley	MR	R	R	R	R	R	S	S	S	S	S	S	S	S	S	
S327889	2016		F	22	Lisbon and Tagus Valley	XDR	R	R	R	R	R	R	S	R	R	R	S	R	S	S	cluster3
S320857	2016		M	43	Lisbon and Tagus Valley	MR	R	R	R	R	R	S	S	S	S	S	S	S	S	S	
S316569	2016		M	54	North region	MR	R	R	R	S	S	R	R	S	S	S	S	R	S	S	
S326782	2016		M	62	North region	MR	R	R	R	S	S	S	S	R	S	S	S	S	S	S	cluster3
S340248	2016		M	20	Lisbon and Tagus Valley	MR	R	R	R	S	R	S	S	R	S	S	S	S	S	S	cluster2
TB31393	2016		F	40	Lisbon and Tagus Valley	MR	R	R	R	S	S	S	S	S	S	S	S	S	S	S	
S332846	2016		M	70	Centre region	MR	R	R	R	S	R	S	S	S	S	S	S	S	S	S	
S347401	2016		F	20	Lisbon and Tagus Valley	MR	R	R	R	R	S	S	S	R	S	S	S	S	S	S	
ACC	2016		F	40	Lisbon and Tagus Valley	MR	R	R	R	S	R	S	S	R	S	S	S	S	S	S	cluster2
S333605	2016		M	62	North region	MR	R	R	R	S	S	S	S	R	S	S	S	S	S	S	cluster3
S348387	2016		M	Desc	Centre region	MR	S	R	R	S	S	S	S	S	S	S	S	S	S	S	
S352139	2017		F	22	Centre region	MR	R	R	R	R	R	S	S	S	S	S	S	S	S	S	cluster6
S375001	2017		F	22	Centre region	MR	S	R	R	R	S	S	S	R	S	S	S	S	S	S	cluster5
S374686	2017		F	34	Lisbon and Tagus Valley	MR	S	R	R	R	R	S	S	R	S	S	S	S	S	S	
S381277	2017		M	30	North region	MR	R	R	R	R	S	S	S	S	S	S	S	S	S	S	
TB33470	2017		M	52	Lisbon and Tagus Valley	MR	S	R	R	R	R	S	S	R	S	S	S	S	S	S	cluster1
S389865	2017		F	59	North region	MR	R	R	R	R	R	S	S	S	S	S	S	S	S	S	
S387683	2017		M	56	Lisbon and Tagus Valley	XDR	R	R	R	R	R	R	S	R	R	R	S	R	S	S	cluster3
S399045	2017		F	41	North region	MR	R	R	R	S	S	S	S	R	R	R	S	S	S	S	
S396397	2017		M	51	Lisbon and Tagus Valley	XDR	R	R	R	R	R	R	R	R	S	S	S	R	S	S	cluster1
S399986	2017		M	58	Lisbon and Tagus Valley	MR	R	R	R	R	S	S	S	R	R	R	S	S	S	S	cluster3
TB34192	2017		M	62	Lisbon and Tagus Valley	MR	R	R	R	S	S	S	S	R	S	S	S	S	S	S	cluster2
	2017	only LPA	M	54	North region	MR	NA	R	R	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	

Using Miru-VNTR, seven different clusters were identified (table 3.1), ranging from 2 to 14 strains. Overall, the proportion of MDR-TB cases attributable to recent transmission in the study period (2014-2017), on the basis of genetic data, was 63.4% (45/71).

From the analysis of the molecular data we observed a decreasing tendency of the cases that can be potentially related to recent transmission. In fact, in 2014, we found 6 clusters ranging from 2 to 4 strains, corresponding to a clustering rate of 72,7%. The major cluster was from strains isolated in the North region and the remaining clusters were mainly from LTV strains. Regarding the MDR-TB strains isolated in 2015, 3 clusters were found, with a clustering rate of 55,0%, ranging from 2 to 5 strains. All the clustered strains were from LTV with the exception of one strain that belonged to a patient from the North region. In 2016, only 2 clusters were found with 2 and 5 strains, with a clustering rate of 38,9%. The minor cluster was from a mother/child and the larger included strains from LTV and the North regions. Finally, in 2017, 2 clusters were identified with 2 strains each, all from LTV region, corresponding to a clustering rate of 36,4%.

When linking the epidemiological and the molecular data, we did not find a good agreement. After adjustment for confirmed epidemiological links, the overall cluster rate (2014-2017) decreased to from to 63.4% to 14.9 %.

This study has a limitation related to the possible heterogeneity of the epidemiological enquiries. It has however the strength of collecting all MDR-TB samples in the country for 4 years to be analyzed in the NRL.

We observed, in the studied period, a decreasing tendency both in the number of MDR-TB cases and the clustering rates despite a poor agreement between laboratory and epidemiological data. The centralization of the MDR-TB cases in reference centres seems to be effective, although there is a need for a better molecular tool, with higher discriminatory power, and a better inclusion of epidemiological data when discussing these clusters.

Chapter IV

Development of genomic-based surveillance methodologies

Evaluation of a gene-by-gene approach for prospective whole-genome sequencing-based surveillance of multidrug resistant *Mycobacterium tuberculosis*

Manuscript published with minor changes in
2019, Tuberculosis, 115: 81-88
DOI: 10.1016/j.tube.2019.02.006
Rita Macedo, Miguel Pinto, Vítor Borges, Alexandra Nunes, Olena Oliveira, Isabel Portugal, Raquel Duarte, João Paulo Gomes
Evaluation of a gene-by-gene approach for prospective whole-genome sequencing-based surveillance of multidrug resistant *Mycobacterium tuberculosis*

RM contributed to the design of the study, performed the experimental work, interpreted data and wrote the manuscript.

4. Evaluation of a gene-by-gene approach for prospective whole-genome sequencing-based surveillance of multidrug resistant *Mycobacterium tuberculosis*

Abstract

Whole-genome sequencing (WGS) offer unprecedented resolution for tracking *Mycobacterium tuberculosis* transmission and antibiotic-resistance spread. Still, the establishment of standardized WGS-based pipelines and the definition of epidemiological clusters based on genetic relatedness are under discussion. We report the implementation of a dynamic gene-by-gene approach, fully relying on freely available software, for prospective WGS-based tuberculosis surveillance. Its application for detecting transmission chains was demonstrated by retrospectively analysing all M/XDR strains isolated in 2013-2017 in Portugal. We observed a good correlation between genetic relatedness and epidemiological links, with strongly epilinked clusters displaying mean pairwise allele differences (AD) always below 0.3% (ratio of mean AD over the total number of shared loci between same-cluster strains). This data parallels the genetic distances acquired by the core-SNV analysis, while providing higher resolution and epidemiological concordance than MIRU-VNTR genotyping. The dynamic analysis of strain sub-sets (i.e., increasing the number of shared loci within each sub-set) also strengthens the confidence in detecting epilinked clusters. This gene-by-gene strategy offers several practical benefits (e.g., amenability to standardization, reliance on freely-available software, scalability and low computational requirements) and our data further consolidated it as the method of choice for a timely, standardized and robust prospective WGS-based laboratory surveillance of M/XDR-TB cases.

Keywords

Mycobacterium tuberculosis; Whole-genome sequencing; Surveillance; Gene-by-gene approach; Multidrug-resistance

4.1 Introduction

In 2015, the World Health Organization (WHO) proposed to fight the global drug resistant tuberculosis (TB) crisis as one of its five high priority actions (169). In 2017, of the 10.4 million that fell ill with TB worldwide, 1.7 million died from the disease, with 460 000 cases caused by multidrug-resistant strains (MDR-TB; i.e., resistant to, at least, rifampicin and isoniazid) of which 8.5% were also extensively drug-resistant (XDR-TB; i.e., MDR-TB with additional resistance to any fluoroquinolone and one or more of amikacin, kanamycin or capreomycin) (31). Globally, TB incidence is falling at a rate of about 2% per year but in order to reach the 2020 milestones of the "End TB Strategy" (31) this rate needs to accelerate to a 4–5% annual decline. In Portugal, TB incidence has been steadily decreasing in the last years, with an average of about 5% per year (151). From 2013 to 2017, about 10.000 new TB cases were reported to the Portuguese General Health Directorate (GHD), and the proportion of patients with MDR-TB remained stationary at 1% of total cases (151,222). Considering that human-to-human transmission is the major cause of this epidemic, M/XDR-TB are more difficult and expensive to treat and have poorer survival rates when compared to drug sensitive TB, monitoring and controlling these multi-drug resistant cases is key for successful TB control programs and for achieving the targets of the "End-TB Strategy" (31).

Targeted interventions to stop transmission, especially from M/XDR-TB cases, requires in depth epidemiological knowledge that can only be provided by a combination of effective genotyping with classical epidemiological enquiries. Concerning *M. tuberculosis* complex (MTBC) strains, there are three methods that are commonly used for typing purposes: IS6110 Restriction Fragment Length Polymorphism (RFLP) (152), spoligotyping (interspaced palindromic repeats) (153), and Mycobacterial Interspersed Repetitive Units-Variable Number of Tandem Repeats (MIRU-VNTR) (154). These methods have been widely applied to answer a variety of questions, from the investigation of cross-contamination in the laboratory to the investigation of outbreaks or population-based studies (161,223–227). Still, although traditional typing techniques provide standardized and easily computable typing results, these have limited discriminatory power. In this context, whole-genome sequencing (WGS) has emerged as a very powerful tool for the surveillance, outbreak investigation and drug resistance monitoring of human infectious pathogens including MTBC. As the ongoing technological developments are rapidly decreasing costs, WGS has the potential to become the ultimate tool for diagnostics and pathogen typing (161,228–230), providing comprehensive genetic information, virtually including all possible genomic targets, as well as additional valuable information on drug resistance, virulence and genome evolution (161,172,174,231). Additionally, efforts are being directed to overcome the

time-consuming *M. tuberculosis* growth prior to WGS, as recent developments show that acquisition of WGS data directly from clinical samples may be possible (232–234).

The recent recommendations for TB molecular surveillance by the European Center for Disease Control and Prevention (ECDC) propose to introduce WGS-based surveillance methodologies in the National Reference Laboratories (NRL) of the EU/EEA regions. While the time between sample collection and WGS data analysis is increasingly being shortened, there is still a lack of standardized guidelines for the use of WGS for molecular epidemiological analysis, which is crucial for Public Health Authorities to act. In particular, two practical issues remain in order to sustain alerts for possible ongoing transmission and to maintain data sharing, either locally or globally: i) the definition of the more suitable approach to apply in a WGS-based surveillance era, and ii) the establishment of cut-offs of genetic relatedness that display the highest congruence with epidemiological data for cluster definition. While the latter will become more consistent as more WGS-based studies are being performed (for which consistent epidemiological data is essential), there is still a discussion regarding the methodology to use, namely gene-by-gene and single nucleotide polymorphism (SNP) based approaches (162,228,229,235). In addition, laboratories already performing WGS-based surveillance essentially rely on in-house command-line-based pipelines or commercial platforms (e.g., Bionumerics from Applied Maths or Ridom SeqSphere+ from Ridom Bioinformatics), which may not be accessible to all laboratories, thus potentially delaying the implementation of a harmonized and globally accepted strategy.

In Portugal, since 2014, there are specific centres for the diagnosis, consultancy, monitoring, and treatment of the M/XDR-TB cases. As such, these centres constitute a major driving force for linking the epidemiological survey performed within the community by the Public Health Authorities (218) and the systematic molecular genotyping performed by the NRL at the National Institute of Health (NIH). Since the Portuguese NRL receives all the strains isolated from all the MDR-TB patients from Portugal (mandatory since 2007) (236), it is of utmost importance to set up a centralized and robust molecular typing system that potentiate the establishment of correlations between genetic and epidemiological information towards the detection/monitoring of the resistance profiles and transmission chains (220,221). In light of the transition to a WGS-based laboratory surveillance in Portugal, the present study reports the implementation of a dynamic gene-by-gene approach for WGS-based TB surveillance, through the evaluation of potential transmission chains among the M/XDR-TB cases isolated during the last 5 years (2013 to 2017). We also aimed to contribute for the understanding on how cut-offs of genetic relatedness can be applied to strengthen epidemiological investigation and public health actions.

4.2 Material and methods

4.2.1 Sample dataset characterization and MIRU-VNTR genotyping

All MDR-TB strains isolated in Portugal are mandatorily sent to the TB NRL from the Portuguese NIH for drug susceptibility testing (DST) and genotyping (236,237). Whenever a TB case is notified, public health services intervene in the community in order to identify other potentially associated TB cases or contacts at risk. An epidemiological enquiry is deployed and the information is stored in the Public Health Departments, at regional level. In addition, whenever a cluster is identified by molecular-based laboratory typing techniques, public health authorities are asked to confirm if there is an epidemiological link between the cases. This workflow was applied during the course of the present study with novel WGS-derived links being investigated retrospectively.

From the NIH collection of ~10 000 TB strains isolated at country level from 2013 to 2017, a total of 96 M/XDR TB strains were identified and notified to the Portuguese GHD. From these, 83 (86.5%) strains were available for molecular and genomic analysis (the remaining 13 were diagnosed based on molecular methods with no culture isolation). For 48 strains, DST, 24-loci MIRU-VNTR genotyping and WGS had been previously performed on behalf of a recent study focused on in silico prediction of antibiotic resistance (231). For the remaining isolates, DST was performed for first- (isoniazid-INH, rifampicin-RMP, etambutol-EMB, streptomycine-STR and pyrazinamide-PZA) and second-line (amikacin-AMK, kanamycin-KAN, capreomycin-CAP, ofloxacin-OFX, moxifloxacin-MOX, ethionamide-ETH, linezolid-LNZ, cicloserine-CICLO and para-aminosalicylic acid-PAS) anti-TB drugs, using MGIT960 system (Becton Dickinson), according to manufacturer´s instructions, or the proportion method using solid media (CICLO and PAS).

For the traditional MIRU-VNTR genotyping and WGS, total DNA was extracted from solid cultures using commercial extraction kits (QIAmp, Qiagen) after an initial step of cell inactivation where samples were subjected to 95ºC for 1h followed by enzyme digestion for 3h with proteinase K. The 24-loci MIRU-VNTR experimental genotyping was performed, as previously described (231), by standardized protocols using MIRU-VNTR typing kit (Genoscreen) according to manufacturer´s instructions. Dendrograms were constructed using the online free software MIRU-VNTR plus (238,239). For WGS, high-quality DNA samples (quantified using Qubit, ThermoFisher) were subjected to dual-indexed NexteraXT Illumina library preparation using the KAPA HiFi Hot Start Ready Mix PCR Kit (KAPA Biosystems) in the indexing step to improve amplification of the GC-rich genome regions, as previously described (231). Libraries were subsequently subjected to cluster generation and paired-end sequencing (2×250bp) on an

Illumina MiSeq (Illumina), available at the Portuguese NIH. All data for isolates used in the present study, including drug susceptibility, genotyping and demographic data, are summarized in Supplementary Table S4.1.

4.2.2 Genome de novo assembly

All 83 genomes were *de novo* assembled using the INNUca v3.1 pipeline (https://github.com/B-UMMI/INNUca), which consists of integrated modules for reads QA/QC, *de novo* assembly and post-assembly optimization steps. Briefly, after reads' quality analysis (FastQC v0.11.5 - http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and cleaning (Trimmomatic v0.36) (240), genomes are assembled with SPAdes v3.11 (241) and subsequently improved using Pilon v1.18 (242). Based on a comparative analysis of the impact of performing "post-assembly polishment" of the assemblies using Pilon (data not shown), this step was skipped when Pilon introduced changes in more than 1/3 of the contigs. The affected genomes (n=8) were also among the ones presenting high genome fragmentation (i.e., more than 200 contigs). Draft genome sizes, mean depth of coverage, number of contigs and sequencing read size are described in Supplementary Table S4.1. *In silico* DST were performed using TB-profiler v0.3.0 (184) as previously described (231). MIRU-VNTR profiles were also predicted *in silico* with a recently developed tool, MIRU-profiler (first release), using default parameters (243).

4.2.3 Gene-by-gene analysis

For gene-by-gene analysis, two panel of loci were retrieved (27th of July 2018) from the RIDOM Nomenclature Server (https://www.cgmlst.org/ncs), developed by Kohl and colleagues (228,229). These panels include 2891 genes from the *M. tuberculosis* complex "cgMLSTschema" and additional 755 "accessory" loci. In the present study, we inferred allelic diversity against either only the "short schema" (cgMLST 2891 loci) or the "extended schema" (cgMLST plus accessory loci; 3646 loci). For simplification purposes, the terms "short" and "extended" will be used throughout the manuscript to describe both schemas.

Allele calling was performed for the 83 genomes against both schemas, using chewBBACA v2.0.11 (244) with default parameters and a training file generated by Prodigal v2.6.3 from the H37Rv reference genome (RefSeq Accession **NC_000962.3**). After allele calling using a Blast Score Ratio of 0.6, exact and inferred matches were used to construct an allelic profile matrix, where other allelic classifications (see https://github.com/B-UMMI/chewBBACA/wiki) were assumed as "missing" loci.

Genomes with less than 90% (2602 loci) called in the short schema were removed for subsequent phylogenetic inferences. This occurred only for 3 out of the 83 genomes under analysis (Supplementary Table S4.1). Minimum spanning trees (MST) were constructed taking advantage of goeBURST algorithm (245) implemented in the PHYLOViZ online web-based tool (246), based on 100% shared loci between all strains (i.e., shared-genome MLST). This approach was applied to both schemas and every cluster under evaluation. To take advantage of accessory genome loci, likely increasing the resolution power (which may be key for discriminating cases during outbreak investigation), we took advantage of PHYLOViZ online 2.0 Beta version (http://online2.phyloviz.net/), which allows maximizing the shared genome in a dynamic manner, i.e., for each sub-set of strains under comparison, the maximum number of shared loci between them is automatically used for tree construction. In this regard, all allelic distance thresholds used during cluster investigation were expressed as percentages of allele differences (AD) over the total number of shared loci under comparison. Initial potential clusters to explore were defined by applying an allelic distance cut-off of 0.56% to the initial tree enrolling 80 strains (i.e., 12 AD out of 2202 loci shared in the "short schema"; or 15 AD out of 2657 shared loci in the "extended schema"). The use of this conservative threshold relies on literature data that pointed out 12 AD as an accurate value for cluster investigation in gene-by-gene based surveillance era for *M. tuberculosis* (228,229). Finally, in order to evaluate the clustering agreement between MIRU-VNTR and the gene-by-gene approach, the Wallace's coefficient, with 95% confidence intervals, was calculated using the web-based framework "Comparing partitions" (http://www.comparingpartitions.info/) (247).

4.2.4 Core-Single Nucleotide Variant (SNV)-based analysis

A core-SNV-based analysis was also performed as it is one of the current methodologies used for MTB laboratory surveillance. For this, we applied Snippy v3.1 in the set of all 83 *M. tuberculosis* isolates.(https://github.com/tseemann/snippy). Briefly, quality improved read data (after Trimmomatic processing) of all isolates were individually mapped against the H37Rv reference genome (**NC_000962.3**), and SNP calling was performed on variant sites with the following criteria: minimum proportion of reads differing from the reference of 90%, a minimum mapping quality of 20 and a minimum coverage for SNP calling of 10. Core-SNPs were extracted using Snippy's core module (*snippy-core*) ensuring that all genomes reached at least 95% of aligned bases with the reference. To achieve this, a minimum coverage of 6 was set for five out of the 83 genomes and one genome was excluded from the analysis (TB82). In summary, this

analysis enrols 82 strains (TB30 and TB35 had been excluded in gene-by-gene approach). Core-SNV falling within known *M. tuberculosis* genomic regions with high GC content or repetitive elements (Supplementary Table S4.2), as well as known SNVs in resistance-associated positions, were excluded (compiled by Kohl and colleagues, available at https://github.com/ngs-fzb/MTBseq_source/tree/master/var/res), as their inclusion would likely bias the phylogeny. Following the same rationale applied for gene-by-gene approach, initial potential clusters for investigation in core-SNV approach were defined based on a conservative maximum number of differences of 12 (228,229).

4.2.5 Data availability

All raw sequence reads used in the present study were deposited in the European Nucleotide Archive under the study accession number **PRJEB29446** (detailed in Supplementary Table S1), which will be the Bioproject including all MTB strains analysed for prospective WGS-based surveillance in Portugal. As such, reads that were previously released (BioProject **SRP131205**) have also been deposited in **PRJEB29446**.

4.3 Results

4.3.1 MIRU-VNTR analysis

Our first approach consisted in analyzing the M/XDR-TB strains by MIRU-VNTR genotyping, the conventional method used for the molecular epidemiology of MTBC strains. Overall, we found 41 distinct MIRU-VNTR profiles, constituting 10 different genomic clusters involving 2 to 15 strains each (Supplementary Table S4.1), thus yielding a clustering rate of 63.9% (53/83). Detailed analysis of these clusters will be integrated in the subsequent WGS-based analysis (see next sections). Although the *in silico* determination of VNTR profiles is challenging, a recently developed software (MIRU-profiler) was tested on our assembly data. This *in silico* approach correctly assigned a mean of ~15±3 (±SD) out of 24 VNTR profiles (ranging from 3 up to 20) for each strain, with a mean of 5±3 mismatched (ranging from 0 up to 16) and 4±3 “not detected” loci (ranging from 0 up to 16). In order to understand if these results correlate with the quality of WGS data, MIRU-profiler results were analysed against both the mean depth of coverage (after read quality improvement) (Supplementary Fig. S4.1A) and the number of contigs (Supplementary Fig. S4.1B). As expected, genomes relying on high depth of coverage and with

lower number of contigs yielded more correctly assigned loci. Still, while the number of correct assignments positively correlates with depth of coverage, decreasing this critical parameter seems to lead to more mismatches than loci not found. In an opposite scenario, high genome fragmentation (i.e., increasing number of contigs) seems to correlate mostly with high number of unassigned loci.

4.3.2 Gene-by-gene analysis

To investigate phylogenetic relationships between strains and their potential epidemiological link ("epilink"), we applied a gene-by-gene approach by taking advantage of publicly available schemas [using both a short (2891 loci) and an extended schema (3646 loci) (228,229)] and a freely available software for allele calling (chewBBACA) (244). As both schemas contain "accessory" genes, i.e., genes not present in *M. tuberculosis spp tuberculosis*, only genomes yielding more than 90% of called loci (80 out of 83 strains) in the short schema were subjected to gene-by-gene analysis (Supplementary Table S4.1). In a first step, a MST was generated for each schema based on allelic diversity found among the shared loci: 2202 for the short (Supplementary Fig. S4.2A) and 2657 for the extended schema (Fig. 4.1A). As a conservative approach to generate potential clusters to be subjected to the dynamic MST analysis, the same cut-off of 0.56% was applied for both short and extended schemas (i.e., strains with ≤ 12 and ≤ 15 AD, respectively, were kept interconnected), resulting in seven (Supplementary Fig. S4.2A) and eight (Fig. 4.1A) clusters, respectively. The additional cluster in the extended schema (Cluster 4) was borderline in the short schema (the two enrolled strains, TB54 and TB11, displayed 14 AD; Supplementary Fig. S4.2). The content of the remaining clusters overlapped between both approaches. Of note, at the applied cut-off, the probability of two strains having the same MIRU-VNTR profile also belonging to the same cluster is about 85% (CI 69%-100%) (using either the short or the extended schema), while the probability of two strains that belong to the same cluster (using either the short or extended) also share the same MIRU-VNTR profile is around 50% (CI 35%-64%) (calculations using the Adjusted Wallace coefficient and 95% confidence intervals).

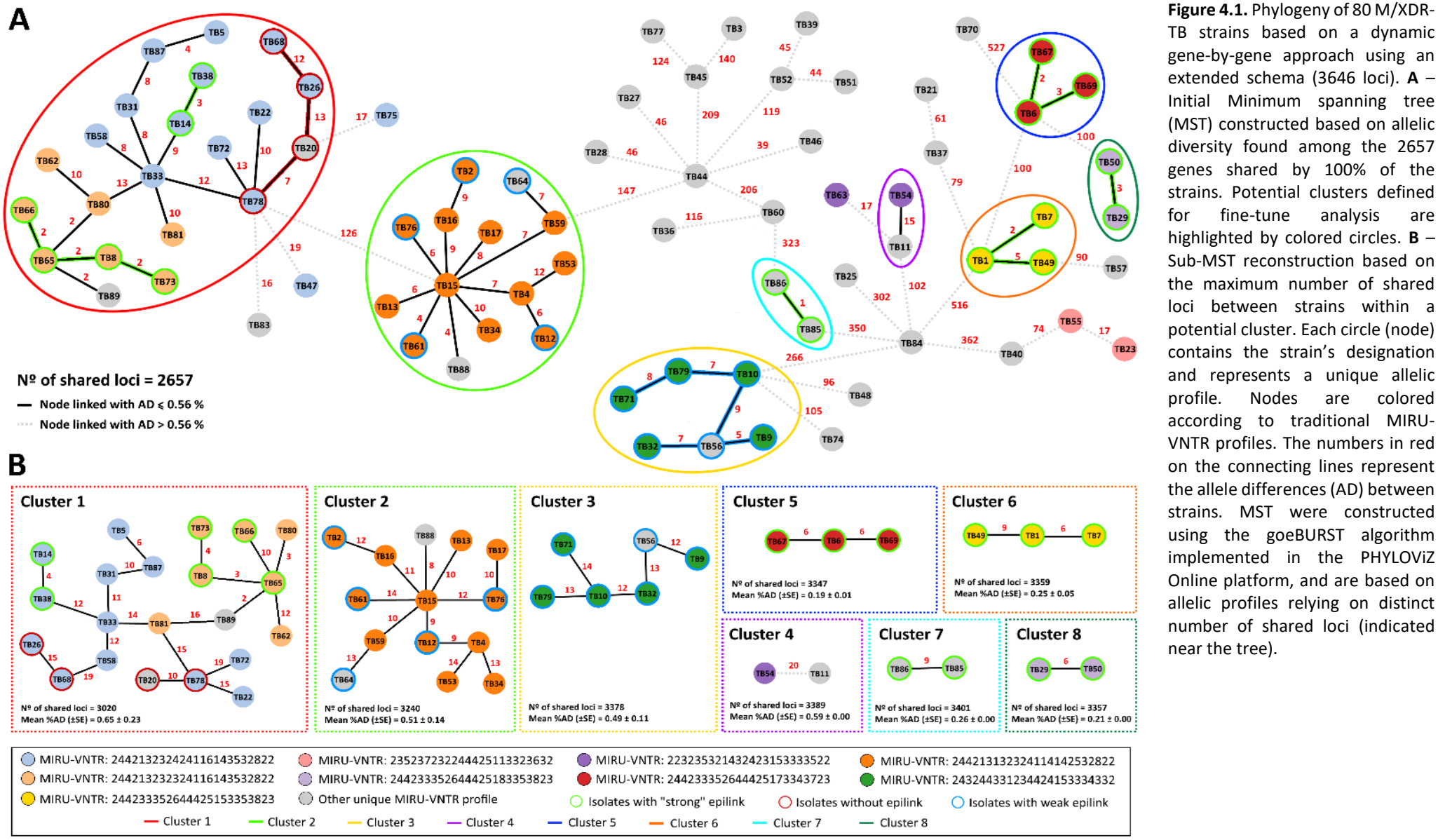


Figure 4.1. Phylogeny of 80 M/XDR-TB strains based on a dynamic gene-by-gene approach using an extended schema (3646 loci). **A** – Initial Minimum spanning tree (MST) constructed based on allelic diversity found among the 2657 genes shared by 100% of the strains. Potential clusters defined for fine-tune analysis are highlighted by colored circles. **B** – Sub-MST reconstruction based on the maximum number of shared loci between strains within a potential cluster. Each circle (node) contains the strain's designation and represents a unique allelic profile. Nodes are colored according to traditional MIRU-VNTR profiles. The numbers in red on the connecting lines represent the allele differences (AD) between strains. MST were constructed using the goeBURST algorithm implemented in the PHYLOViZ Online platform, and are based on allelic profiles relying on distinct number of shared loci (indicated near the tree).

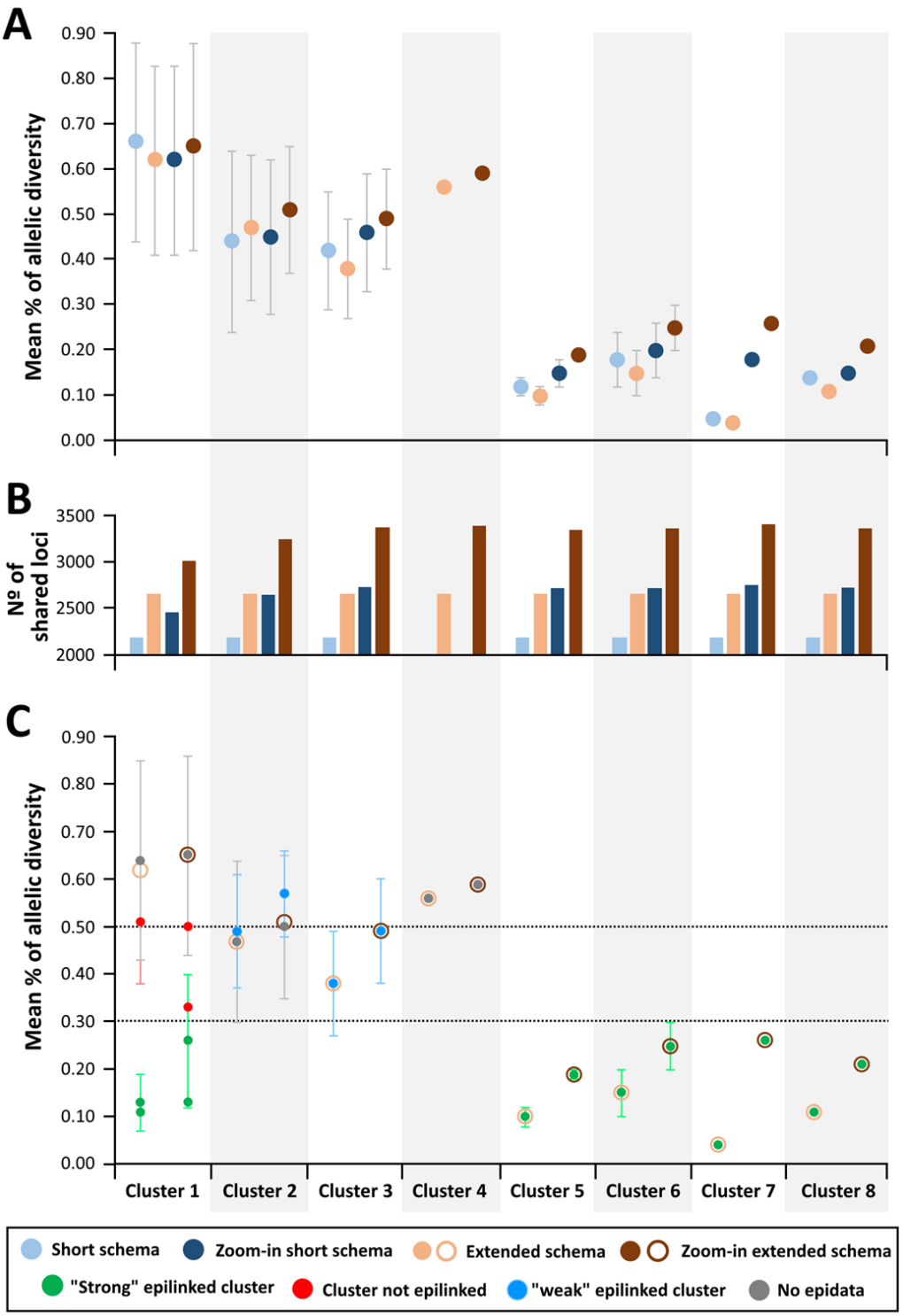


Figure 4.2. Allelic diversity with potential and confirmed clusters. **A** – Comparison of the mean percentage of allelic diversity (with standard deviation) observed within each potential cluster using the short and extended schemas, either with or without sub-minimum spanning tree (MST) reconstruction. **B** – Number of shared loci analyzed for each cluster detected by each approach. **C** – Comparison of the mean percentage of allelic diversity (with standard deviation) observed within each potential cluster, after its fragmentation into sub-sets according to the available epidemiological data. Calculations are presented when using the extended schema, with or without dynamic sub-MST reconstruction. Epilinks are highlighted by different colors.

Within each cluster, the mean percentage of pairwise AD (i.e., the ratio of mean AD over the total number of shared loci) fell within the same magnitude for the "initial MST" (containing all 80 isolates) regardless of the schema used (Fig. 4.2A). A similar analysis after reconstruction of a sub-MST for each cluster, maximizing the number of shared loci between the sub-set of strains, revealed a general increase in mean percentage of AD, with this trend being more marked for the extended schema (Fig. 4.2A), indicating that accessory genes contributed to diversity. As the extended schema enrolls a much higher number of shared loci (Fig. 4.2B), this dynamic approach increases sensitivity, thus allowing a better evaluation of the close-relatedness of strains and, consequently, making cluster definition and exclusion of outliers more robust. For instance, the sub-MST of the additional cluster 4 in the extended schema consolidated the weak genetic link between the two enrolled strains (Fig. 4.1B). On the other hand, for cluster 5, despite the addition of almost 700 shared loci to the analysis, the absolute AD between strains increased by up to 4 AD, thus strengthening the potential link between the three strains. In general, regardless the schema that was used, the diversity within each cluster showed two patterns: 4 clusters (clusters 1 to 4) with mean %AD of 0.4-0.7 and 4 clusters (clusters 5 to 8) with mean %AD of 0.0-0.3 (Fig. 4.2A).

After correlating phylogenetic data with epidemiological data provided by the Public Health Authorities for each potential cluster, strong or weak epilinks could be established among strains within the analyzed clusters. In particular, while all strains within clusters 5 to 8 (with mean %AD below 0.3) had a strong epilink (the authorities confirmed these "strong" epi links as family members or very close contacts) (Fig. 4.2C), a weak epidemiological link was reported for some strains of cluster 2 and all strains of cluster 3 (representing strains from patients of the same geographic region and/or same chest clinic attendance), which could explain the higher mean pairwise %AD observed for the latter. In cluster 1, the largest cluster analyzed, while the overall mean %AD was high, very strong epilinks were reported for two clusters of patients (Fig. 4.2C; Fig. 4.1B). Concordantly, the mean %AD among epilinked strains was below 0.3%, while confirmed non-epilinked strains had %AD within ~0.3-0.5 (Fig. 4.2C). If we applied this empirical observed mean %AD (0.3%) as a linkage cut-off to the sub-MST of cluster 1, three sub-clusters would be obtained (Fig. 4.1B), including the two already confirmed and a novel cluster. In fact, this fine-tune analysis raises the hypothesis that not only TB5 and TB87 might also be epilinked, but also that TB80 and TB89 may be included in the confirmed epilinked cluster containing strains TB66, TB65, TB73 and TB8. Unfortunately, both potential epilinks could not be retrospectively traced. Regarding cluster 2 and 3, although only weak links or no data were retrieved, we cannot discard some possible links among the clustered strains, as observed when applying this cut-off. Of note, adjusted Wallace coefficient analysis at 0.3% threshold showed

that the probability of two strains having the same MIRU-VNTR profile also belonging to the same cluster decreased to about 28% (using either the short or the extended schema), while same-cluster strains have about 60% probability of belonging to the same MIRU-VNTR profile. Nevertheless, a proper inference of the level of congruence will require further large-scale studies with more diverse datasets.

4.3.3 Core-SNP-based analysis

A core-SNV-based approach matrix was used to infer phylogenetic relationships between isolates, after filtering repetitive loci and resistance-associated SNVs. The resulting MST, enrolling 82 strains, is presented in Supplementary Fig. S4.3. As expected, we observed a strong correlation between the mean %AD and the mean number of SNPs (Fig. 4.3) within each of the eight clusters investigated by the gene-by-gene approach (extended schema). Most importantly, when looking at the 10 clusters with epidemiological support (including eight with confirmed epilink and two with confirmed non-epilink), we observed two threshold patterns, where all clusters with strong epilinks have less than 6 SNPs/0.3 %AD, whereas the weak or confirmed non-epilinks have pairwise distances ranging from 6-16 SNPs/0.3-0.7 %AD, sustaining the parallelism with the gene-by-gene approach.

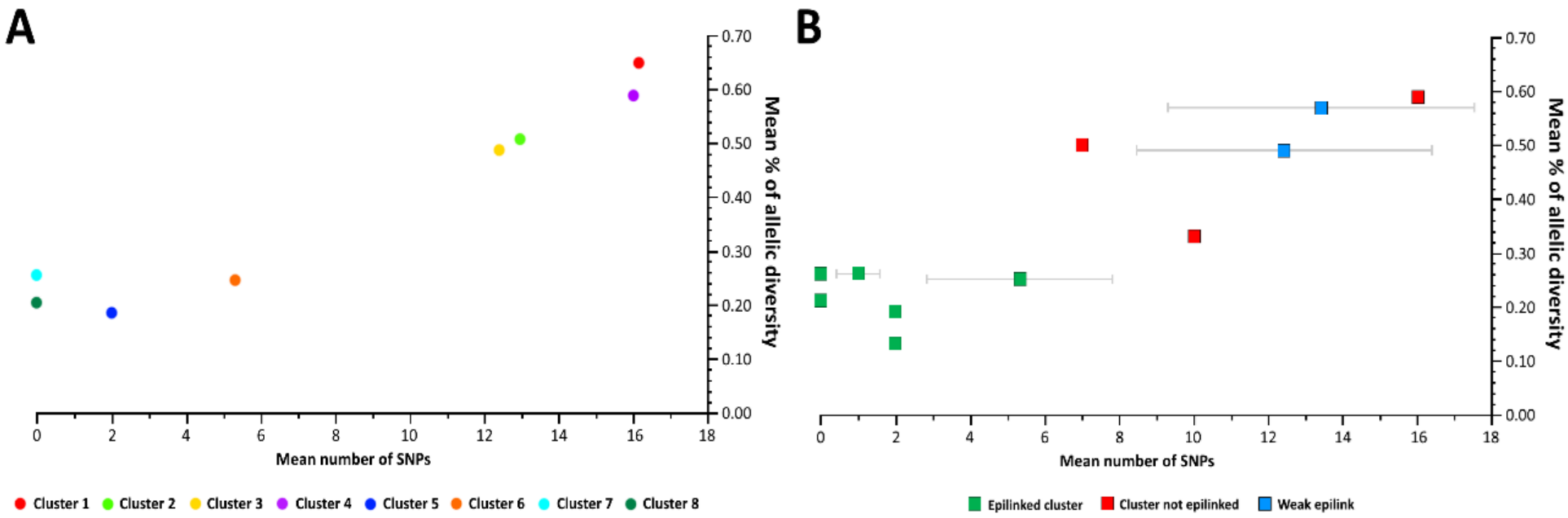


Figure 4.3. Genetic diversity within clusters evaluated by the extended gene-by-gene and the core-SNV approach. Mean pairwise percentage of allelic diversity versus mean pairwise SNP distance within (**A**) each potential cluster defined in the extended gene-by-gene approach using a threshold of 0.56% (Fig. 4.1) and (**B**) their fragmentation according to the available epidemiological data.

4.4 Discussion

The current study describes the implementation of a WGS-based laboratory workflow for TB surveillance in the Portuguese NRL, in the frame of the demanding supra-national short-term transition from traditional genotyping methods (MIRU-VNTR). A collection of 83 M/XDR-TB strains, isolated between 2013 and 2017, was characterized by traditional genotyping as well as by distinct methods for WGS data analysis (i.e., a gene-by-gene approach using both a short schema and an extended schema, and a core-SNV strategy) in order to set up a WGS-based TB surveillance focused on unveiling possible ongoing transmission chains to be flagged for action and further analysis by the health authorities. This study is the first to be performed in Portugal with these specific aims and involves the M/XDR strains isolated in the last 5 years. In general, the same clusters were essentially detected regardless the WGS-based approaches, both in strain number and content (Fig. 4.1; Supplementary Fig. S4.2; Supplementary Fig. S4.3). As expected, strains with the same MIRU-VNTR profile revealed high genetic relatedness at the genome level, even though we could unveil intricate genetic links between strains with distinct MIRU-VNTR patterns. Additionally, our attempt to extract MIRU-VNTR profiles *in silico* (using MIRU-profiler; (243)) yielded unsatisfactory results (Supplementary Fig. S4.1). Although we believe that efforts in this field are needed in order to maintain backward compatibility with traditional genotyping data, our results consolidate the expectation that repetitive genomic regions are still challenging to be extracted from short-read WGS data.

Although similar results were obtained using a gene-by-gene analysis with both short and extended schemas, we observed that the increase in the number of shared loci analyzed may have an important role in the sensitivity (Fig. 4.2), which may be key for outbreak investigation. In fact, the introduction of accessory genes in the schema not only allows performing a dynamic analysis of epidemiological clusters, but also increases the confidence in results, as the link between strains is strengthened when AD decrease or remain very similar after generating sub-MST. This concept of dynamically reconstructing phylogenies for sub-sets of strains, maximizing the number of shared loci, as well as the bioinformatics tools needed for its operationalization, have been set-up on behalf of a recent internationally collaborative project (INNUENDO; https://sites.google.com/site/theinnuendoproject/) that precisely aimed at developing analytical frameworks for the use of WGS in routine surveillance and epidemiologic investigations. Noteworthy, by providing an in-depth perspective on the accessory genomes content, this strategy also opens new possibilities to enrich our knowledge on MTBC biology and population genetics, by identifying for instance, lineage specific genetic markers to use for rapid

cluster identification (233,248,249) or by linking specific phenotypes with gene presence/absence profiles (250).

In the present study, a good correlation was observed between the magnitude of strain genetic relatedness (defined both by gene-by-gene and core-SNV-based strategies) and the strength of the epidemiological link between patients, as reported by the Public Health Authorities. After conservatively defining potential clusters using the implemented dynamic gene-by-gene approach, we could subsequently observe that the mean pairwise distance among strongly epilinked strains was always below 0.3%AD, while confirmed non-epilinked and weakly epilinked strains had >0.3%AD. This trend strongly correlated with data acquired through the core-SNV strategy, with strains sharing strong epilinks displaying mean pairwise distances of less than 6 SNPs and strains with weak or confirmed non-epilinks with up to 16 SNPs (Fig. 4.3). Of note, we found very close related strains within this threshold for which epidemiological data was not available, raising the hypothesis that some retrospective links may have been missed. These results are in agreement with previous studies, where genetic distance among strains from recent transmission chains reveal a maximum number of 12 SNPs/AD (143,183,228,229,251,252), and also corroborate the expectation that applying lower thresholds will increase the concordance with epidemiological data. Several studies have already applied even more stringent thresholds (234,235). Still, we believe that cut-offs should not be applied in a static manner, but instead should be dynamically fit to the sub-set of strains/lineages under investigation, which fits the evolutionary and epidemiological dynamics of *M. tuberculosis* that relies on a low mutation rate coupled with a long-term transmissibility. In fact, this can be operationally achieved with the gene-by-gene strategy described in this study as the maximum number of shared loci is automatically used for each potential cluster, with allelic distance thresholds being expressed as percentages of allele differences (AD) over the total number of shared loci under comparison. Moreover, this dynamic analysis should always be contingent on the epidemiological data, when available. In this context, this study illustrates the need to reinforce the conventional epidemiological tracing (e.g., by improving clinical inquiries and strengthening the communication workflow between stakeholders) in order to potentiate the benefits of an enhanced WGS-based laboratory surveillance focused on detecting/confirming more robust contact tracing investigation.

While the advantages of WGS-based typing analysis are undeniable for routine TB laboratory surveillance, there are still several drawbacks that need to be overcome before a multi-country harmonized WGS-based surveillance (253), such as: i) the lack of standardized pipelines; ii) the need of empirical-derived thresholds for cluster definition and, iii) the existence of an internationally agreed nomenclature to be used in order to facilitate data sharing, especially for

cross-border outbreaks (173,213). Also, there is still an ongoing debate on which approach will be the most reliable method, i.e., gene-by-gene-based, SNP-based or a combination of both (143,162,228,229,251,252). In opposition to core SNP-based approaches, our main method of choice for future WGS-based TB surveillance in the Portuguese NIH, the gene-by-gene strategy, offers several benefits such as: amenability to standardization, higher scalability, likely lower computational requirements and less constrains to implement a harmonized nomenclature (254,255). In addition, considering that our methodology fully relies on freely available software, in contrast with pay-per-use platforms (such as RIDOM SeqSphere) currently applied by several reference laboratories, we believe that the present study's approach may constitute a valuable alternative for most laboratories, especially those with less resources.

In summary, this study stands as the "starting point" for the application of a freely-available standardized routine WGS-based laboratory surveillance of M/XDR-TB cases. Together with the centralization of the diagnosis and molecular typing in the NIH and a strong articulation with the Public Health Authorities, the described framework constitutes the driving force towards a faster and robust prospective surveillance of M/XDR-TB cases, from antibiotic resistance prediction to transmission chain detection. As such, future studies based on this approach will be crucial to consolidate the benefits of this technological transition for Public Health.

Acknowledgements

We would like to thank to Prof. João A. Carriço and Bruno Ribeiro-Gonçalves for making the PHYLOViZ online 2.0 Beta version available for the analysis performed in the present study.

O. Oliveira is supported by the Project NORTE-08-5369-FSE-000041, financed by the Operational Program NORTE 2020 and co-financed by the European Social Fund through a doctoral grant (grant number UMINHO/BD/47/2016).

Chapter V

Whole-genome-sequencing of *M. tuberculosis* directly from clinical samples

On-going work

Rita Macedo, Bioinformatics team

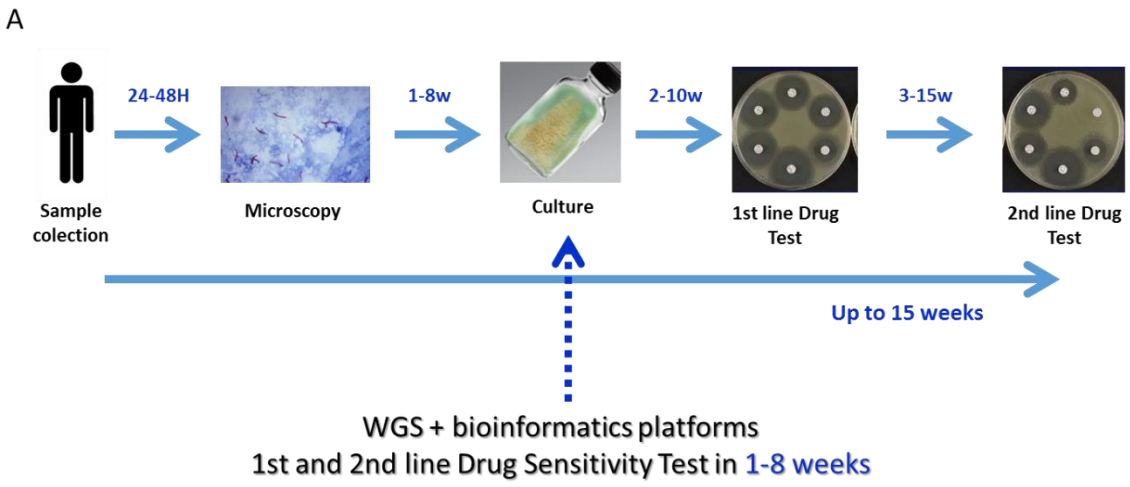
RM contributed to the design of the study, wrote the chapter and is performing most of the experimental work.

5. Whole-genome-sequencing of *Mycobacterium tuberculosis* directly from clinical samples

5.1 Introduction

The gold standard for the routine clinical diagnosis and DST for *M. tuberculosis* is culture-based, which requires months for visible growth. As such, antimicrobial resistance detection can be delayed (until 6 months after culture becomes positive), leading to a potential prolonged suboptimal antibiotic treatment. As mentioned in the above chapters of this Ph.D. thesis, there are several molecular assays already endorsed by the WHO, but they fall short on targets and resistance-related regions or genes, to assure a correct prediction of resistance. The potential of WGS as a diagnostic assay has been repeatedly demonstrated and enables a comprehensive identification of all known-resistant mutations for all TB drugs as well as it can provide reliable contact tracing information (183,213,228,256–258). For these reasons, we have implemented a WGS-based methodology as a routine for early positive cultures identification, resistance prediction and surveillance (231,259). Moreover, WGS approach performed at a comparable cost to phenotypic assays offering short turnaround times.

However, as early positive cultures may still represent some weeks of bacterial growth, generating WGS information directly from samples avoiding the time-consuming need for culture would constitute a tremendous achievement. (Figure 5.1). But this approach has a major potential hurdle, because biological samples contain variable amounts of human cells (mixed with *M. tuberculosis* cells) that account for up to 99.9% of the total DNA present (166,167,260,261).



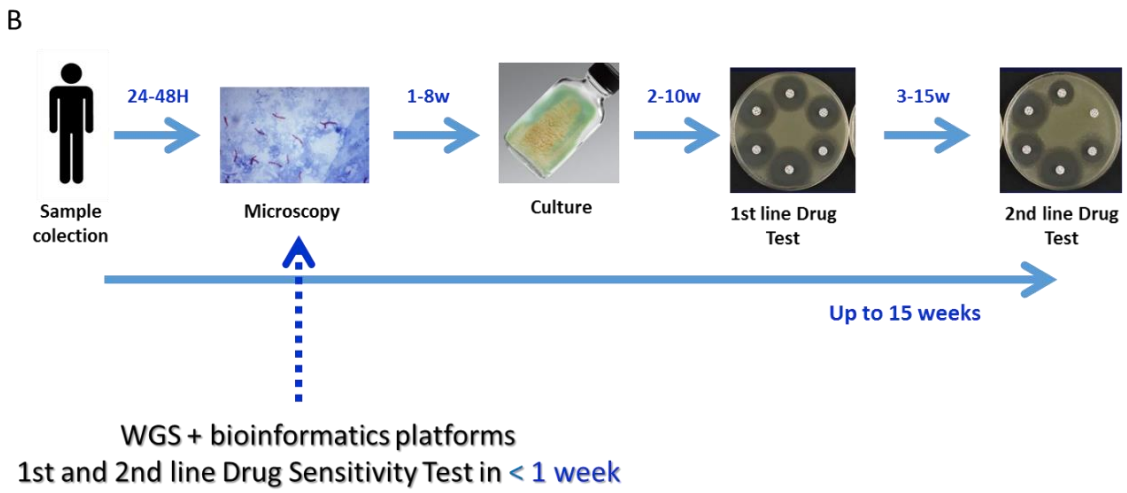


Figure 5.1. Diagnostics workflows and time-to-results shift using the traditional phenotypic methodologies and WGS-based approaches directly from cultures (A) or from smear-positive sputum samples (B).

Several studies have already described the use a specific protocol, which uses SureSelect™ (Agilent) custom designed 120-mer RNA oligonucleotides that span the entire microbial genome, and can recover (by hybridization) low copy numbers of DNA directly from clinical samples with sufficiently high sensitivity and specificity to enable efficient WGS (262,263).

M. tuberculosis is particularly appropriate for the use of diagnostic WGS with enrichment, since, unlike the majority of pathogenic organisms, it has a well characterized clonal nature, with low levels of sequence variation and does not undergo recombination or horizontal transfer (264); thus, a stable set of oligonucleotide baits can be created and sequence data can be mapped against a reference genome (167). However, mycobacterial cells may aggregate because of the high mucus content of respiratory samples, meaning that the volume and AFB count may not represent the total quantity of mycobacteria available (166,167,261). Direct samples therefore require pre-processing for homogenization and enrichment purposes, and to deplete other cells/DNA.

The main objective of this study is to provide a proof-of-concept regarding the application of WGS-based methodologies directly from sputum samples, for resistance prediction and surveillance, in *M. tuberculosis*.

5.2 Materials and Methods

5.2.1 Samples

Smear-positive sputum samples retrieved from pulmonary TB patients and that were sent to the NRL-TB from the Portuguese NIH were tested. Samples were decontaminated using N-acetyl-l-cysteine/NAOH (1% NaOH final concentration) and resuspended after centrifugation in 2 mL phosphate buffer (pH 6.8). After inoculation for phenotypic testing, all the remaining sputum specimens were kept frozen at −20°C until further use. Only samples with AFB scoring of 3+ (visually quantified according to WHO guidelines) were considered.

5.2.2 Phenotypic resistance profiles

All isolates were tested for susceptibility to first-line drugs rifampicin (RIF), isoniazid (INH), ethambutol (EMB), pyrazinamide (PZA), and streptomycin (STR). Isolates resistant to at least RIF and INH (i.e., representing multidrug-resistant TB [MDR-TB]) were additionally tested for susceptibility to kanamycin (KAN), amikacin (AMK), ofloxacin (OFL), capreomycin (CAP), ethionamide (ETH), prothionamide (PTH), and para-aminosalicylate sodium (PAS). Drug susceptibility testing (DST) was carried out on an automated liquid medium-based system, Bactec MGIT960 (Becton Dickinson), using standard drug concentrations (in micrograms per milliliter) as follows: for STR, 1.0; for INH, 0.1; for RIF, 1.0; for EMB, 5.0; for PZA, 100.0; for OFL, 2.0; for AMK, 1.0; for CAP, 2.5; for KAN, 5.0; for ETH, 5.0; for PTH, 2.5; and for PAS, 4.0.

5.2.3 DNA extraction

In order to measure the effect of human DNA depletion prior to DNA extraction on the success of WGS directly from clinical samples, two DNA extraction protocols were tested in twin aliquots of each biological sample. These two protocols differed in the inclusion of an initial sonication step and treatment with a DNase/RNase solution.

Briefly, the bacterial suspension used for inoculation was repelleted by centrifugation at 14,000 × g for 10 min. Supernatants were decanted, and pelleted cells were resuspended in 0.2 mL PBS buffer. Microorganisms were heat killed at 95°C for 1 hour; after inactivation, the samples were subjected to two cycles of ultrasounds for 1 min each (S30 Elmasonic). After sonication, 0.2 mL

of a DNase/RNase solution was added and incubated for 30 minutes at 37ºC. The samples were then centrifuged at 14,000 × g for 10 min, the supernatant (containing the digested nucleic acids from the host cells) discarded and the pellet ressuspended in 0.2 mL of a DNase/RNase solution, subjected to 2 additional cycles of sonication for 1 min each (S30 Elmasonic) and incubated for 30 minutes at 37ºC. The DNase/RNase solution was inactivated at 65ºC for 30 min followed by 1 min on ice, centrifuged at 14,000 × g for 10 min and 0.2 mL PBS buffer was added to the pellet. TE 1X buffer (160 μL) and lysozyme (40 μL) were added and the samples were incubated for 1h at 37ºC. Next, 180 μL lysis buffer (Bioline) and 20 μL proteinase K were added to the samples, followed by incubation for 3h at 56ºC. The remaining protocol steps were performed according to manufacturer´s instructions using the Isolate II Genomic DNA kit (Bioline).

In parallel, the twin aliquots of the same samples were extracted without the human DNA depletion steps, using the lysozyme solution, proteinase K and the Isolate II Genomic DNA kit (Bioline), as described above. At the end of the process, we proceeded with the subsequent experimental steps for four pairs of aliquots, corresponding to four different sputum samples.

5.2.4 Generation of standard curves for real-time quantitative PCR (qPCR)

To quantify the number of *M. tuberculosis* genomes in each sample, a plasmid standard curve was generated as previously described for other pathogens (263,265,266). Primers for the conserved MTBC single-copy gene *katG* were designed based on constant regions (primers KatG-A TTACCGCTGGGCGTGTTC and KatG-B TCACGAAGAAGTCGTTGGTCAGT designed using Primer Express software; Applied Biosystems), according to the sequence of MTB H37Rv strain (Genbank # AL123456). Briefly, an amplified fragment (58 bp) of *katG* was cloned into the pCR® 2.1 vector using the TOPO TA technology (Invitrogen, MA, USA) according to the manufacturer's instructions. After transformation of DH5α *E. coli* with the cloned vector and subsequent overnight propagation, plasmid DNA was purified and transformation was confirmed by PCR and sequencing. The plasmid copy number was calculated according to the formula: Nº plasmid/mL = (Avogadro's Nº x Plasmid conc. (g/mL))/MW of 1 mol of plasmids (g). The standard curve consisted of eight-serial plasmid dilutions (~1 to 1 x 10^8 plasmid copies/μL). The number of human cells/sample was quantified by a similarly generated plasmid standard curve using an amplified fragment (73 bp) of a single copy human gene (*b-actin*) cloned in a similar vector, according to Gomes *et al.* 2006 (primers B-actin-3 GGTGCATCTCTGCCTTACAGATC and B-actin-4 ACAGCCTGGATAGCAACGTACAT).

5.2.5 qPCR for quantification of MTB *vs* human cells

The real-time quantification was performed using the Light-Cycler® 480 SYBR Green chemistry and optical plates (Roche Diagnostics, Basel, Switzerland). The qPCR reagents consisted of 2 x SYBR Green I Master Mix, 400 nM of each primer and 5 µL of DNA sample in a final volume of 25 µL. The thermocycling profile was: 10 min/95 °C followed by 40 cycles of 15 s/95 °C and 1 min/ 60 °C. Specificity was checked by generating the dissociation melting curves. Absolute quantification of bacterial and human genomes was calculated in relation to the respective plasmid standard curve. The relative load of MTBC cells in each sample was determined as the ratio between the number of *katG* and *b-actin* copies.

5.2.6 DNA capture directly from clinical samples

In order to capture the *M. tuberculosis* DNA, directly from clinical samples, complementary RNA oligonucleotide "baits", 120 bp in size, were designed to span the ~4.5 Mb of the *M. tuberculosis* genome. As such, the reference genome sequence of the MTBC H37Rv strain (Genbank #AL123456) was *in silico* fragmented into 120 bp sequences twice, to ensure an overlap of 60 bp between sequences. Due to their rich GC content, which could hamper DNA capture, all MTBC genes of the PE, PPE and PE-PGRS family were also independently fragmented into 120 bp sequences, in order to increase capture sensitivity. All resulting sequences were BLASTn searched against the Human Genomic + Transcript database to excluded homologous sequences to the human genome. Overall, a total of 42,278 RNA probes were generated and this custom bait library was then uploaded to the SureDesign software (https://earray.chem.agilent.com/suredesign/) and synthesized by Agilent Technologies. During synthesis, the 2199 sequences complementary to the PE, PPE and PE-PGRS family were unbalanced 8:1 to potentiate capture. Before enrichment and WGS, DNA samples were quantified using Qubit HS kit (Invitrogen, Life Technologies) to calibrate the input to 10-200 ng DNA.

5.2.7 SureSelect$^{XT\ HS}$ target enrichment: library preparation, hybridization, and whole genome sequencing

Prior to the preparation of sequencing libraries, DNA was fragmented with an enzymatic procedure.

For the libraries preparation we used SureSelect$^{XT\ HS}$ Target Enrichment System for Illumina Paired-End Multiplexed Sequencing Library (Agilent). Instead of the mechanical shearing (referred in Step 2 of the abovementioned protocol), we used the SureSelect$^{XT\ HS}$ Low input Enzymatic Fragmentation kit (Agilent) for DNA shearing, according to manufacturer´s instructions. We then continued the library preparation at "Step 3. Repair and dA-Tail the DNA Ends" of the protocol.

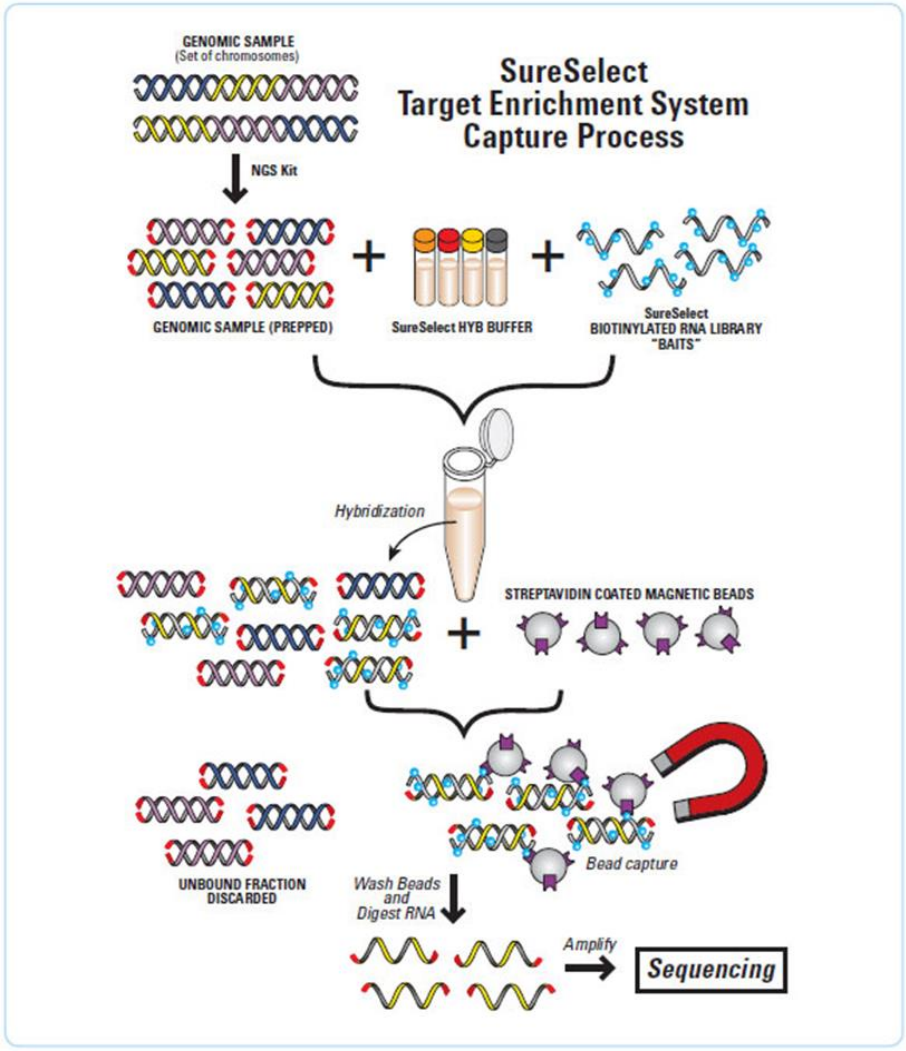


Figure 5.2. Schematic protocol of the *M. tuberculosis* DNA enrichment SureSelect$^{XT\ HS}$ target enrichment prior to WGS.

5.2.8 WGS analysis

To access the percentage of on-target reads, after quality analysis (FastQC v0.11.5 - http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and trimming (Trimmomatic v0.36) (240), the improved reads were mapped (Bowtie 2 v2.2.9) against the reference MTB H37Rv strain genome (Genbank #AL123456). Likewise, the reads were also mapped against the human reference genome GRCh38.p12 (assembly #GCA_000001405.27).

All genomes were *de novo* assembled using the INNUca v4.0.1 pipeline (https://github.com/B-UMMI/INNUca), which consists of integrated modules for reads QA/QC, *de novo* assembly and post-assembly optimization steps. Briefly, after reads' quality analysis and cleaning, genomes are assembled with SPAdes v3.9.0 (241) and subsequently improved using Pilon v1.18 (242). To allow a higher horizontal coverage of the genomes, the parameters *estimatedMinimumCoverage* and *assemblyMinCoverageContigs* were set to the value 5. In order to access the suitability of the obtained genomes directly from the clinical samples for subsequent *in silico* DST and surveillance purposes, we subsequently used TB-profiler v0.3.0 (267) as previously described (231), and performed gene-by-gene analysis, according to the study that is presented in the chapter IV of this PhD thesis (259).

5.3 Results

Four sputum samples were subjected to the workflow presented at the Methods section (point 5.2.3), which includes two parallel DNA extraction protocols. As shown in Figure 5.3, the human DNA-depletion protocol was very successful for all samples analysed and did not compromise the yield of *M. tuberculosis* DNA extraction. One of the samples (TB 36163A) did not meet the concentration criteria required for further processing (10-200 ng) and was thus excluded. As such, four human DNA-depleted samples and three non-depleted were processed and analysed.

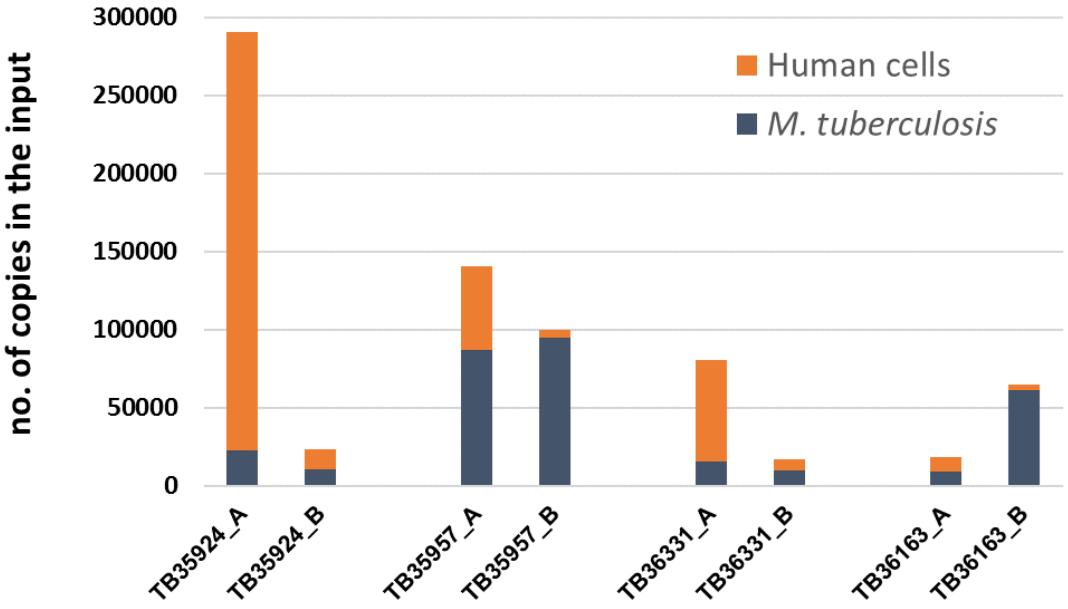


Figure 5.3. Results of the number of copies of human and *M. tuberculosis* after DNA extraction protocol with (B) or without (A) the human-DNA depletion step.

The percentage of reads mapping against the reference MTB H37Rv strain genome ranged from 56.69% to 97.22%, with 96.9-97.7% of the genome being covered by at least one read. Conversely, the percentage of reads mapping against the human genome ranged from 0.22% to 5.1% (Figure 5.4). Although all the human DNA-depleted samples (labeled "B") presented a lower level of human DNA when compared to their non-depleted pair, these values were not significantly lower and thus the success of the methodology seemed to be independent of this previous depletion.

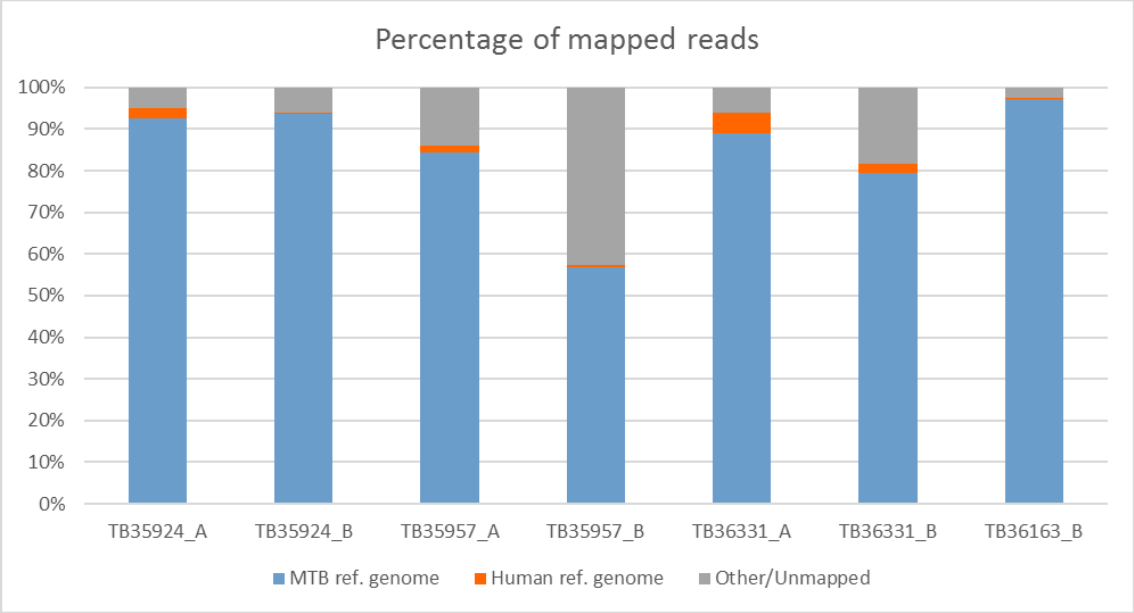


Figure 5.4. Percentage of reads mapping against the *M. tuberculosis* (Genbank #AL123456) and human (assembly #GCA_000001405.27) reference genomes. Samples labeled with "A" were extracted without human DNA depletion. Samples labeled with "B" were subjected to human DNA depletion.

All samples were analysed using TB-profiler v0.3.0 and genotypic and phenotypic data were concordant. With the exception of TB36313, all the other samples were fully susceptible to all first-, second-, and third-line anti-TB drugs. TB36313 was phenotypically resistant to STR, which was confirmed genotypically by the identification of a mutation in the *rrs* gene.

Since the main objective of this section was to establish a proof-of-concept on the efficiency of MTBC DNA capture directly from clinical sputum samples, for this preliminary assay, we decided to sequence the libraries in an Illumina MiSeq platform, using the Reagent Micro Kit v2 (300-cycles). Surprisingly, given the high specificity of the RNA "baits" used for the capture and enrichment protocol (reflected in the high percentage of on-target reads), it was possible to assemble the genomes with a depth of coverage ranging from 21.2x to 32.4x, which was homogenously distributed throughout the genome (Figure 5.5). Hence, we can assume that if a higher output sequencing kit was used, the coverages would be significantly higher and comparable to those that are generally obtained from pure cultures. Furthermore, we were able to retrieve sequence data from GC-rich regions, namely, the PE/PPE family genes, for which the implemented routine WGS-based protocol (using pure cultures and mentioned in the previous chapters) does not yield sufficient depth of coverage sequencing (Figure 5.5). This is likely due to the specific enrichment of RNA "baits" spanning these regions, in a 8:1 ratio. The rRNA operon, however, presents a high coverage most likely because of the presence of several non-MTB microorganisms in the sputum samples. The sequence conservation of this operon within bacteria may thus justify a high dedication of reads to this genomic region.

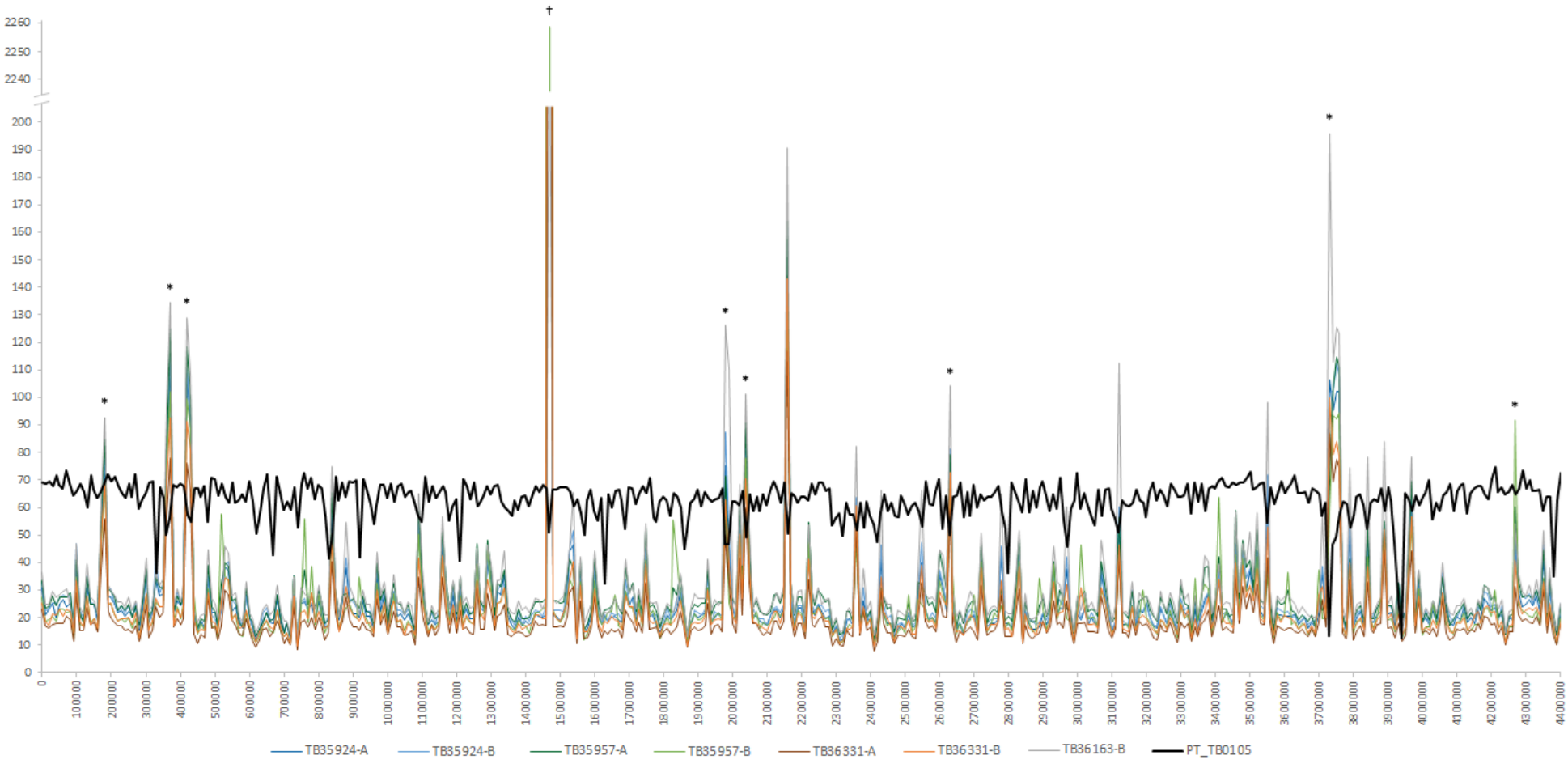


Figure 5.5. Depth of coverage of the genomes sequenced directly from sputum samples. All sequences were mapped against the MTB H37Rv reference genome (Genbank #AL123456). An additional genome (PT_TB0105, in black), previously sequenced from a pure culture and published in a previous work (259), was added for comparison purposes. High depth of coverage peaks (presenting >3-fold the mean depth of coverage) in the sputum sequenced samples were inspected: peaks marked with an "*" refer to PE/PPE genes and the peak marked with a "†" indicates the rRNA genes operon.

It was also possible to perform a gene-by-gene analysis, as described in Chapter IV of the present thesis (259). In order to frame these samples within the set of Portuguese MDR-TB strains previously analysed, after allele calling, and for simplicity purposes, we chose the assemblies that performed better from each pair of samples. This approach allowed the comparison of these four genomes with the remaining 112 strains, using a set 1709 shared loci, showing the potential of WGS directly from clinical samples for real-time surveillance purposes (Figure 5.6).

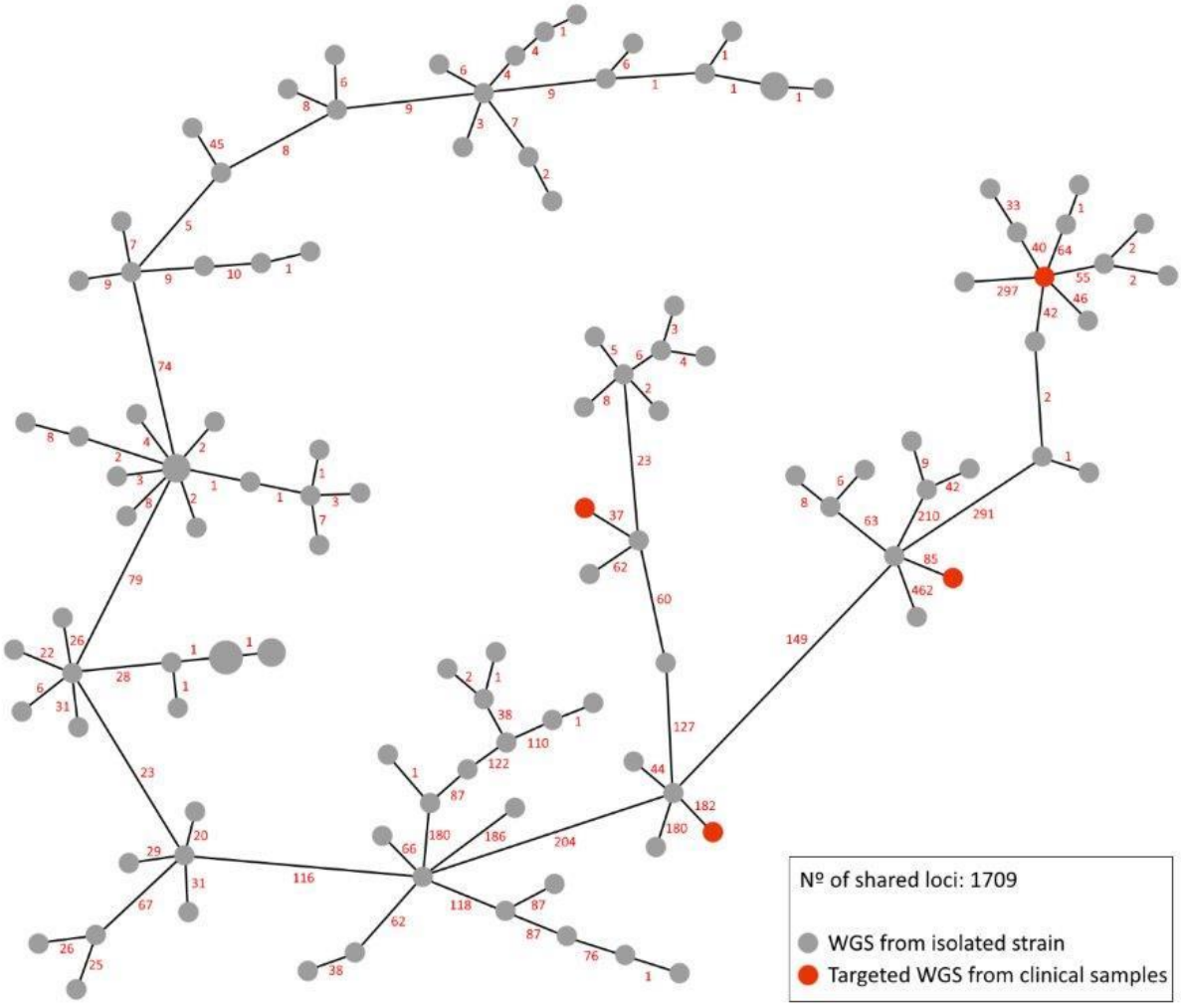


Figure 5.6. MST of all MTBC strains used for surveillance purposes highlighting (marked as red dots) the phylogenetic position of the genomes that were captured directly from clinical samples.

5.4 Discussion / Perspectives

The capture of large genomic sequences directly from clinical samples, such as the baits-based Agilent SureSelect procedure, has been widely used in genetic-related diagnoses such hemoglobinopathy cases (268) but, in the microbiology field, it has been used mostly for viruses (269).

Here, we aimed to provide a proof-of-concept on the utility of WGS-based methodologies directly from sputum samples, for resistance prediction and surveillance, in *M. tuberculosis* using a similar enrichment approach.

Overcoming the constraints of time-consuming laboratory procedures for the isolation of MTBC strains in culture, we were able to provide a methodology that allowed not only the identification and prediction of genotypic resistance, but also the possibility to integrate this approach in a "real-time" MDR-TB surveillance for the rapid articulation with the public health authorities. Within a 5-day wet-lab procedure, after DNA isolation directly from sputum samples, we were able to retrieve all the information needed for routine diagnostic purposes, skipping the 1-3 weeks period required for culture isolation.

Furthermore, unlike WGS performed from DNA isolated from pure cultures, which potentially leads to the loss of information on the existence of MTBC sub-populations, targeted sequencing directly from clinical samples will likely provide information on the real scenario of the *in vivo* sub-populations that might co-exist during the infection period. Although this is a preliminary work, in which we have obtained relatively low coverage genomes exclusively due to the choice for a low output sequencing flow cell, future assays aiming for higher coverages will allow deeper and refined analysis, such as the potential prediction of transmission chains through the analysis of sub-populations.

Furthermore, a rough estimation on the total cost of this methodology showed, comparing to culture WGS-based analysis, an increase of > 150 euros. Thus, an estimated final cost would roughly be between 180 and 220 euros per sample, depending on the desirable coverage and on the flow-cell that is used. Considering the improvement that is potentiated by a faster and accurate diagnostic, which will clearly result in proper therapeutic options and public health interventions, it is mandatory to perform a cost/benefit study in order to evaluate the usefulness of this approach. Also, and as it has already happened with WGS technology, it is expected a reduction in the costs of reagents in a near future. Overall, despite the important impact of the implementation of this approach in individual and public health and the quite promising results that we have obtained so far, more samples with heterogeneous *M. tuberculosis* loads will have

to be tested in order to obtain a more complete picture about its usefulness in laboratory practice for TB surveillance and diagnosis. We expect to provide results in the next couple of months.

Chapter VI

Identification of new mutations associated with decreased susceptibility to anti-TB drugs

On-going work

Rita Macedo, João Paulo Gomes.

RM contributed to the design of the study, performed the experimental work, interpreted data and wrote the chapter.

This study is being conducted after submission of the forms required by the Ethics Committee of the National Institute of Health Dr. Ricardo Jorge. All procedures are in accordance with the ethical standards and with the Helsinki Declaration, as revised in 2008.

6. Identification of new mutations associated with decreased susceptibility to anti-TB drugs

6.1 Introduction

Although the molecular mechanisms responsible for the resistance to the first and some of the second-line drugs were already studied and are well documented, the same does not happen with most of second and third-line anti-TB drugs, such as linezolide (LZD), PAS, cycloserine and the recently acquired bedaquiline (BDQ) and delamanid (DLM). They are still poorer investigated and, as such, the correlation between genotypic and phenotypic DST has very low sensitivity.

LZD, the first oxazolidinone approved for clinical use, has demonstrated excellent activity against drug-resistant strains of MTBC. LZD is rapidly absorbed after oral dosing. It readily distributes to well-perfused regions of the body, penetrates well into bronchoalveolar tissues, and acts by inhibiting protein synthesis at an early stage of translation (270). Recent case series have reported improvement among MDR-TB patients whose treatment regimens included LZD (271,272). However, the toxicities are severe, primarily including reversible myelosuppression and neuropathy, mainly determined by dose and duration, and require discontinuation of LZD in several cases (270). Mutations in the *rplC* and *rrl* target genes appear to be associated with either high or low-level LZD resistance respectively (273).

BDQ and DLM are the first new anti-TB drugs to become available in 40 years and have recently received a marketing approval by the Food and Drug Administration and European Medicine Agency (77,274).

BDQ-containing regimens increase by 12 times the probability of culture conversion in MDR-TB cases and prevent the emergence of further resistances to the drugs included in the backbone regimens. It has been shown to reduce the time to culture conversion in the first 6 months and the safety and tolerability profile seems to be good compared with other antituberculosis drugs (77,79).

BDQ it is a diarylquinoline with potent activity against *M. tuberculosis*, including *in vitro* induced dormant cells, which targets the subunit c of the ATP synthase (275–277). Although it has been more recently introduced, some evidences points out that resistance to this drug appears to emerge rapidly. Three genes are so far known to be involved in the emergence of resistance to BDQ: *atpE*, Rv0678 and *pepQ*. *atpE* codes for the BDQ target and mutations in this gene are thought to abrogate the binding of this drug to its target, resulting in high-level resistance (276).

DLM-containing regimens showed a short- and long-term efficacy in terms of culture conversion. The positive microbiological features are associated with the relevant improvement of a strong

epidemiological indicator- the reduction of mortality rates; the proportion of individuals who died after a ≥6 month´s exposure to DLM was 1% versus 8% in those not exposed (77). Also, the percentage of individuals who culture-converted at two months was about 45% versus nearly 30% in the control group (77).

DLM is a dihydro-nitroimidazooxazole derivative with a highly potent activity against either actively growing or dormant *M. tuberculosis* (278,279). As PZA or INH, DLM is a pro-drug that requires activation to its reactive form, which is catalysed by the deazaflavin(F420)-dependent nitroreductase (Ddn) and leads to the synthesis inhibition of methoxy-mycolic acids and keto-mycolic acids (278,280,281). As such, loss of Ddn function by mutations on *ddn* gene is one of the main genetic mechanisms leading to DLM resistance. Other resistance mechanisms include non-synonymous mutations in *fbiA*, *fbiB* and *fbiC*, coding for enzymes involved in the synthesis of F420 that produce non-functional forms thereby compromising DLM bioactivation (281).

Despite the identification of the above-cited genetic markers thought to be associated with the resistance to LZD, BDQ e DLM, since its regular use is only occurring in the present, more extensive research regarding the precise resistance mechanisms should be performed. As such, we aimed to disclose the genetic basis of resistance for these three drugs through the application of *in vitro* selective pressure scenarios, which have been accurately validated for other bacteria (282).

6.2 Materials and Methods

6.2.1. Bacterial strain and culture conditions

A fully susceptible MTBC strain was subjected to sequential sub-MIC concentrations of LZD, BDQ and DLM using 7H9 liquid broth in the MGIT system according to manufacturer´s instructions. We started with concentrations corresponding to 1:8 MIC (1,0 μg/μL for LZD, 0,06 μg/μL for DLM and 1,0 μg/μL for BDQ) (171) and, after observing the fully adaptation of the strain, i.e., when the time of growth seemed to be independent of the introduction of the antibiotic, we duplicated the dosage until the strains were fully adapted at drugs concentrations higher than the pre-determined MIC. The procedure was stopped when no increase in MIC values were observed after three weeks of growth. Each test was performed in triplicate and using a control (with no antibiotic) and the aliquots were kept frozen after growth for further WGS analysis. Periodically, DST was performed to the "supposedly mutant" strain to confirm its susceptibility

and conversion to resistance. At the end of these selection procedures, the culture was plated onto solid media, single colonies were chosen for susceptibility testing and independent mutants from each reference drug were selected for sequencing.

6.2.2 DNA extraction and quantification

The bacterial cells were first resuspended in TE 1X buffer (160μL) and lysozyme (40μL) and were incubated for 1h at 37ºC. After this lysis step, 180 μL lysis buffer (Bioline) and 20 μL proteinase K were added and incubated for 3h at 56ºC. The remaining protocol was performed according to manufacturer´s instructions using the Isolate II Genomic DNA kit (Bioline). Quantification and quality assessment of the purified DNA was performed using Qubit Fluorometer with hsDNA Assay Kit (Thermo Fisher Scientific) and agarose gel electrophoresis (0,8%), respectively.

6.2.3 WGS and genome assemblies

For each strain, WGS was performed as previously described [28]. Briefly, high-quality DNA samples were used to prepare dual-indexed Nextera XT Illumina libraries using the KAPA HiFi HotStart ReadyMix PCR Kit (KAPA Biosystems) in the indexing step to improve amplification of the GC-rich genome regions. Libraries were subsequently subjected to cluster generation and paired-end sequencing (2×250bp) on a MiSeq Illumina platform (Illumina Inc.), according to the manufacturer's instructions.

All genomes were *de novo* assembled using the INNUca v3.1 pipeline (https://github.com/B-UMMI/INNUca), which consists of integrated modules for reads QA/QC, *de novo* assembly and post-assembly optimization steps. Briefly, after reads' quality analysis (FastQC v0.11.5 - http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and cleaning (Trimmomatic v0.36) (240), genomes are assembled with SPAdes v3.11 (241) and subsequently improved using Pilon v1.18 (242).

6.3 Preliminary results and future perspectives

We started these experiments with LZD due to constraints regarding the acquisition of BDQ and DLM. So far, and after 9 months of sub-culturing, we have achieved a LZD dosage of 2xMIC, but the growth rate is still far from the expected if the cells were already fully adapted, so the experiments are still ongoing. The first passages were achieved in short periods of time, i.e., from the 1:8 MIC until 1:2 MIC we needed around 3 months; to achieve the MIC it took an extra 2 months and, since then, we do not seem to be able to grow easy a 2xMIC *M. tuberculosis* culture. Although not comparable to other microorganisms, such as *E. coli*, which in only few weeks can convert to a resistant phenotype 1000X MIC (282), we must acknowledge that the means of *in vitro* evolution and adaptation are pretty much the same (i.e., random emergence of point mutations). The major hurdle is the fact that *M. tuberculosis* cells typically take as much as 24h to divide, making it extremely difficult to perform these experiments. This disadvantage is indeed potentiated under selective pressure scenarios (as the one in the present study), as the bacteria must adapt to a hostile environment created by the presence of antibiotics, which certainly strongly affects the duplication time. Nevertheless, we have already selected colonies growing at 2X MIC for future WGS analysis and screening for possible existing mutations. Even if these emergent clones are not yet the best fitted, the identification of their mutation profile will certainly help the understanding of the evolutionary pathway that takes place until a well-adapted clone is established.

The experiments with DLM and BDQ started this early february, and, so far, we have not been able to surpass the 1:8 MIC.

Although really demanding, especially due to the constraints associated with *M. tuberculosis* species, we hope this work will contribute to the detection (and confirmation) of mutations associated with resistant phenotypes. This will contribute, not only to a better understanding of the resistance mechanisms, but also to the update of the mutations' databases used by the bioinformatics platforms (as the ones used in Chapter II of this thesis) that allow a rapid *in silico* screening of resistance targets.

Chapter VII

Final overview, concluding remarks and future directions

7. Final overview, concluding remarks and future directions

The main goal of this PhD dissertation was to acknowledge the potential of the use of WGS-based methodologies for routine diagnostic and epidemiological surveillance of *M. tuberculosis* strains and to develop specific pipelines to be implemented in the Portuguese NRL. As such, and given the major constraints regarding the isolation and time of growth of MTB strains, specially concerning the time to have a DST result, our first approach, as described in Chapter II, was to evaluate the efficiency of the available online-free software platforms for *in silic*o resistance prediction. We conducted a study involving a set of 54 M/XDR- and five susceptible-TB strains, corresponding to about 300 phenotypic hits, simultaneously evaluating the major four free online platforms, TB Profiler, PhyResSE, Mykrobe Predictor and TGS-TB. Overall, the sensitivity of resistance prediction ranged from 84.3% using Mykrobe predictor to 95.2% using TB profiler, while specificity was higher and homogeneous among platforms. TB profiler revealed the best performance robustness (sensitivity, specificity, PPV and NPV above 95%), and thus it was the adopted platform for use in the NRL at the NIH as a tool for first-line clinical guidance for therapeutic decision. We have also observed a few discrepancies between phenotype and genotype, where, in some cases, it was possible to pinpoint some "candidate" mutations highlighting the need for their confirmation through mutagenesis assays and potential review of the anti-TB genetic databases. In addition, and specially with the introduction of new therapeutic options, it is mandatory to conduct these experiments (as it was pointed out in the Chapter VI of this thesis) for the revision and inclusion of new possible genetic markers underlying *M. tuberculosis* resistance in the available databases, so that they can reunite all possible mutations underlying antibiotic resistance. This tremendous development of the bioinformatics algorithms and reduced time frame for reporting a result to the clinician will certainly trigger the technological transition where WGS-based bioinformatics platforms could replace phenotypic drug susceptibility testing for TB. Although, at this stage, we have only implemented this approach at the culture level, studies are ongoing in order to use it directly from clinical samples isolated from TB patients (as it is demonstrated in Chapter V). This study resulted already in a paper that was published in an international scientific journal, as mentioned above.

In chapter III, we studied the usefulness of the routinely and most widely used methodology for MTBC surveillance, MIRU-VNTR, in order to evaluate the rate of MDR-TB cases clustering and the possible association links that could underlie a direct transmission. We have evaluated the activity, since 2014, of specific reference centres for the diagnosis, consultancy, monitoring, and

treatment of the M/XDR-TB cases. Besides the clinical approach, these centres also link the information on molecular genotyping performed by the NRL with the epidemiological surveys performed by the Public Health Authorities. Although, with this analysis, we have observed a decreasing tendency both in the number of MDR-TB cases and the clustering rates, there is a poor agreement between laboratory and epidemiological data. The centralization of the MDR-TB cases in reference centres seems to be effective, but there is certainly a need for a better molecular tool, with higher discriminatory power, and a better inclusion of epidemiological data when discussing these clusters. This "better molecular tool" was already developed and described in Chapter IV of this thesis and it will be discussed in the next paragraphs. To better understand the transmission patterns from all culture-confirmed MDR-TB cases notified in Portugal from 2014 to 2017, we started a collaboration with a group from Universidade do Minho to deeper investigate this phenomenon. The results showed a unique MDR-TB transmission scenario, where MDR strains likely arose and are being transmitted within local chains despite the efforts from the Public Health authorities to identify and control all the cases. With 63% of these cases in cluster (using MIRU-VNTR molecular methodology), it appoints to a very high degree of primary transmission. An evolutionary analysis of these same strains suggested that clustered strains arose locally from endemic strains (still under investigation). While in Europe LAM strains are more common and Beijing sub-lineages are mostly associated with MDR-TB, in Portugal LAM is the common MTBC lineage clade but also associated with higher prevalence of MDR cases. One hypothesis is that MDR-TB in Portugal is evolving within an autochthonous chain of transmission. In fact, according to the MIRU-VNTR profiles, we have found clusters that have been previously identified (283,284) and associated with MDR-TB cases, such as cluster Q1, Lisboa3-A and Lisboa3-B, in the LTV region. Despite the efforts to track and contain MDR-TB strains in Portugal, their transmission remain to be disclosed, stressing the need to reinforce surveillance and Public Health containment strategies. This specific work is also part of a PhD thesis from a colleague from Universidade do Minho and will be soon submitted for publication.

As stated above, there is a need for a more accurate molecular surveillance tool that can allow a better match between the laboratory and the epidemiological information. For this reason, in Chapter IV, we aimed to report the implementation of a dynamic gene-by-gene approach, fully relying on freely available software, for prospective WGS-based tuberculosis surveillance. It is well established that WGS offers unprecedented resolution for tracking MTBC transmission and antibiotic-resistance prediction, however, standardized pipelines and the definition of epidemiological cut-offs for cluster detection are still under discussion. Our implemented approach and its application for detecting transmission chains was demonstrated by

retrospectively analysing all M/XDR strains isolated between 2013 and 2017 in Portugal. We observed a good correlation between genetic relatedness and epidemiological links, obtaining mean pairwise allele differences (AD) below 0.3% for strongly epi-linked clusters. We observed the same scenario by applying the core-SNV analysis, while providing higher resolution and epidemiological concordance than MIRU-VNTR genotyping. In addition, "zooming in" each strains' sub-set (i.e., increasing the number of shared *loci* within each sub-set) also strengthens the confidence in detecting epi-linked clusters. This gene-by-gene strategy offers several practical benefits (e.g., amenability to standardization, reliance on freely-available software, scalability and low computational requirements) and our data further consolidated it as the method of choice for a timely, standardized and robust prospective WGS-based laboratory surveillance of M/XDR-TB cases.

In Chapter V, we intended to overcome the limitations of performing our WGS-based approaches only after isolation of MTBC strains in culture. As such, we aimed at capturing the genomes of MTBC directly from the sputum positive clinical samples, without the need for culture propagation. To achieve this, we bioinformatically designed RNA baits that span the entire genome, and can recover (by hybridization) low copy numbers of DNA directly from the samples with sufficiently high sensitivity and specificity to enable efficient WGS. Although, at this stage, we simply aimed at establishing a proof-of-concept, the results were quite surprising in the sense that we were able to recover the full MTBC genomes from all tested sputum samples. In fact, the percentage of dedicated reads to this bacterium ranged from 56,7% to 96,7% and all samples had 99% of their genome covered by, at least, one read. Surprisingly, these results are comparable with the ones obtained with pure cultures, raising tremendous expectations about its applicability in the routine practice of the NRL. We anticipate that, if the extension of this study to many more samples yields similar results, this approach shows-up as a cutting-edge laboratory procedure for determination of the antibiotic resistance profiles and even for epidemiological surveillance. The potential decrease in the time-to-results to about five days after receiving the biological sample will constitute a hallmark in the TB diagnosis practice, as it would yield unequivocal gains both in the individual and in the Public Health strategies.

Finally, in chapter VI, we aimed to disclose the genetic basis of resistance for the newest introduced drugs for TB treatment, LZD, BDQ and DLM, through the application of *in vitro* selective pressure scenarios. It has taken many years to create a database with most of the mutations associated with the susceptibility decrease to the currently used TB drugs. Even though it is believed that many mutations remain to be disclosed, as several studies have pointed additional putative genetic markers of resistance that not belong to the available databases, this issue becomes even more relevant for the drugs that were introduced in the

treatment procedures only very recently. As such, we used *in vitro* selective pressure scenarios, for which the proof-of-concept for this purpose was already demonstrated for some bacteria, to disclose the genetic basis of resistance to these drugs. This approach consisted on the propagation of a fully susceptible MTBC strain using increasing dosages (starting with sub-MIC concentrations) of each drug until it fully adapts to concentrations two or three times higher than the MICs defined for DST analysis. This study is still ongoing and the time when the first resistant clones will emerge cannot be predicted at this stage.

Overall, we believe this PhD dissertation contributes for a better understanding of how WGS-based methodologies can surpass the difficulties of TB diagnosis and surveillance and, in particular, the ability to provide the clinicians with a much more rapid information regarding resistance prediction and an eventual transmission chain.

As immediate future perspectives, we will continue the work underlying the use of WGS directly on clinical samples in order to be fully validated and implemented as routine as the culture WGS-based approach already is. Additionally, the propagation of the strains using antibiotic selective pressure is also ongoing with the ultimate goal of the enrichment of the publicly available databases of genetic markers of resistance.

References

References

1. KGaA WVG& C, editor. Handbook of tuberculosis. 2008.
2. Wayne, L. G. and GPK. Bergey's Manual of Systematic Bacteriology. P. H. A. Sneath, N. S. Mair, M. E. Sharpe and JGH, editor. Williams & Wilkins; 1986. 1436-1457 p.
3. Goodfellow, M. and TC. The Biology of the Actinomycetes. M. Goodfellow, M. Mordarski and STW, editor. London: Academic Press; 1984. 8-164 p.
4. David HL. The Mycobacteria: A Source Book, Part A. Wayne GPK and LG, editor. New york: Marcel Dekker; 1984. 537-545 p.
5. Wayne LG. The Mycobacteria: A Source Book, Part A. G. P. Kubica and L. G. Wayne, editor. New york: Marcel Dekker; 1984. 25-65 p.
6. Skerman, V. B. D., V. McGowan and PHAS. Approved lists of bacterial names. Int J Syst Bacteriol. 1980;30:225–420.
7. World Health Organization. Buruli ulcer (*Mycobacterium ulcerans* infection) . 2013. Available from: http://www.who.int/mediacentre/factsheets/fs199/en/index.html
8. Hermon-Taylor J and FE-Z. The *Mycobacterium avium* subspecies paratuberculosis problem and its relation to the causation of Crohn disease. Publishing I, editor. London; 2004.
9. World Health Organization. Leprosy. . 2012. Available from: http://www.who.int/mediacentre/factsheets/fs101/en/. 101.
10. Medjahed H, Gaillard J-L, Reyrat J-M. *Mycobacterium abscessus*: a new player in the mycobacterial field. Trends Microbiol . 2010 Mar;18(3):117–23. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0966842X09002637
11. Bentley, S. D., I. Comas, J. M. Bryant, D. Walker, N. H. Smith, S. R. Harris, S. Thurston, S. Gagneux, J. Wood, M. Antonio, M. A. Quail, F. Gehre, R. A. Adegbola JP and BC de J. The genome of *Mycobacterium africanum* West African 2 reveals a lineage-specific locus and genome erosion common to the *M. tuberculosis* complex. PLoS Negl Trop Dis. 2012;6(2):e1552.
12. Supply P, Marceau M, Mangenot S, Roche D, Rouanet C, Khanna V, et al. Genomic analysis of smooth tubercle bacilli provides insights into ancestry and pathoadaptation of *Mycobacterium tuberculosis*. Nat Genet . 2013 Feb 6;45(2):172–9. Available from: http://www.nature.com/articles/ng.2517
13. Alexander KA, Laver PN, Michel AL, Williams M, van Helden PD, Warren RM, et al. Novel *mycobacterium tuberculosis* complex pathogen, *M. Mungi*. Emerg Infect Dis . 2010 Aug ;16(8):1296–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20678329

14. Parsons SDC, Drewe JA, Gey van Pittius NC, Warren RM, van Helden PD. Novel Cause of Tuberculosis in Meerkats, South Africa. Emerg Infect Dis . 2013 Dec ;19(12):2004–7. Available from: http://wwwnc.cdc.gov/eid/article/19/12/13-0268_article.htm

15. van Ingen J, Rahim Z, Mulder A, Boeree MJ, Simeone R, Brosch R, et al. Characterization of *Mycobacterium orygis* as *M. tuberculosis* complex subspecies. Emerg Infect Dis . 2012 Apr ;18(4):653–5. Available from: http://wwwnc.cdc.gov/eid/article/18/4/11-0888_article.htm

16. Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, et al. Deciphering the biology of *mycobacterium tuberculosis* from the complete genome sequence. Nature. 1998;393(6685):537–44.

17. Brosch R, Gordon S V., Marmiesse M, Brodin P, Buchrieser C, Eiglmeier K, et al. A new evolutionary scenario for the *Mycobacterium tuberculosis* complex. Proc Natl Acad Sci . 2002 Mar 19 ;99(6):3684–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/11891304

18. Hershberg R, Lipatov M, Small PM, Sheffer H, Niemann S, Homolka S, et al. High Functional Diversity in *Mycobacterium tuberculosis* Driven by Genetic Drift and Human Demography. Blaser MJ, editor. PLoS Biol . 2008 Dec 16 ;6(12):e311. Available from: https://dx.plos.org/10.1371/journal.pbio.0060311

19. Harris SR, Berg S, Luo T, Thwaites G, Bothamley G, Aseffa A, et al. Out-of-Africa migration and Neolithic coexpansion of *Mycobacterium tuberculosis* with modern humans. Nat Genet . 2013 Oct 1 ;45(10):1176–82. Available from: http://www.nature.com/articles/ng.2744

20. Comas I, Gagneux S. The past and future of tuberculosis research . Manchester M, editor. Vol. 5, PLoS Pathogens. 2009 . p. e1000600. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19855821

21. Bentley SD, Horstmann RD, Drobniewski F, Harris SR, Nikolayevskyy V, Casali N, et al. Microevolution of extensively drug-resistant tuberculosis in Russia. Genome Res . 2012 Apr 1 ;22(4):735–45. Available from: http://genome.cshlp.org/cgi/doi/10.1101/gr.128678.111

22. van Soolingen D, Qian L, de Haas PE, Douglas JT, Traore H, Portaels F, et al. Predominance of a single genotype of *Mycobacterium tuberculosis* in countries of east Asia. J Clin Microbiol . 1995 Dec ;33(12):3234–8. Available from: http://www.ncbi.nlm.nih.gov/pubmed/8586708

23. Eldholm V, Balloux F. Antimicrobial Resistance in *Mycobacterium tuberculosis*: The Odd One Out . Vol. 24, Trends in Microbiology. 2016 . p. 637–48. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0966842X16000767

24. Dos Vultos T, Mestre O, Rauzier J, Golec M, Rastogi N, Rasolofo V, et al. Evolution and diversity of clonal bacteria: The paradigm of *Mycobacterium tuberculosis*. Ahmed N,

editor. PLoS One . 2008 Feb 6 ;3(2):e1538. Available from: https://dx.plos.org/10.1371/journal.pone.0001538

25. Bos KI, Harkins KM, Herbig A, Coscolla M, Weber N, Comas I, et al. Pre-Columbian mycobacterial genomes reveal seals as a source of New World human tuberculosis. Nature . Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.; 2014 Oct 20;514(7523):494–7. Available from: https://doi.org/10.1038/nature13591

26. Orgeur M, Brosch R. Evolution of virulence in the *Mycobacterium tuberculosis* complex. Curr Opin Microbiol . 2018 Feb ;41:68–75. Available from: http://www.ncbi.nlm.nih.gov/pubmed/29216510

27. Hershkovitz I, Donoghue HD, Minnikin DE, Besra GS, Lee OY-C, Gernaey AM, et al. Detection and molecular characterization of 9000-year-old *Mycobacterium tuberculosis* from a neolithic settlement in the Eastern mediterranean. Ahmed N, editor. PLoS One . 2008 Oct 15 ;3(10):e3426. Available from: https://dx.plos.org/10.1371/journal.pone.0003426

28. Daniel TM. The history of tuberculosis. Respir Med . 2006 Nov ;100(11):1862–70. Available from: http://www.ncbi.nlm.nih.gov/pubmed/16949809

29. Smith I. *Mycobacterium tuberculosis* pathogenesis and molecular determinants of virulence . Vol. 16, Clinical Microbiology Reviews. 2003 . p. 463–96. Available from: http://www.ncbi.nlm.nih.gov/pubmed/12857778

30. Zumla A, Mwaba P, Huggett J, Kapata N, Chanda D, Grange J. Reflections on the white plague . Vol. 9, The Lancet Infectious Diseases. 2009 . p. 197–202. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1473309909700453

31. WHO. Global tuberculosis report 2018 . 2018. Available from: https://www.who.int/tb/publications/global_report/en/

32. Young DB, Perkins MD, Duncan K, Barry CE. Confronting the scientific obstacles to global control of tuberculosis . Vol. 118, Journal of Clinical Investigation. 2008 . p. 1255–65. Available from: http://www.jci.org/articles/view/34614

33. Hopewell PC BB. Respiratory Medicine. 2nd ed. Murray JF NJ, editor. Philadelphia: WB Saunders Company; 1994. 1094-1160 p.

34. Sutherland I. Recent studies in the epidemiology of tuberculosis, based on the risk of being infected with tubercle bacilli. Adv Tuberc Res . 1976 ;19:1–63. Available from: http://www.ncbi.nlm.nih.gov/pubmed/823803

35. Marais BJ, Gie RP, Schaaf HS, Hesseling AC, Obihara CC, Starke JJ, et al. The natural history of childhood intra-thoracic tuberculosis: a critical review of literature from the pre-chemotherapy era. Int J Tuberc Lung Dis . 2004 Apr ;8(4):392–402. Available from: http://www.ncbi.nlm.nih.gov/pubmed/15141729

36. Leung CC, Li T, Lam TH, Yew WW, Law WS, Tam CM, et al. Smoking and tuberculosis among the elderly in Hong Kong. Am J Respir Crit Care Med . 2004 Nov 1 ;170(9):1027–33. Available from: http://www.atsjournals.org/doi/abs/10.1164/rccm.200404-512OC

37. Leung CC, Lam TH, Ho KS, Yew WW, Tam CM, Chan WM, et al. Passive smoking and tuberculosis. Arch Intern Med . 2010 Feb 8 ;170(3):287–92. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20142576

38. Leung CC, Lam TH, Chan WM, Yew WW, Ho KS, Leung G, et al. Lower risk of tuberculosis in obesity. Arch Intern Med . 2007 Jun 25 ;167(12):1297–304. Available from: http://www.ncbi.nlm.nih.gov/pubmed/17592104

39. Chan CK, Leung GM, Chang KC, Leung CC, Chan WM, Ho KS, et al. Diabetic Control and Risk of Tuberculosis: A Cohort Study. Am J Epidemiol . 2008 Apr 29 ;167(12):1486–94. Available from: http://www.ncbi.nlm.nih.gov/pubmed/18400769

40. Daley CL, Hahn JA, Moss AR, Hopewell PC, Schecter GF. Incidence of tuberculosis in injection drug users in San Francisco: impact of anergy. Am J Respir Crit Care Med . 1998 Jan ;157(1):19–22. Available from: http://www.atsjournals.org/doi/abs/10.1164/ajrccm.157.1.9701111

41. Selwyn PA, Sckell BM, Alcabes P, Friedland GH, Klein RS, Schoenbaum EE. High risk of active tuberculosis in HIV-infected drug users with cutaneous anergy. JAMA . ;268(4):504–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/1619742

42. Small PM, Shafer RW, Hopewell PC, Singh SP, Murphy MJ, Desmond E, et al. Exogenous reinfection with multidrug-resistant *Mycobacterium tuberculosis* in patients with advanced HIV infection. N Engl J Med . 1993 Apr 22 ;328(16):1137–44. Available from: http://www.nejm.org/doi/abs/10.1056/NEJM199304223281601

43. Philips JA, Ernst JD. Tuberculosis Pathogenesis and Immunity. Annu Rev Pathol Mech Dis . 2011 Feb 28 ;7(1):353–84. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22054143

44. Kaufmann SHE. How can immunology contribute to the control of tuberculosis? . Vol. 1, Nature Reviews Immunology. 2001 . p. 20–30. Available from: http://www.nature.com/articles/nri35095558

45. Knechel NA. Tuberculosis: Pathophysiology, clinical features, and diagnosis. Crit Care Nurse . 2009 Apr 1 ;29(2):34–43. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19339446

46. Holscher C, Solovic I, Zellweger JP, Milburn H, Cardona PJ, Bossink A, et al. LTBI: latent tuberculosis infection or lasting immune responses to *M. tuberculosis*? A TBNET consensus statement. Eur Respir J . 2009 May 1 ;33(5):956–73. Available from: http://erj.ersjournals.com/cgi/doi/10.1183/09031936.00120908

47. Dye C, Williams BG. The population dynamics and control of tuberculosis . Vol. 328, Science. 2010 . p. 856–61. Available from: http://www.sciencemag.org/cgi/doi/10.1126/science.1185449

48. Zumla A, Raviglione M HR. Current Concepts: Tuberculosis. N Engl J Med . 2013 Nov 13 ;259(20):976–8. Available from: http://www.ncbi.nlm.nih.gov/pubmed/13600608

49. Lawn SD, Zumla AI. Tuberculosis. Lancet . 2011 Jul 2 ;378(9785):57–72. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0140673610621733

50. Tiemersma EW, van der Werf MJ, Borgdorff MW, Williams BG, Nagelkerke NJD. Natural history of tuberculosis: Duration and fatality of untreated pulmonary tuberculosis in HIV negative patients: A systematic review . Pai M, editor. Vol. 6, PLoS ONE. 2011 . p. e17601. Available from: https://dx.plos.org/10.1371/journal.pone.0017601

51. Ray S, Talukdar A, Kundu S, Khanra D, Sonthalia N, Etyang AO, et al. Diagnosis and management of miliary tuberculosis: current state and future perspectives Medical causes of admissions to hospital among adults in Africa: a systematic review. Ther Clin Risk Manag . 2013 Jan ;9:9–26. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23326198

52. Rodrigo T, Caylà JA, García de Olalla P, Galdós-Tangüis H, Jansà JM, Miranda P, et al. Characteristics of tuberculosis patients who generate secondary cases. Int J Tuberc Lung Dis . 1997 Aug ;1(4):352–7. Available from: http://www.ncbi.nlm.nih.gov/pubmed/9432392

53. Riley RL, Mills CC, Nyka W, Weinstock N, Storey PB, Sultan LU, et al. Aerial dissemination of pulmonary tuberculosis. Am Rev Tuberc . 1957 Dec ;76(6):931–41. Available from: http://www.ncbi.nlm.nih.gov/pubmed/13488004

54. Leung CC, Lange C, Zhang Y. Tuberculosis: Current state of knowledge: An epilogue . Vol. 18, Respirology. 2013 . p. 1047–55. Available from: http://doi.wiley.com/10.1111/resp.12156

55. Han J, Kwon OJ, Yi CA, Lee JY, Lee KS, Kim TS, et al. CT Scan Features as Predictors of Patient Outcome After Bronchial Intervention in Endobronchial TB. Chest . 2010 Aug ;138(2):380–5. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0012369210604217

56. Skoura E, Zumla A, Bomanji J. Imaging in tuberculosis. Int J Infect Dis . 2015 Mar ;32:87–93. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25809762

57. Davis JL, Cattamanchi A, Cuevas LE, Hopewell PC, Steingart KR. Diagnostic accuracy of same-day microscopy versus standard microscopy for pulmonary tuberculosis: A systematic review and meta-analysis. Lancet Infect Dis . 2013 Feb ;13(2):147–54. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23099183

58. Steingart KR, Henry M, Ng V, Hopewell PC, Ramsay A, Cunningham J, et al. Fluorescence versus conventional sputum smear microscopy for tuberculosis: a systematic review .

Vol. 6, Lancet Infectious Diseases. 2006 . p. 570–81. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1473309906705783

59. Al-Aghbari N, Anderson de Cuevas RM, Aseffa A, Ramsay A, Emenyonu EN, Faragher B, et al. LED Fluorescence Microscopy for the Diagnosis of Pulmonary Tuberculosis: A Multi-Country Cross-Sectional Evaluation. Wilson D, editor. PLoS Med . 2011 Jul 12 ;8(7):e1001057. Available from: https://dx.plos.org/10.1371/journal.pmed.1001057

60. Wayne LG. Microbiology of tubercle bacilli. Am Rev Respir Dis . 1982 Mar ;125(3 Pt 2):31–41. Available from: http://www.ncbi.nlm.nih.gov/pubmed/6803631

61. Dunlap NE, Bass J, Fujiwara P, Hopewell P, Horsburgh CR, Salfinger M, et al. Diagnostic standards and classification of tuberculosis in adults and children . Vol. 161, American Journal of Respiratory and Critical Care Medicine. 2000 . p. 1376–95. Available from: http://www.ncbi.nlm.nih.gov/pubmed/10764337

62. Steingart KR, Schiller I, Horne DJ, Pai M, Boehme CC, Dendukuri N. Xpert® MTB/RIF assay for pulmonary tuberculosis and rifampicin resistance in adults . Vol. 2014, Cochrane Database of Systematic Reviews. 2014 . p. CD009593. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24448973

63. Richardson M, Warren R, Donegan S, Peter J, Dheda K, Theron G, et al. The diagnostic accuracy of the GenoType ® MTBDR sl assay for the detection of resistance to second-line anti-tuberculosis drugs. In: Theron G, editor. Cochrane Database of Systematic Reviews . Chichester, UK: John Wiley & Sons, Ltd; 2014 . p. CD010705. Available from: http://doi.wiley.com/10.1002/14651858.CD010705.pub2

64. WHO. Automated Real-Time Nucleic Acid Amplification Technology for Rapid and Simultaneous Detection of Tuberculosis and Rifampicin Resistance: Xpert MTB/RIF Assay for the Diagnosis of Pulmonary and Extrapulmonary TB in Adults and Children: Policy Update. 2013;

65. National Committee for Clinical Laboratory Standards(NCCLS). Susceptibility testing for mycobacteria, Nocardia, and other aerobic actinomycetes. Wayne P, editor. 2000.

66. Zhang H, Li D, Zhao L, Fleming J, Lin N, Wang T, et al. Genome sequencing of 161 *Mycobacterium tuberculosis* isolates from China identifies genes and intergenic regions associated with drug resistance. Nat Genet . 2013 Oct 1 ;45(10):1255–60. Available from: http://www.nature.com/articles/ng.2735

67. Rosenblatt MB. Pulmonary tuberculosis: evolution of modern therapy. Bull N Y Acad Med . 1973 Mar ;49(3):163–96. Available from: http://www.ncbi.nlm.nih.gov/pubmed/4572586

68. Sakula A. Carlo Forlanini, inventor of artificial pneumothorax for treatment of pulmonary tuberculosis. Thorax . 1983 May ;38(5):326–32. Available from: http://www.ncbi.nlm.nih.gov/pubmed/6348993

69. Dheda K, Migliori GB. The global rise of extensively drug-resistant tuberculosis: Is the time to bring back sanatoria now overdue? Vol. 379, The Lancet. 2012 . p. 773–5. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0140673611610623

70. VC B. In Chemotherapy of tuberculosis. VC B, editor. London; 1964.

71. Migliori GB, Sotgiu G, Centis R, Grzemska M, Falzon D, Getahun H RM. Antituberculosis therapy and current global guidelines. In Current & emerging diagnostics, therapeutics & vaccines for tuberculosis. Future Med. SHE K, editor. London; 2011. 42-63 p.

72. Sotgiu G, Centis R, D'Ambrosio L MG. Medical treatment of pulmonary tuberculosis. Eur Respir. 2013. 11-19 p.

73. Schatz A, Bugie E WS. Streptomycin, a substance exhibiting antibiotic activity against gram-positive and gram-negative bacteria. ProcSoc Exp Biol Med. 1944 . p. 66–9. Available from: https://www.ncbi.nlm.nih.gov/pubmed/?term=Streptomycin%2C+a+substance+exhibiting+antibiotic+activity+against+Gram-positive+and+Gram-negative+bacteria

74. Crofton J. Some principles in the chemotherapy of bacterial infections. II. Br Med J . 1969 Apr 26 ;2(5651):209–12. Available from: http://www.ncbi.nlm.nih.gov/pubmed/5780427

75. Council EAMR. Controlled clinical treatment of short course (6 months) regime of chemotherapy for treatment of pulmonary tuberculosis. Third report. Lancet. 1974;2:237–48.

76. Sensi P. History of the development of rifampin. Rev Infect Dis . ;5 Suppl 3:S402-6. Available from: http://www.ncbi.nlm.nih.gov/pubmed/6635432

77. Sotgiu G, Centis R, D'Ambrosio L, Battista Migliori G. Tuberculosis treatment and drug regimens. Cold Spring Harb Perspect Med . 2015 May 1 ;5(5):a017822–a017822. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25573773

78. WHO Geneva. Toman's tuberculosis: Case detection, treatment and monitoring, second edition. 2004.

79. Sotgiu G, Migliori GB. Facing multi-drug resistant tuberculosis. Pulm Pharmacol Ther . Elsevier Ltd; 2015;32:144–8. Available from: http://dx.doi.org/10.1016/j.pupt.2014.04.006

80. Raviglione MC, Smith IM. XDR Tuberculosis — Implications for Global Public Health. N Engl J Med . 2007 Feb 15 ;356(7):656–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/17301295

81. Dheda K, Warren R, Mastrapa B, Badri M, Pietersen E, Sirgel FA, et al. Long-term outcomes of patients with extensively drug-resistant tuberculosis in South Africa: a cohort study. Lancet . 2014 Apr 5 ;383(9924):1230–9. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0140673613626756

82. Klopper M, Hayes C, Gey van Pittius NC, Trollip AP, Sirgel FA, Chabula-Nxiweni M, et al. Emergence and Spread of Extensively and Totally Drug-Resistant Tuberculosis, South Africa. Emerg Infect Dis . 2013 Mar ;19(3):449–55. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23622714

83. WHO. WHO best-practice statement on the off-label use of bedaquiline and delamanid for the treatment of multidrug-resistant tuberculosis. 2017.

84. David HL. Probability distribution of drug-resistant mutants in unselected populations of *Mycobacterium tuberculosis*. Appl Microbiol . 1970 Nov ;20(5):810–4. Available from: http://www.ncbi.nlm.nih.gov/pubmed/4991927

85. Council BMR. VARIOUS combinations of isoniazid with streptomycin or with P.A.S. in the treatment of pulmonary tuberculosis; seventh report to the Medical Research Council by their Tuberculosis Chemotherapy Trials Committee. Br Med J . 1955 Feb 19 ;1(4911):435–45. Available from: http://www.ncbi.nlm.nih.gov/pubmed/13230514

86. Perdigão J, Portugal I. Genetics and roadblocks of drug resistant tuberculosis. Infect Genet Evol . 2018 Sep 24 [cited 2019 Mar 8]; Available from: https://linkinghub.elsevier.com/retrieve/pii/S1567134818307366

87. Hausner TP, Geigenmüller U, Nierhaus KH. The allosteric three-site model for the ribosomal elongation cycle. New insights into the inhibition mechanisms of aminoglycosides, thiostrepton, and viomycin. J Biol Chem . 1988 Sep 15 [cited 2019 Mar 8];263(26):13103–11. Available from: http://www.ncbi.nlm.nih.gov/pubmed/2843509

88. Rastogi N, Labrousse V, Goh KS. In vitro activities of fourteen antimicrobial agents against drug susceptible and resistant clinical isolates of *Mycobacterium tuberculosis* and comparative intracellular activities against the virulent H37Rv strain in human macrophages. Curr Microbiol . 1996 Sep [cited 2019 Mar 8];33(3):167–75. Available from: http://www.ncbi.nlm.nih.gov/pubmed/8672093

89. Peloquin CA, Berning SE, Nitta AT, Simone PM, Goble M, Huitt GA, et al. Aminoglycoside Toxicity: Daily versus Thrice-Weekly Dosing for Treatment of Mycobacterial Diseases. Clin Infect Dis . 2004 Jun 1 [cited 2019 Mar 8];38(11):1538–44. Available from: https://academic.oup.com/cid/article-lookup/doi/10.1086/420742

90. Meier A, Sander P, Schaper KJ, Scholz M, Böttger EC. Correlation of molecular resistance mechanisms and phenotypic resistance levels in streptomycin-resistant *Mycobacterium tuberculosis*. Antimicrob Agents Chemother . 1996 Nov [cited 2019 Mar 8];40(11):2452–4. Available from: http://www.ncbi.nlm.nih.gov/pubmed/8913445

91. Cambau E, Viveiros M, Machado D, Raskine L, Ritter C, Tortoli E, et al. Revisiting susceptibility testing in MDR-TB by a standardized quantitative phenotypic assessment in a European multicentre study. J Antimicrob Chemother . 2015 Mar 1 [cited 2019 Mar 8];70(3):686–96. Available from: https://academic.oup.com/jac/article-lookup/doi/10.1093/jac/dku438

92. Feuerriegel S, Oberhauser B, George A, 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 May 30 [cited 2019 Mar 8];12(1):90. Available from: http://bmcmicrobiol.biomedcentral.com/articles/10.1186/1471-2180-12-90

93. Martin A, Ribeiro MO, Palomino JC, da Silva PEA, Zaha A, Ribeiro AW, et al. Streptomycin Resistance and Lineage-Specific Polymorphisms in *Mycobacterium tuberculosis gidB* Gene. J Clin Microbiol . 2011 Jul [cited 2019 Mar 8];49(7):2625–30. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21593257

94. Perdigão J, Macedo R, Silva C, Machado D, Couto I, Viveiros M, et al. From multidrug-resistant to extensively drug-resistant tuberculosis in Lisbon, Portugal: The stepwise mode of resistance acquisition. J Antimicrob Chemother. 2013;68(1).

95. 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;20(5).

96. Heym B, Zhang Y, Poulet S, Young D, Cole ST. Characterization of the *katG* gene encoding a catalase-peroxidase required for the isoniazid susceptibility of *Mycobacterium tuberculosis*. J Bacteriol . 1993 Jul ;175(13):4255–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/8320241

97. Bardou F, Raynaud C, Ramos C, Lanéelle MA, Lanéelle G. Mechanism of isoniazid uptake in *Mycobacterium tuberculosis*. Microbiology . 1998 Sep 1 ;144(9):2539–44. Available from: http://www.ncbi.nlm.nih.gov/pubmed/9782502

98. Heifets LB, Lindholm-Levy PJ, Flory M. Comparison of bacteriostatic and bactericidal activity of isoniazid and ethionamide against *Mycobacterium avium* and *Mycobacterium tuberculosis*. Am Rev Respir Dis . 1991 Feb ;143(2):268–70. Available from: http://www.atsjournals.org/doi/abs/10.1164/ajrccm/143.2.268

99. Viveiros M, Bettencourt R, Victor TC, Jordaan AM, Leandro C, Ordway D, et al. Isoniazid-Induced Transient High-Level Resistance in *Mycobacterium tuberculosis*. Antimicrob Agents Chemother. 2002;46(9):2804–10.

100. 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 May [cited 2019 Mar 8];8(3):521–4. Available from: http://www.ncbi.nlm.nih.gov/pubmed/8392139

101. Zhang Y, Heym B, Allen B, Young D, Cole S. The catalase—peroxidase gene and isoniazid resistance of *Mycobacterium tuberculosis*. Nature . 1992 Aug 13 [cited 2019 Mar 8];358(6387):591–3. Available from: http://www.ncbi.nlm.nih.gov/pubmed/1501713

102. Afanas'ev M V, Ikryannikova LN, Il'ina EN, Sidorenko S V, Kuz'min A V, Larionova EE, et al. Molecular characteristics of rifampicin- and isoniazid-resistant *Mycobacterium*

tuberculosis isolates from the Russian Federation. J Antimicrob Chemother . 2007 Jun 1 [cited 2019 Mar 8];59(6):1057–64. Available from: http://academic.oup.com/jac/article/59/6/1057/713130/Molecular-characteristics-of-rifampicin-and

103. Campbell PJ, Morlock GP, Sikes RD, Dalton TL, Metchock B, Starks AM, et al. Molecular detection of mutations associated with first- and second-line drug resistance compared with conventional drug susceptibility testing of *Mycobacterium tuberculosis*. Antimicrob Agents Chemother . 2011 May [cited 2019 Mar 8];55(5):2032–41. Available from: http://aac.asm.org/lookup/doi/10.1128/AAC.01550-10

104. Zhang M, Yue J, Yang YP, Zhang HM, Lei JQ, Jin RL, et al. Detection of mutations associated with isoniazid resistance in *Mycobacterium tuberculosis* isolates from China. J Clin Microbiol . 2005 Nov 1 [cited 2019 Mar 8];43(11):5477–82. Available from: http://www.ncbi.nlm.nih.gov/pubmed/16272473

105. Müller B, Streicher EM, Hoek KGP, Tait M, Trollip A, Bosman ME, et al. *inhA* promoter mutations: A gateway to extensively drug-resistant tuberculosis in South Africa? Int J Tuberc Lung Dis . 2011 Mar [cited 2019 Mar 8];15(3):344–51. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21333101

106. Perdigão J, MacEdo R, João I, Fernandes E, Brum L, Portugal I. Multidrug-resistant tuberculosis in Lisbon, Portugal: A molecular epidemiological perspective. Microb Drug Resist. 2008;14(2).

107. Vall-Spinosa A, Lester W, Moulding T, Davidson PT, McClatchy JK. Rifampin in the Treatment of Drug-Resistant *Mycobacterium tuberculosis* Infections. N Engl J Med . 1970 Sep 17;283(12):616–21. Available from: http://www.ncbi.nlm.nih.gov/pubmed/4988918

108. Wehrli W, Knüsel F, Schmid K, Staehelin M. Interaction of rifamycin with bacterial RNA polymerase. Proc Natl Acad Sci U S A . 1968 Oct;61(2):667–73. Available from: http://www.ncbi.nlm.nih.gov/pubmed/4879400

109. Campbell EA, Korzheva N, Mustaev A, Murakami K, Nair S, Goldfarb A, et al. Structural mechanism for rifampicin inhibition of bacterial rna polymerase. Cell . 2001 Mar 23;104(6):901–12. Available from: http://www.ncbi.nlm.nih.gov/pubmed/11290327

110. Stottmeier KD, Kubica GP, Woodley CL. Antimycobacterial activity of rifampin under in vitro and simulated in vivo conditions. Appl Microbiol . 1969;17(6):861–5. Available from: http://www.ncbi.nlm.nih.gov/pubmed/4978926

111. Mitchison DA. Basic mechanisms of chemotherapy. Chest . 1979 Dec ;76(6 Suppl):771–81. Available from: http://www.ncbi.nlm.nih.gov/pubmed/92392

112. Rusch-Gerdes S, Heep M, Niemann S, Richter E, Brandstatter B, Rieger U, et al. Frequency of rpoB Mutations Inside and Outside the Cluster I Region in Rifampin-Resistant Clinical *Mycobacterium tuberculosis* Isolates. J Clin Microbiol . 2002 Jan

1;39(1):107–10. Available from: http://jcm.asm.org/cgi/doi/10.1128/JCM.39.1.107-110.2001

113. Kapur V, Li LL, Iordanescu S, Hamrick MR, Wanger A, Kreiswirth BN, et al. Characterization by automated DNA sequencing of mutations in the gene (*rpoB*) encoding the RNA polymerase beta subunit in rifampin-resistant *Mycobacterium tuberculosi*s strains from New York City and Texas. J Clin Microbiol . 1994 Apr;32(4):1095–8. Available from: http://www.ncbi.nlm.nih.gov/pubmed/8027320

114. Herrera L, Jiménez S, Valverde A, García-Aranda MA, Sáez-Nieto JA. Molecular analysis of rifampicin-resistant *Mycobacterium tuberculosis* isolated in Spain (1996-2001). Description of new mutations in the *rpoB* gene and review of the literature. Int J Antimicrob Agents . 2003 May; 21(5):403–8. Available from: http://www.ncbi.nlm.nih.gov/pubmed/12727071

115. Bakonyte D, Baranauskaite A, Cicenaite J, Sosnovskaja A, Stakenas P. Molecular characterization of isoniazid-resistant *Mycobacterium tuberculosis* clinical isolates in Lithuania. Antimicrob Agents Chemother . 2003 Jun; 47(6):2009–11. Available from: http://www.ncbi.nlm.nih.gov/pubmed/12760887

116. Campbell PJ, Morlock GP, Sikes RD, Dalton TL, Metchock B, Starks AM, et al. Molecular Detection of Mutations Associated with First- and Second-Line Drug Resistance Compared with Conventional Drug Susceptibility Testing of *Mycobacterium tuberculosis*. Antimicrob Agents Chemother. 2011 May; 55(5):2032–41. Available from: http://aac.asm.org/lookup/doi/10.1128/AAC.01550-10

117. Lee ASG, Lim IHK, Tang LLH, Wong SY. High Frequency of Mutations in the *rpoB* Gene in Rifampin-Resistant Clinical Isolates of *Mycobacterium tuberculosis* from Singapore. J Clin Microbiol . 2005 Apr 1; 43(4):2026–7. Available from: http://jcm.asm.org/cgi/doi/10.1128/JCM.43.4.2026-2027.2005

118. Siu GKH, Zhang Y, Lau TCK, Lau RWT, Ho P-L, Yew W-W, et al. Mutations outside the rifampicin resistance-determining region associated with rifampicin resistance in *Mycobacterium tuberculosis*. J Antimicrob Chemother . 2011;66(4):730–3. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21393153

119. Yao C, Zhu T, Li Y, Zhang L, Zhang B, Huang J, et al. Detection of *rpoB*, *katG* and *inhA* gene mutations in *Mycobacterium tuberculosis* clinical isolates from Chongqing as determined by microarray. Clin Microbiol Infect . 2010 Nov;16(11):1639–43. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1198743X14605577

120. Jamieson FB, Guthrie JL, Neemuchwala A, Lastovetska O, Melano RG, Mehaffy C. Profiling of rpoB Mutations and MICs for Rifampin and Rifabutin in *Mycobacterium tuberculosis*. J Clin Microbiol . 2014 Jun 1;52(6):2157–62. Available from: http://jcm.asm.org/cgi/doi/10.1128/JCM.00691-14

121. van Ingen J, Aarnoutse R, de Vries G, Boeree MJ, van Soolingen D. Low-level rifampicin-resistant *Mycobacterium tuberculosis* strains raise a new therapeutic challenge [Short

communication]. Int J Tuberc Lung Dis . 2011 Jul 1;15(7):990–2. Available from: http://openurl.ingenta.com/content/xref?genre=article&issn=1027-3719&volume=15&issue=7&spage=990

122. Manson AL, Cohen KA, Abeel T, Desjardins CA, Armstrong DT, Barry CE, et al. Genomic analysis of globally diverse *Mycobacterium tuberculosis* strains provides insights into the emergence and spread of multidrug resistance. Nat Genet . 2017 Mar 16;49(3):395–402. Available from: http://www.nature.com/articles/ng.3767

123. Takayama K, Kilburn JO. Inhibition of synthesis of arabinogalactan by ethambutol in *Mycobacterium smegmatis*. Antimicrob Agents Chemother . 1989 Sep;33(9):1493–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/2817850

124. Beggs WH, Auran NE. Uptake and binding of 14C-ethambutol by tubercle bacilli and the relation of binding to growth inhibition. Antimicrob Agents Chemother . 1972 Nov;2(5):390–4. Available from: http://www.ncbi.nlm.nih.gov/pubmed/4207958

125. FORBES M, KUCK NA, PEETS EA. Mode of action of ethambutol. J Bacteriol . 1962 Nov;84:1099–103. Available from: http://www.ncbi.nlm.nih.gov/pubmed/13958686

126. 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 May;3(5):567–70. Available from: http://www.ncbi.nlm.nih.gov/pubmed/9142129

127. Jadaun GPS, Das R, Upadhyay P, Chauhan DS, Sharma VD, Katoch VM. Role of *embCAB* gene mutations in ethambutol resistance in *Mycobacterium tuberculosis* isolates from India. Int J Antimicrob Agents . 2009 May;33(5):483–6. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0924857908005268

128. Myoung HJ, Cho SN, Bai GH, Kim SJ, Kim JD, Bang HE, et al. Mutations in the *embB* Locus among Korean Clinical Isolates of *Mycobacterium tuberculosis* Resistant to Ethambutol. Yonsei Med J . 2002 Feb;43(1):59. Available from: http://www.ncbi.nlm.nih.gov/pubmed/11854934

129. Plinke C, Rusch-Gerdes S, Niemann S. Significance of Mutations in *embB* Codon 306 for Prediction of Ethambutol Resistance in Clinical *Mycobacterium tuberculosis* Isolates. Antimicrob Agents Chemother. 2006 May 1;50(5):1900–2. Available from: http://aac.asm.org/cgi/doi/10.1128/AAC.50.5.1900-1902.2006

130. Park YK, Ryoo SW, Lee SH, Jnawali HN, Kim C-K, Kim HJ, et al. Correlation of the phenotypic ethambutol susceptibility of *Mycobacterium tuberculosis* with *embB* gene mutations in Korea. J Med Microbiol . 2012 Apr 1;61(Pt_4):529–34. Available from: http://jmm.microbiologyresearch.org/content/journal/jmm/10.1099/jmm.0.037614-0

131. Zhang Y, Scorpio A, Nikaido H, Sun Z. Role of acid pH and deficient efflux of pyrazinoic acid in unique susceptibility of *Mycobacterium tuberculosis* to pyrazinamide. J Bacteriol . 1999 Apr;181(7):2044–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/10094680

132. den Hertog AL, Menting S, Pfeltz R, Warns M, Siddiqi SH, Anthony RM. Pyrazinamide Is Active against *Mycobacterium tuberculosis* Cultures at Neutral pH and Low Temperature. Antimicrob Agents Chemother . 2016 Aug;60(8):4956–60. Available from: http://aac.asm.org/lookup/doi/10.1128/AAC.00654-16

133. Scorpio A, Zhang Y. Mutations in *pncA*, a gene encoding pyrazinamidase/nicotinamidase, cause resistance to the antituberculous drug pyrazinamide in tubercle bacillus. Nat Med . 1996 Jun;2(6):662–7. Available from: http://www.ncbi.nlm.nih.gov/pubmed/8640557

134. Morlock GP, Crawford JT, Butler WR, Brim SE, Sikes D, Mazurek GH, et al. Phenotypic characterization of *pncA* mutants of *Mycobacterium tuberculosis*. Antimicrob Agents Chemother . 2000 Sep;44(9):2291–5. Available from: http://www.ncbi.nlm.nih.gov/pubmed/10952570

135. Felnagle EA, Podevels AM, Barkei JJ, Thomas MG. Mechanistically distinct nonribosomal peptide synthetases assemble the structurally related antibiotics viomycin and capreomycin. Chembiochem . 2011 Aug 16;12(12):1859–67. Available from: http://doi.wiley.com/10.1002/cbic.201100193

136. Maus CE, Plikaytis BB, Shinnick TM. Mutation of *tlyA* Confers Capreomycin Resistance in *Mycobacterium tuberculosis*. Antimicrob Agents Chemother . 2005 Feb 1;49(2):571–7. Available from: http://www.ncbi.nlm.nih.gov/pubmed/15673735

137. Heifets L, Simon J, Pham V. Capreomycin is active against non-replicating *M. tuberculosis*. Ann Clin Microbiol Antimicrob . 2005 Apr 1;4(1):6. Available from: http://www.ncbi.nlm.nih.gov/pubmed/15804353

138. Bauskenieks M, Pole I, Skenders G, Jansone I, Broka L, Nodieva A, et al. Genotypic and phenotypic characteristics of aminoglycoside-resistant *Mycobacterium tuberculosis* isolates in Latvia. Diagn Microbiol Infect Dis . 2015 Mar;81(3):177–82. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0732889314004957

139. Chen W, Biswas T, Porter VR, Tsodikov O V, Garneau-Tsodikova S. Unusual regioversatility of acetyltransferase Eis, a cause of drug resistance in XDR-TB. Proc Natl Acad Sci U S A . 2011 Jun 14;108(24):9804–8. Available from: http://www.pnas.org/cgi/doi/10.1073/pnas.1105379108

140. Engström A, Perskvist N, Werngren J, Hoffner SE, Juréen P. Comparison of clinical isolates and in vitro selected mutants reveals that *tlyA* is not a sensitive genetic marker for capreomycin resistance in *Mycobacterium tuberculosis*. J Antimicrob Chemother . 2011 Jun 1;66(6):1247–54. Available from: https://academic.oup.com/jac/article-lookup/doi/10.1093/jac/dkr109

141. Badet B, Hughes P, Kohiyama M, Forterre P. Inhibition of DNA replication in vitro by pefloxacin. FEBS Lett . 1982 Aug 23;145(2):355–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/6751875

142. Malik S, Willby M, Sikes D, Tsodikov O V, Posey JE. New insights into fluoroquinolone resistance in *Mycobacterium tuberculosis*: functional genetic analysis of *gyrA* and *gyrB* mutations. Via LE, editor. PLoS One . 2012 Jun 28;7(6):e39754. Available from: http://dx.plos.org/10.1371/journal.pone.0039754

143. Chen Y-Y, Chang J-R, Huang W-F, Kuo S-C, Su I-J, Sun J-R, et al. Genetic Diversity of the *Mycobacterium tuberculosis* Beijing Family Based on SNP and VNTR Typing Profiles in Asian Countries. Tyagi AK, editor. PLoS One . 2012 Jul 12 ;7(7):e39792. Available from: https://dx.plos.org/10.1371/journal.pone.0039792

144. Von Groll A, Martin A, Jureen P, Hoffner S, Vandamme P, Portaels F, et al. Fluoroquinolone Resistance in *Mycobacterium tuberculosis* and Mutations in *gyrA* and gyrB. Antimicrob Agents Chemother . 2009 Oct 1;53(10):4498–500. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19687244

145. Takiff HE, Salazar L, Guerrero C, Philipp W, Huang WM, Kreiswirth B, et al. Cloning and nucleotide sequence of *Mycobacterium tuberculosis gyrA* and *gyrB* genes and detection of quinolone resistance mutations. Antimicrob Agents Chemother . 1994 Apr;38(4):773–80. Available from: http://www.ncbi.nlm.nih.gov/pubmed/8031045

146. Feuerriegel S, Cox HS, Zarkua N, Karimovich HA, Braker K, Rusch-Gerdes S, et al. Sequence Analyses of Just Four Genes To Detect Extensively Drug-Resistant *Mycobacterium tuberculosis* Strains in Multidrug-Resistant Tuberculosis Patients Undergoing Treatment. Antimicrob Agents Chemother . 2009 Aug 1;53(8):3353–6. Available from: http://aac.asm.org/cgi/doi/10.1128/AAC.00050-09

147. WHO Geneva. TB: A Global Emergency. 1994.

148. WHO. Global tuberculosis report. 2017.

149. Tacconelli E, Carrara E, Savoldi A, Harbarth S, Mendelson M, Monnet DL, et al. Discovery, research, and development of new antibiotics: the WHO priority list of antibiotic-resistant bacteria and tuberculosis. Lancet Infect Dis . 2018 Mar;18(3):318–27. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1473309917307533

150. ECDC. Tuberculosis surveillance and monitoring in Europe 2018. 2018.

151. Directorate PGH. Tuberculose em Portugal: Desafios e estratégias. 2018.

152. Van Embden JDA, Cave MD, Crawford JT, Dale JW, Eisenach KD, Gicquel B, et al. Strain identification of *Mycobacterium tuberculosis* by DNA fingerprinting: Recommendations for a standardized methodology . Vol. 31, Journal of Clinical Microbiology. American Society for Microbiology (ASM); 1993 . p. 406–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/8381814

153. Kamerbeek J, Schouls L, Kolk A, van Agterveld M, van Soolingen D, Kuijper S, et al. Simultaneous detection and strain differentiation of *Mycobacterium tuberculosis* for diagnosis and epidemiology. J Clin Microbiol . 1997 Apr ;35(4):907–14. Available from: http://www.ncbi.nlm.nih.gov/pubmed/9157152

154. Supply P, Allix C, Lesjean S, Cardoso-Oelemann M, Rusch-Gerdes S, Willery E, et al. Proposal for Standardization of Optimized Mycobacterial Interspersed Repetitive Unit-Variable-Number Tandem Repeat Typing of *Mycobacterium tuberculosis*. J Clin Microbiol . 2006 Dec 1 ;44(12):4498–510. Available from: http://www.ncbi.nlm.nih.gov/pubmed/17005759

155. Supply P, Lesjean S, Savine E, Kremer K, van Soolingen D, Locht C. Automated high-throughput genotyping for study of global epidemiology of *Mycobacterium tuberculosis* based on mycobacterial interspersed repetitive units. J Clin Microbiol . 2001 Oct 1 ;39(10):3563–71. Available from: http://jcm.asm.org/cgi/doi/10.1128/JCM.39.10.3563-3571.2001

156. de Beer JL, Kremer K, Kodmon C, Supply P, van Soolingen D. First Worldwide Proficiency Study on Variable-Number Tandem-Repeat Typing of *Mycobacterium tuberculosis* Complex Strains. J Clin Microbiol . 2012 Mar 1 ;50(3):662–9. Available from: http://jcm.asm.org/cgi/doi/10.1128/JCM.00607-11

157. Gardy JL, Johnston JC, Sui SJH, Cook VJ, Shah L, Brodkin E, et al. Whole-Genome Sequencing and Social-Network Analysis of a Tuberculosis Outbreak. N Engl J Med . 2011 Feb 24 ;364(8):730–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21345102

158. Liu L, Li Y, Li S, Hu N, He Y, Pong R, et al. Comparison of next-generation sequencing systems. In: The Role of Bioinformatics in Agriculture . 2014 . p. 1–25. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22829749

159. M. Grueber. Economic Impact of the Human Genome Project. Hum Gene Ther . 2011 Jul ;22(7):777–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21756073

160. Collins FS, Morgan M, Patrinos A. The Human Genome Project: Lessons from large-scale biology . Vol. 300, Science. 2003 . p. 286–90. Available from: http://www.ncbi.nlm.nih.gov/pubmed/12690187

161. Roetzer A, Diel R, Kohl TA, Rückert C, Nübel U, Blom J, et al. Whole genome sequencing versus traditional genotyping for investigation of a *Mycobacterium tuberculosis* outbreak: a longitudinal molecular epidemiological study. PLoS Med . Public Library of Science; 2013 [cited 2018 Mar 21];10(2):e1001387. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23424287

162. Walker TM, Ip CL, Harrell RH, Evans JT, Kapatai G, Dedicoat MJ, et al. Whole-genome sequencing to delineate *Mycobacterium tuberculosis* outbreaks: A retrospective observational study. Lancet Infect Dis . 2013 Feb ;13(2):137–46. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1473309912702773

163. Brown T, Nejentsev S, Kontsevaya I, Nikolayevskyy V, Balabanova Y, Parkhill J, et al. Evolution and transmission of drug-resistant tuberculosis in a Russian population. Nat Genet . 2014 Mar 26 ;46(3):279–86. Available from: http://www.nature.com/articles/ng.2878

164. Gagneux S, Borrell S, Kato-Maeda M, Niemann S, Roetzer A, Comas I, et al. Whole-genome sequencing of rifampicin-resistant *Mycobacterium tuberculosis* strains identifies compensatory mutations in RNA polymerase genes. Nat Genet . 2011 Jan 18 ;44(1):106–10. Available from: http://www.nature.com/articles/ng.1038

165. Murray M, Kieser KJ, Streicher EM, Borowsky ML, Posey JE, Kurepina N, et al. Genomic analysis identifies targets of convergent positive selection in drug-resistant *Mycobacterium tuberculosis*. Nat Genet . 2013 Oct 1 ;45(10):1183–9. Available from: http://www.nature.com/articles/ng.2747

166. Votintseva AA, Bradley P, Pankhurst L, Del Ojo Elias C, Loose M, Nilgiriwala K, et al. Same-Day Diagnostic and Surveillance Data for Tuberculosis via Whole-Genome Sequencing of Direct Respiratory Samples. J Clin Microbiol . American Society for Microbiology; 2017 Mar 8 [cited 2018 Mar 15];55(5):1285–98. Available from: http://www.ncbi.nlm.nih.gov/pubmed/28275074

167. Brown AC, Bryant JM, Einer-Jensen K, Holdstock J, Houniet DT, Chan JZM, et al. Rapid whole-genome sequencing of *mycobacterium tuberculosis* isolates directly from clinical samples. J Clin Microbiol. 2015;53(7):2230–7.

168. ECDC. Expert opinion on whole genome sequencing for public health surveillance. 2016.

169. WHO. Global tuberculosis report 2015 . 2015. Available from: http://www.who.int/iris/handle/10665/191102

170. Pfyffer GE, Wittwer F. Incubation time of mycobacterial cultures: How long is long enough to issue a final negative report to the clinician? . Vol. 50, Journal of Clinical Microbiology. 2012 . p. 4188–9. Available from: http://jcm.asm.org/cgi/doi/10.1128/JCM.02283-12

171. WHO. Companion handbook to the WHO guidelines for the programmatic management of drug-resistant tuberculosis. 2014.

172. Boehme CC, Ruesch-Gerdes S, Jones M, Shenai S, Krapp F, Allen J, et al. Rapid Molecular Detection of Tuberculosis and Rifampin Resistance. N Engl J Med . 2010 Sep 9 ;363(11):1005–15. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20825313

173. Satta G, Lipman M, Smith GP, Arnold C, Kon OM, McHugh TD. *Mycobacterium tuberculosis* and whole-genome sequencing: how close are we to unleashing its full potential? Clin Microbiol Infect . 2018 Jun ;24(6):604–9. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1198743X17306237

174. Satta G, Atzeni A, McHugh TD. *Mycobacterium tuberculosis* and whole genome sequencing: a practical guide and online tools available for the clinical microbiologist. Clin Microbiol Infect . 2017 Feb ;23(2):69–72. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1198743X16303925

175. Faksri K, Tan JH, Chaiprasert A, Teo Y-Y, Ong RT-H. Bioinformatics tools and databases for whole genome sequence analysis of *Mycobacterium tuberculosis*. Infect Genet Evol . 2016 Nov ;45:359–68. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27637931

176. Schleusener V, Köser CU, Beckert P, Niemann S, Feuerriegel S. *Mycobacterium tuberculosis* resistance prediction and lineage classification from genome sequencing: Comparison of automated analysis tools. Sci Rep . Nature Publishing Group; 2017;7(September 2016):1–9. Available from: http://dx.doi.org/10.1038/srep46327

177. Nikolayevskyy V, Kranzer K, Niemann S, Drobniewski F. Whole genome sequencing of *Mycobacterium tuberculosis* for detection of recent transmission and tracing outbreaks: A systematic review. Tuberculosis . 2016 May ;98:77–85. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1472979215302705

178. Zhang Y, Yew WW. Mechanisms of drug resistance in *Mycobacterium tuberculosis*. Int J Tuberc Lung Dis . 2009 Nov ;13(11):1320–30. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19861002

179. Louw GE, Warren RM, Gey van Pittius NC, McEvoy CRE, Van Helden PD, Victor TC. A Balancing Act: Efflux/Influx in Mycobacterial Drug Resistance. Antimicrob Agents Chemother . 2009 Aug 1 ;53(8):3181–9. Available from: http://aac.asm.org/cgi/doi/10.1128/AAC.01577-08

180. 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 Feb 22 ;50(2):307–16. Available from: http://www.ncbi.nlm.nih.gov/pubmed/29358649

181. Manson AL, Cohen KA, Abeel T, Desjardins CA, Armstrong DT, Barry CE, et al. Genomic analysis of globally diverse *Mycobacterium tuberculosis* strains provides insights into the emergence and spread of multidrug resistance. Nat Genet . NIH Public Access; 2017 Mar [cited 2018 Mar 15];49(3):395–402. Available from: http://www.ncbi.nlm.nih.gov/pubmed/28092681

182. Papaventsis D, Casali N, Kontsevaya I, Drobniewski F, Cirillo DM, Nikolayevskyy V. Whole genome sequencing of *Mycobacterium tuberculosis* for detection of drug resistance: a systematic review. Clin Microbiol Infect . 2017 Feb ;23(2):61–8. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27665704

183. Walker TM, Kohl TA, Omar S V., Hedge J, Del Ojo Elias C, Bradley P, et al. Whole-genome sequencing for prediction of *Mycobacterium tuberculosis* drug susceptibility and resistance: A retrospective cohort study. Lancet Infect Dis. 2015;15(10):1193–202.

184. Coll F, McNerney R, Preston MD, Guerra-Assunção JA, Warry A, Hill-Cawthorne G, et al. Rapid determination of anti-tuberculosis drug resistance from whole-genome sequences. Genome Med; 2015;7(1):1–10.

185. 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. Carroll KC, editor. J Clin Microbiol . 2015 Jun ;53(6):1908–14. Available from: http://jcm.asm.org/lookup/doi/10.1128/JCM.00025-15

186. Sekizuka T, Yamashita A, Murase Y, Iwamoto T, Mitarai S, Kato S, et al. TGS-TB: Total Genotyping Solution for *Mycobacterium tuberculosis* Using Short-Read Whole-Genome Sequencing. Ahmed N, editor. PLoS One . 2015 Nov 13 ;10(11):e0142951. Available from: https://dx.plos.org/10.1371/journal.pone.0142951

187. Bradley P, Gordon NC, Walker TM, Dunn L, Heys S, Huang B, et al. Rapid antibiotic-resistance predictions from genome sequence data for *Staphylococcus aureus* and *Mycobacterium tuberculosis*. Nat Commun . Nature Publishing Group; 2015;6:1–14. Available from: http://dx.doi.org/10.1038/ncomms10063

188. Zhang Y, Yew W-W. Mechanisms of drug resistance in *Mycobacterium tuberculosis*: update 2015. Int J Tuberc Lung Dis . 2015 Nov 1 ;19(11):1276–89. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26467578

189. Whitfield MG, Soeters HM, Warren RM, York T, Sampson SL, Streicher EM, et al. A Global Perspective on Pyrazinamide Resistance: Systematic Review and Meta-Analysis. Mokrousov I, editor. PLoS One . 2015 Jul 28 ;10(7):e0133869. Available from: https://dx.plos.org/10.1371/journal.pone.0133869

190. Chatterjee A, Nilgiriwala K, Saranath D, Rodrigues C, Mistry N. Whole genome sequencing of clinical strains of *Mycobacterium tuberculosis* from Mumbai, India: A potential tool for determining drug-resistance and strain lineage. Tuberculosis . 2017 Dec ;107:63–72. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1472979217301786

191. Miotto P, Tessema B, Tagliani E, Chindelevitch L, Starks AM, Emerson C, et al. A standardised method for interpreting the association between mutations and phenotypic drug resistance in *Mycobacterium tuberculosis*. Eur Respir J . 2017;50(6):1701354. Available from: http://erj.ersjournals.com/lookup/doi/10.1183/13993003.01354-2017

192. Phelan J, O'Sullivan DM, Machado D, Ramos J, Whale AS, O'Grady J, et al. The variability and reproducibility of whole genome sequencing technology for detecting resistance to anti-tuberculous drugs. Genome Med . 2016 Dec 22 ;8(1):132. Available from: http://genomemedicine.biomedcentral.com/articles/10.1186/s13073-016-0385-x

193. Borges V, Nunes A, Sampaio DA, Vieira L, Machado J, Simões MJ, et al. *Legionella pneumophila* strain associated with the first evidence of person-to-person transmission of Legionnaires' disease: a unique mosaic genetic backbone. Sci Rep . 2016 Sep 19 ;6(1):26261. Available from: http://www.nature.com/articles/srep26261

194. Nikolayevskyy V, Hillemann D, Richter E, Ahmed N, van der Werf MJ, Kodmon C, et al. External Quality Assessment for Tuberculosis Diagnosis and Drug Resistance in the European Union: A Five Year Multicentre Implementation Study. Hozbor DF, editor.

PLoS One . 2016 Apr 7 ;11(4):e0152926. Available from: http://dx.plos.org/10.1371/journal.pone.0152926

195. Feuerriegel S, Koser CU, Niemann S. Phylogenetic polymorphisms in antibiotic resistance genes of the *Mycobacterium tuberculosis* complex. J Antimicrob Chemother . 2014 May 1 ;69(5):1205–10. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24458512

196. 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 Dec 1 ;5(1):4812. Available from: http://www.nature.com/articles/ncomms5812

197. Plinke C, Cox HS, Zarkua N, Karimovich HA, Braker K, Diel R, et al. *embCAB* sequence variation among ethambutol-resistant *Mycobacterium tuberculosis* isolates without embB306 mutation. J Antimicrob Chemother. 2010;65(7):1359–67.

198. Margaryan H, Rüsch-Gerdes S, Hayrapetyan A, Mirzoyan A. Ethambutol-resistance testing by mutation detection using MTBDRsl. Int J Mycobacteriology . 2016 Dec ;5:S50. Available from: http://www.ncbi.nlm.nih.gov/pubmed/28043606

199. Ellington MJ, Ekelund O, Aarestrup FM, Canton R, Doumith M, Giske C, et al. The role of whole genome sequencing in antimicrobial susceptibility testing of bacteria: report from the EUCAST Subcommittee. Clin Microbiol Infect . 2017 Jan ;23(1):2–22. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1198743X16305687

200. Bradley P, Gordon NC, Walker TM, Dunn L, Heys S, Huang B, et al. Rapid antibiotic-resistance predictions from genome sequence data for *Staphylococcus aureus* and *Mycobacterium tuberculosis*. Nat Commun . 2015 Dec 21 ;6(1):10063. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26686880

201. Mphahlele M, Syre H, Valvatne H, Stavrum R, Mannsaker T, Muthivhi T, et al. Pyrazinamide Resistance among South African Multidrug-Resistant *Mycobacterium tuberculosis* Isolates. J Clin Microbiol . 2008 Oct 1 ;46(10):3459–64. Available from: http://www.ncbi.nlm.nih.gov/pubmed/18753350

202. Martinez E, Holmes N, Jelfs P, Sintchenko V. Genome sequencing reveals novel deletions associated with secondary resistance to pyrazinamide in *MDR Mycobacterium tuberculosis*. J Antimicrob Chemother. 2015;70(9):2511–4.

203. Blanchard JS. Molecular Mechanisms of Drug Resistance in *Mycobacterium Tuberculosis*. Annu Rev Biochem . 1996 Jun ;65(1):215–39. Available from: http://www.ncbi.nlm.nih.gov/pubmed/8811179

204. 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 Jul ;2(1):29–35. Available from: http://www.ncbi.nlm.nih.gov/pubmed/4208567

205. 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 . 1994 Jan 14 ;263(5144):227–30. Available from: http://www.ncbi.nlm.nih.gov/pubmed/8284673

206. 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 Oct ;46(2):453–66. Available from: http://www.ncbi.nlm.nih.gov/pubmed/12406221

207. Lee H, Cho SN, Bang HE, Lee JH, Bai GH, Kim SJ, et al. Exclusive mutations related to isoniazid and ethionamide resistance among *Mycobacterium tuberculosis* isolates from Korea. Int J Tuberc Lung Dis . 2000 May ;4(5):441–7. Available from: http://www.ncbi.nlm.nih.gov/pubmed/10815738

208. de Welzen L, Eldholm V, Maharaj K, Manson AL, Earl AM, Pym AS. Whole-Transcriptome and -Genome Analysis of Extensively Drug-Resistant *Mycobacterium tuberculosis* Clinical Isolates Identifies Downregulation of ethA as a Mechanism of Ethionamide Resistance. Antimicrob Agents Chemother . 2017 Dec ;61(12). Available from: http://aac.asm.org/lookup/doi/10.1128/AAC.01461-17

209. 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 Dec ;47(12):3799–805. Available from: http://www.ncbi.nlm.nih.gov/pubmed/14638486

210. Baulard AR. Activation of the pro-drug ethionamide is regulated in mycobacteria. J Biol Chem . 2000 Jun 26 ;275(36):28326–31. Available from: http://www.ncbi.nlm.nih.gov/pubmed/10869356

211. DeBarber AE, Mdluli K, Bosman M, Bekker LG, Barry CE. Ethionamide activation and sensitivity in multidrug-resistant *Mycobacterium tuberculosis*. Proc Natl Acad Sci U S A . 2000 Aug 15 ;97(17):9677–82. Available from: http://www.ncbi.nlm.nih.gov/pubmed/10944230

212. Abe C, Kobayashi I, Mitarai S, Wada M, Kawabe Y, Takashima T, et al. Biological and Molecular Characteristics of *Mycobacterium tuberculosis* Clinical Isolates with Low-Level Resistance to Isoniazid in Japan. J Clin Microbiol . 2008 Jul 1 ;46(7):2263–8. Available from: http://jcm.asm.org/cgi/doi/10.1128/JCM.00561-08

213. Pankhurst LJ, del Ojo Elias C, Votintseva AA, Walker TM, Cole K, Davies J, et al. Rapid, comprehensive, and affordable mycobacterial diagnosis with whole-genome sequencing: A prospective study. Lancet Respir Med . Pankhurst et al. Open Access article distributed under the terms of CC BY; 2016;4(1):49–58. Available from: http://dx.doi.org/10.1016/S2213-2600(15)00466-X

214. Köser CU, Bryant JM, Becq J, Török ME, Ellington MJ, Marti-Renom MA, et al. Whole-Genome Sequencing for Rapid Susceptibility Testing of M. tuberculosis. N Engl J Med .

2013 Jul 18 ;369(3):290–2. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23863072

215. PHE. England world leaders in the use of whole genome sequencing to diagnose TB. 2017.

216. Stucki D, Gagneux S. Single nucleotide polymorphisms in *Mycobacterium tuberculosis* and the need for a curated database. Tuberculosis . 2013 Jan ;93(1):30–9. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1472979212002028

217. Salamon H, Yamaguchi KD, Cirillo DM, Miotto P, Schito M, Posey J, et al. Integration of Published Information Into a Resistance-Associated Mutation Database for *Mycobacterium tuberculosis*. J Infect Dis . 2015 Apr 1 ;211(suppl_2):S50–7. Available from: https://academic.oup.com/jid/article-lookup/doi/10.1093/infdis/jiu816

218. Rito T, Matos C, Carvalho C, Machado H, Rodrigues G, Oliveira O, et al. A complex scenario of tuberculosis transmission is revealed through genetic and epidemiological surveys in Porto. BMC Infect Dis . 2018 Dec 25 ;18(1):53. Available from: https://bmcinfectdis.biomedcentral.com/articles/10.1186/s12879-018-2968-1

219. Directorate PGH. estes de Sensibilidade aos Antituberculosos de 2ª Linha. 2007.

220. Portugal I, Barreiro L, Vultos T, Macedo R, Furtado C, Antunes AF, et al. Molecular epidemiology of *Mycobacterium tuberculosis* in Lisbon. Rev Port Pneumol. 2008;14(2).

221. Macedo R, Antunes AF, Villar M, Portugal I. Multidrug and extensively drug-resistant tuberculosis in Lisbon and Vale do Tejo, Portugal, from 2008 to 2010. Int J Mycobacteriology. 2012;1(3).

222. Directorate PGH. Portugal: Infeção VIH, SIDA e Tuberculose em números . 2017. Available from: https://www.dgs.pt/paginas-de-sistema/saude-de-a-a-z/tuberculose1/relatorios.aspx

223. Mazars E, Lesjean S, Banuls A-L, Gilbert M, Vincent V, Gicquel B, et al. High-resolution minisatellite-based typing as a portable approach to global analysis of *Mycobacterium tuberculosis* molecular epidemiology. Proc Natl Acad Sci . 2001 Feb 13 ;98(4):1901–6. Available from: http://www.pnas.org/cgi/doi/10.1073/pnas.98.4.1901

224. Oelemann MC, Diel R, Vatin V, Haas W, Rusch-Gerdes S, Locht C, et al. Assessment of an Optimized Mycobacterial Interspersed Repetitive- Unit-Variable-Number Tandem-Repeat Typing System Combined with Spoligotyping for Population-Based Molecular Epidemiology Studies of Tuberculosis. J Clin Microbiol . 2007 Mar 1 ;45(3):691–7. Available from: http://jcm.asm.org/cgi/doi/10.1128/JCM.01393-06

225. van Deutekom H, Supply P, de Haas PEW, Willery E, Hoijng SP, Locht C, et al. Molecular Typing of Mycobacterium tuberculosis by Mycobacterial Interspersed Repetitive Unit-Variable-Number Tandem Repeat Analysis, a More Accurate Method for Identifying Epidemiological Links between Patients with Tuberculosis. J Clin Microbiol . 2005 Sep 1

;43(9):4473–9. Available from: http://jcm.asm.org/cgi/doi/10.1128/JCM.43.9.4473-4479.2005

226. Rasoahanitralisoa R, Rakotosamimanana N, Stucki D, Sola C, Gagneux S, Rasolofo Razanamparany V. Evaluation of spoligotyping, SNPs and customised MIRU-VNTR combination for genotyping *Mycobacterium tuberculosis* clinical isolates in Madagascar. Neyrolles O, editor. PLoS One . 2017 Oct 20 ;12(10):e0186088. Available from: https://dx.plos.org/10.1371/journal.pone.0186088

227. Jonsson J, Hoffner S, Berggren I, Bruchfeld J, Ghebremichael S, Pennhag A, et al. Comparison between RFLP and MIRU-VNTR Genotyping of *Mycobacterium tuberculosis* Strains Isolated in Stockholm 2009 to 2011. Supply P, editor. PLoS One . 2014 Apr 14 ;9(4):e95159. Available from: http://dx.plos.org/10.1371/journal.pone.0095159

228. Kohl TA, Diel R, Harmsen D, Rothgänger J, Meywald Walter K, Merker M, et al. Whole-genome-based *Mycobacterium tuberculosis* surveillance: A standardized, portable, and expandable approach. J Clin Microbiol. 2014;52(7):2479–86.

229. Kohl TA, Harmsen D, Rothgänger J, Walker T, Diel R, Niemann S. Harmonized Genome Wide Typing of Tubercle Bacilli Using a Web-Based Gene-By-Gene Nomenclature System. EBioMedicine . 2018 Aug ;34:131–8. Available from: https://linkinghub.elsevier.com/retrieve/pii/S2352396418302731

230. Bravo LTC, Tuohy MJ, Ang C, Destura R V., Mendoza M, Procop GW, et al. Pyrosequencing for Rapid Detection of *Mycobacterium tuberculosis* Resistance to Rifampin, Isoniazid, and Fluoroquinolones. J Clin Microbiol . 2009 Dec 1 ;47(12):3985–90. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19846642

231. Macedo R, Nunes A, Portugal I, Duarte S, Vieira L, Gomes JP. Dissecting whole-genome sequencing-based online tools for predicting resistance in *Mycobacterium tuberculosis* : can we use them for clinical decision guidance? Tuberculosis . 2018 May [cited 2018 Apr 18];110:44–51. Available from: http://linkinghub.elsevier.com/retrieve/pii/S1472979218300672

232. Perdigão J, Silva H, Machado D, Macedo R, Maltez F, Silva C, et al. Unraveling genomic diversity and evolution in lisbon, portugal, a highly drug resistant setting. BMC Genomics. 2014;15(1).

233. Brown-Elliott BA, Simmer PJ, Trovato A, Hyle EP, Droz S, Buckwalter SP, et al. Mycobacterium decipiens sp. nov., a new species closely related to the *Mycobacterium tuberculosis* complex. Int J Syst Evol Microbiol . 2018 Nov 1 ;68(11):3557–62. Available from: http://www.microbiologyresearch.org/content/journal/ijsem/10.1099/ijsem.0.003031

234. Cabibbe AM, Walker TM, Niemann S, Cirillo DM. Whole genome sequencing of *Mycobacterium tuberculosis*. Eur Respir J . 2018 Nov ;52(5):1801163. Available from: http://www.ncbi.nlm.nih.gov/pubmed/30209198

235. Cabibbe AM, Trovato A, De Filippo MR, Ghodousi A, Rindi L, Garzelli C, et al. Countrywide implementation of whole genome sequencing: an opportunity to improve tuberculosis management, surveillance and contact tracing in low incidence countries. Eur Respir J . 2018 Jun ;51(6):1800387. Available from: http://erj.ersjournals.com/lookup/doi/10.1183/13993003.00387-2018

236. Directorate PGH. Circular Normativa Nº 1/DT de 11/01/2007 . 2007. Available from: https://www.dgs.pt/directrizes-da-dgs/normas-e-circulares-normativas/circular-normativa-n-1dt-de-11012007.aspx

237. Directorate PGH. Circular Normativa Nº 12/DSCS/PNT de 17/07/2008 . 2008. Available from: https://www.dgs.pt/...da...e.../circular-normativa-n-12dscspnt-de-17072008-pdf.aspx

238. Weniger T, Krawczyk J, Supply P, Niemann S, Harmsen D. MIRU-VNTRplus: a web tool for polyphasic genotyping of *Mycobacterium tuberculosis* complex bacteria. Nucleic Acids Res . 2010 Jul 1 ;38:W326–31. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20457747

239. Allix-Beguec C, Harmsen D, Weniger T, Supply P, Niemann S. Evaluation and Strategy for Use of MIRU-VNTRplus, a Multifunctional Database for Online Analysis of Genotyping Data and Phylogenetic Identification of *Mycobacterium tuberculosis* Complex Isolates. J Clin Microbiol . 2008 Aug 1 ;46(8):2692–9. Available from: http://jcm.asm.org/cgi/doi/10.1128/JCM.00540-08

240. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics . 2014 Aug 1 ;30(15):2114–20. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24695404

241. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. J Comput Biol . 2012 May ;19(5):455–77. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22506599

242. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al. Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. Wang J, editor. PLoS One . 2014 Nov 19 ;9(11):e112963. Available from: https://dx.plos.org/10.1371/journal.pone.0112963

243. Rajwani R, Shehzad S, Siu GKH. MIRU-profiler: a rapid tool for determination of 24-loci MIRU-VNTR profiles from assembled genomes of *Mycobacterium tuberculosis*. PeerJ . 2018 Jul 11 ;6:e5090. Available from: https://peerj.com/articles/5090

244. Silva M, Machado MP, Silva DN, Rossi M, Moran-Gilad J, Santos S, et al. chewBBACA: A complete suite for gene-by-gene schema creation and strain identification. Microb Genomics . 2018 Mar 1 ;4(3). Available from: http://www.microbiologyresearch.org/content/journal/mgen/10.1099/mgen.0.000166

245. Francisco AP, Bugalho M, Ramirez M, Carriço JA. Global optimal eBURST analysis of multilocus typing data using a graphic matroid approach. BMC Bioinformatics . 2009 May 18 ;10(1):152. Available from: http://www.biomedcentral.com/1471-2105/10/152

246. Ribeiro-Gonçalves B, Francisco AP, Vaz C, Ramirez M, Carriço JA. PHYLOViZ Online: web-based tool for visualization, phylogenetic inference, analysis and sharing of minimum spanning trees. Nucleic Acids Res . 2016 Jul 8 ;44(W1):W246–51. Available from: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkw359

247. Carrico JA, Silva-Costa C, Melo-Cristino J, Pinto FR, de Lencastre H, Almeida JS, et al. Illustration of a Common Framework for Relating Multiple Typing Methods by Application to Macrolide-Resistant *Streptococcus pyogenes*. J Clin Microbiol . 2006 Jul 1 ;44(7):2524–32. Available from: http://jcm.asm.org/cgi/doi/10.1128/JCM.02536-05

248. Vasconcellos SEG, Huard RC, Niemann S, Kremer K, Santos AR, Suffys PN, et al. Distinct genotypic profiles of the two major clades of *Mycobacterium africanum*. BMC Infect Dis . 2010 Dec 29 ;10(1):80. Available from: http://bmcinfectdis.biomedcentral.com/articles/10.1186/1471-2334-10-80

249. Mikheecheva NE, Zaychikova M V, Melerzanov A V, Danilenko VN. A Nonsynonymous SNP Catalog of *Mycobacterium tuberculosis* Virulence Genes and Its Use for Detecting New Potentially Virulent Sublineages. Genome Biol Evol . 2017 Apr 1 ;9(4):887–99. Available from: https://academic.oup.com/gbe/article-lookup/doi/10.1093/gbe/evx053

250. Wilson BA, Garud NR, Feder AF, Assaf ZJ, Pennings PS. The population genetics of drug resistance evolution in natural populations of viral, bacterial and eukaryotic pathogens. Mol Ecol . 2016 Jan ;25(1):42–66. Available from: http://doi.wiley.com/10.1111/mec.13474

251. Schürch AC, Kremer K, Hendriks ACA, Freyee B, McEvoy CRE, van Crevel R, et al. SNP/RD Typing of Mycobacterium tuberculosis Beijing Strains Reveals Local and Worldwide Disseminated Clonal Complexes. Gagneux S, editor. PLoS One . 2011 Dec 5 ;6(12):e28365. Available from: https://dx.plos.org/10.1371/journal.pone.0028365

252. Sengstake S, Bablishvili N, Schuitema A, Bzekalava N, Abadia E, de Beer J, et al. Optimizing multiplex SNP-based data analysis for genotyping of *Mycobacterium tuberculosis* isolates. BMC Genomics . 2014 Jul 7 ;15(1):572. Available from: http://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-15-572

253. Merker M, Kohl TA, Niemann S, Supply P. The Evolution of Strain Typing in the *Mycobacterium tuberculosis* Complex. In: Advances in experimental medicine and biology . 2017 . p. 43–78. Available from: http://www.ncbi.nlm.nih.gov/pubmed/29116629

254. Schürch AC, Arredondo-Alonso S, Willems RJL, Goering RV. Whole genome sequencing options for bacterial strain typing and epidemiologic analysis based on single nucleotide polymorphism versus gene-by-gene–based approaches. Clin Microbiol Infect . 2018 Apr

;24(4):350–4. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1198743X17307103

255. Nadon C, Van Walle I, Gerner-Smidt P, Campos J, Chinen I, Concepcion-Acevedo J, et al. PulseNet International: Vision for the implementation of whole genome sequencing (WGS) for global food-borne disease surveillance. Eurosurveillance . 2017 Jun 8 ;22(23):30544. Available from: http://www.eurosurveillance.org/ViewArticle.aspx?ArticleId=22807

256. Walker TM, Ip CL, Harrell RH, Evans JT, Kapatai G, Dedicoat MJ, et al. Whole-genome sequencing to delineate *Mycobacterium tuberculosis* outbreaks: a retrospective observational study. Lancet Infect Dis . 2013 Feb ;13(2):137–46. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1473309912702773

257. Zignol M, Cabibbe AM, Dean AS, Glaziou P, Alikhanova N, Ama C, et al. Genetic sequencing for surveillance of drug resistance in tuberculosis in highly endemic countries: a multi-country population-based surveillance study. The Lancet Infectious Diseases. 2018.

258. CRyPTIC Consortium and the 100 000 Genomes Project, Allix-Béguec C, Arandjelovic I, Bi L, Beckert P, Bonnet M, et al. Prediction of Susceptibility to First-Line Tuberculosis Drugs by DNA Sequencing. N Engl J Med . 2018 Oct 11 ;379(15):1403–15. Available from: http://www.nejm.org/doi/10.1056/NEJMoa1800474

259. Macedo R, Pinto M, Borges V, Nunes A, Oliveira O, Portugal I, et al. Evaluation of a gene-by-gene approach for prospective whole-genome sequencing-based surveillance of multidrug resistant *Mycobacterium tuberculosis*. Tuberculosis . Churchill Livingstone; 2019 Mar 1 ;115:81–8. Available from: https://www.sciencedirect.com/science/article/pii/S1472979218304748

260. Gnirke A, Melnikov A, Maguire J, Rogov P, LeProust EM, Brockman W, et al. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nat Biotechnol . 2009 Feb 1 ;27(2):182–9. Available from: http://www.nature.com/articles/nbt.1523

261. Votintseva AA, Pankhurst LJ, Anson LW, Morgan MR, Gascoyne-Binzi D, Walker TM, et al. Mycobacterial DNA extraction for whole-genome sequencing from early positive liquid (MGIT) cultures. J Clin Microbiol . American Society for Microbiology (ASM); 2015 Apr;53(4):1137–43. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25631807

262. Christiansen MT, Brown AC, Kundu S, Tutill HJ, Williams R, Brown JR, et al. Whole-genome enrichment and sequencing of *Chlamydia trachomatis* directly from clinical samples. BMC Infect Dis . 2014 Dec 12 ;14(1):591. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25388670

263. Pinto M, Borges V, Antelo M, Pinheiro M, Nunes A, Azevedo J, et al. Genome-scale analysis of the non-cultivable *Treponema pallidum* reveals extensive within-patient

genetic variation. Nat Microbiol . 2017 Jan 17 ;2(1):16190. Available from: http://www.nature.com/articles/nmicrobiol2016190

264. Achtman M. Evolution, Population Structure, and Phylogeography of Genetically Monomorphic Bacterial Pathogens. Annu Rev Microbiol . 2008 Oct ;62(1):53–70. Available from: http://www.annualreviews.org/doi/10.1146/annurev.micro.62.081307.162832

265. Gomes JP, Borrego MJ, Atik B, Santo I, Azevedo J, Brito de Sá A, et al. Correlating *Chlamydia trachomatis* infectious load with urogenital ecological success and disease pathogenesis. Microbes Infect . 2006 Jan ;8(1):16–26. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1286457905002029

266. Florindo C, Ferreira R, Borges V, Spellerberg B, Gomes JP, Borrego MJ. Selection of reference genes for real-time expression studies in *Streptococcus agalactiae*. J Microbiol Methods . 2012 Sep ;90(3):220–7. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0167701212001893

267. Coll F, McNerney R, Preston MD, Guerra-Assunção JA, Warry A, Hill-Cawthorne G, et al. Rapid determination of anti-tuberculosis drug resistance from whole-genome sequences. Genome Med . 2015 Dec 27 ;7(1):51. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26019726

268. Clark BE, Shooter C, Smith F, Brawand D, Thein SL. Next-generation sequencing as a tool for breakpoint analysis in rearrangements of the globin gene clusters. Int J Lab Hematol . 2017 May ;39:111–20. Available from: http://www.ncbi.nlm.nih.gov/pubmed/28447426

269. Li T, Unger ER, Batra D, Sheth M, Steinau M, Jasinski J, et al. Universal Human Papillomavirus Typing Assay: Whole-Genome Sequencing following Target Enrichment. Tang Y-W, editor. J Clin Microbiol . 2017 Mar ;55(3):811–23. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27974548

270. von der Lippe B, Sandven P, Brubakk O. Efficacy and safety of linezolid in multidrug resistant tuberculosis (MDR-TB)—a report of ten cases. J Infect . 2006 Feb ;52(2):92–6. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0163445305001015

271. Condos R, Hadgiangelis N, Leibert E, Jacquette G, Harkin T, Rom WN. Case Series Report of a Linezolid-Containing Regimen for Extensively Drug-Resistant Tuberculosis*. Chest . 2008 Jul ;134(1):187–92. Available from: https://linkinghub.elsevier.com/retrieve/pii/S001236920860166X

272. Yew WW, Chau CH, Wen KH. Linezolid in the treatment of “difficult” multidrug-resistant tuberculosis. Int J Tuberc Lung Dis . 2008 Mar ;12(3):345–6. Available from: http://www.ncbi.nlm.nih.gov/pubmed/18284844

273. Ismail N, Omar SV, Ismail NA, Peters RPH. Collated data of mutation frequencies and associated genetic variants of bedaquiline, clofazimine and linezolid resistance in

Mycobacterium tuberculosis. Data Br . 2018 Oct ;20:1975–83. Available from: https://linkinghub.elsevier.com/retrieve/pii/S2352340918311351

274. Agency EM. Delamanid Assessment Report. 2013.

275. Andries K, Villellas C, Coeck N, Thys K, Gevers T, Vranckx L, et al. Acquired Resistance of *Mycobacterium tuberculosis* to Bedaquiline. van Veen HW, editor. PLoS One . 2014 Jul 10 ;9(7):e102135. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25010492

276. Koul A, Dendouga N, Vergauwen K, Molenberghs B, Vranckx L, Willebrords R, et al. Diarylquinolines target subunit c of mycobacterial ATP synthase. Nat Chem Biol . 2007 Jun 13 ;3(6):323–4. Available from: http://www.ncbi.nlm.nih.gov/pubmed/17496888

277. Koul A, Vranckx L, Dendouga N, Balemans W, Van den Wyngaert I, Vergauwen K, et al. Diarylquinolines Are Bactericidal for Dormant Mycobacteria as a Result of Disturbed ATP Homeostasis. J Biol Chem . 2008 Sep 12 ;283(37):25273–80. Available from: http://www.jbc.org/lookup/doi/10.1074/jbc.M803899200

278. Matsumoto M, Hashizume H, Tomishige T, Kawasaki M, Tsubouchi H, Sasaki H, et al. OPC-67683, a Nitro-Dihydro-Imidazooxazole Derivative with Promising Action against Tuberculosis In Vitro and In Mice. Hopewell P, editor. PLoS Med . 2006 Nov 28 ;3(11):e466. Available from: https://dx.plos.org/10.1371/journal.pmed.0030466

279. Stinson K, Kurepina N, Venter A, Fujiwara M, Kawasaki M, Timm J, et al. MIC of Delamanid (OPC-67683) against *Mycobacterium tuberculosis* Clinical Isolates and a Proposed Critical Concentration. Antimicrob Agents Chemother . 2016 Jun ;60(6):3316–22. Available from: http://aac.asm.org/lookup/doi/10.1128/AAC.03014-15

280. Chen X, Hashizume H, Tomishige T, Nakamura I, Matsuba M, Fujiwara M, et al. Delamanid Kills Dormant Mycobacteria In Vitro and in a Guinea Pig Model of Tuberculosis. Antimicrob Agents Chemother . 2017 Jun ;61(6). Available from: http://aac.asm.org/lookup/doi/10.1128/AAC.02402-16

281. Fujiwara M, Kawasaki M, Hariguchi N, Liu Y, Matsumoto M. Mechanisms of resistance to delamanid, a drug for *Mycobacterium tuberculosis*. Tuberculosis . 2018 Jan ;108:186–94. Available from: http://www.ncbi.nlm.nih.gov/pubmed/29523322

282. Baym M, Lieberman TD, Kelsic ED, Chait R, Gross R, Yelin I, et al. Spatiotemporal microbial evolution on antibiotic landscapes. Science . 2016 Sep 9 [cited 2019 Mar 9];353(6304):1147–51. Available from: http://www.sciencemag.org/cgi/doi/10.1126/science.aag0822

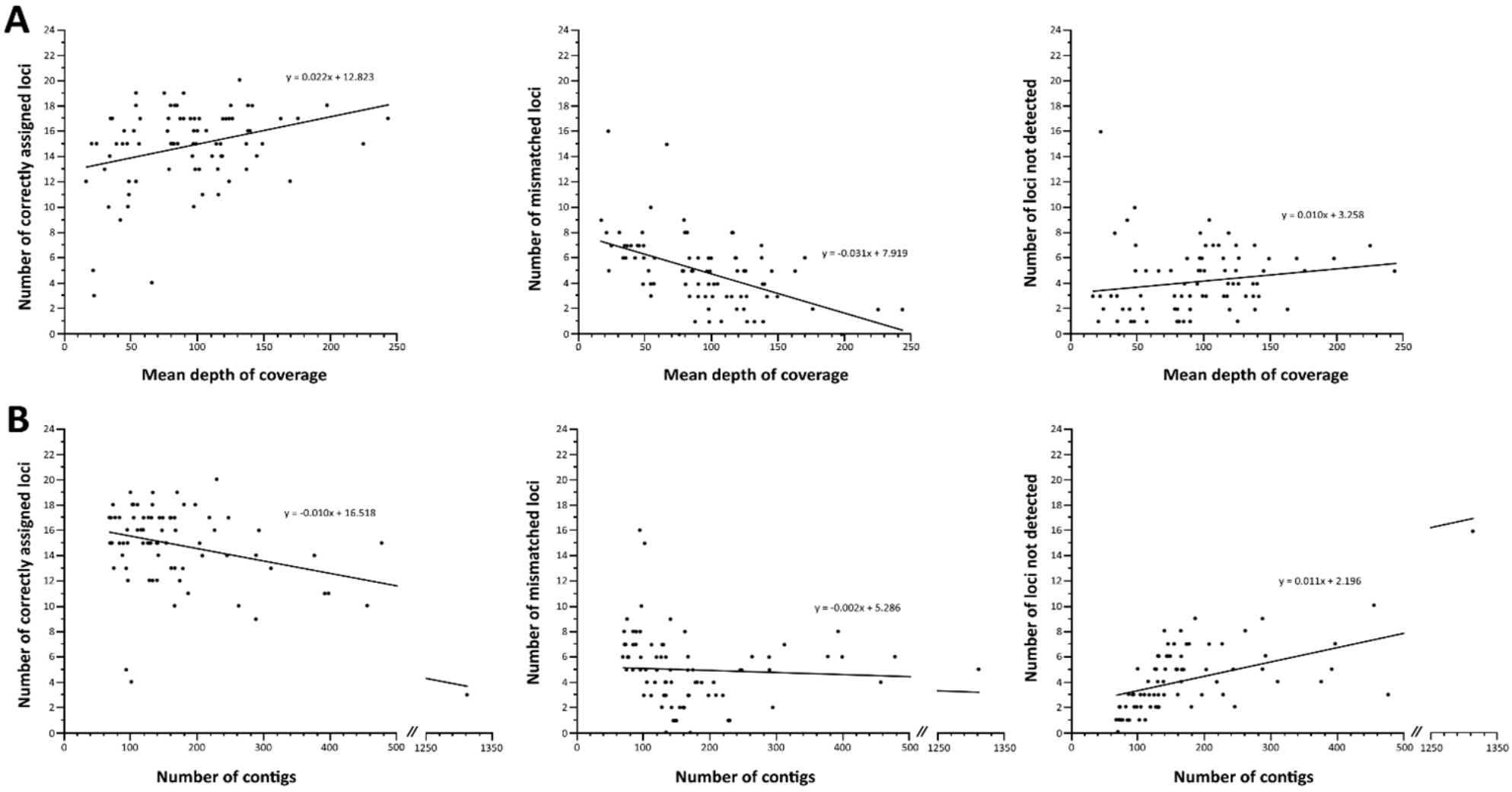
283. Perdigão J, Macedo R, Silva C, Pinto C, Furtado C, Brum L, et al. Tuberculosis drug-resistance in Lisbon, Portugal: A 6-year overview. Clin Microbiol Infect. 2011;17(9).

284. Perdigão J, Silva H, Machado D, Macedo R, Maltez F, Silva C, et al. Unraveling genomic diversity and evolution in lisbon, portugal, a highly drug resistant setting. BMC Genomics. 2014;15(1).

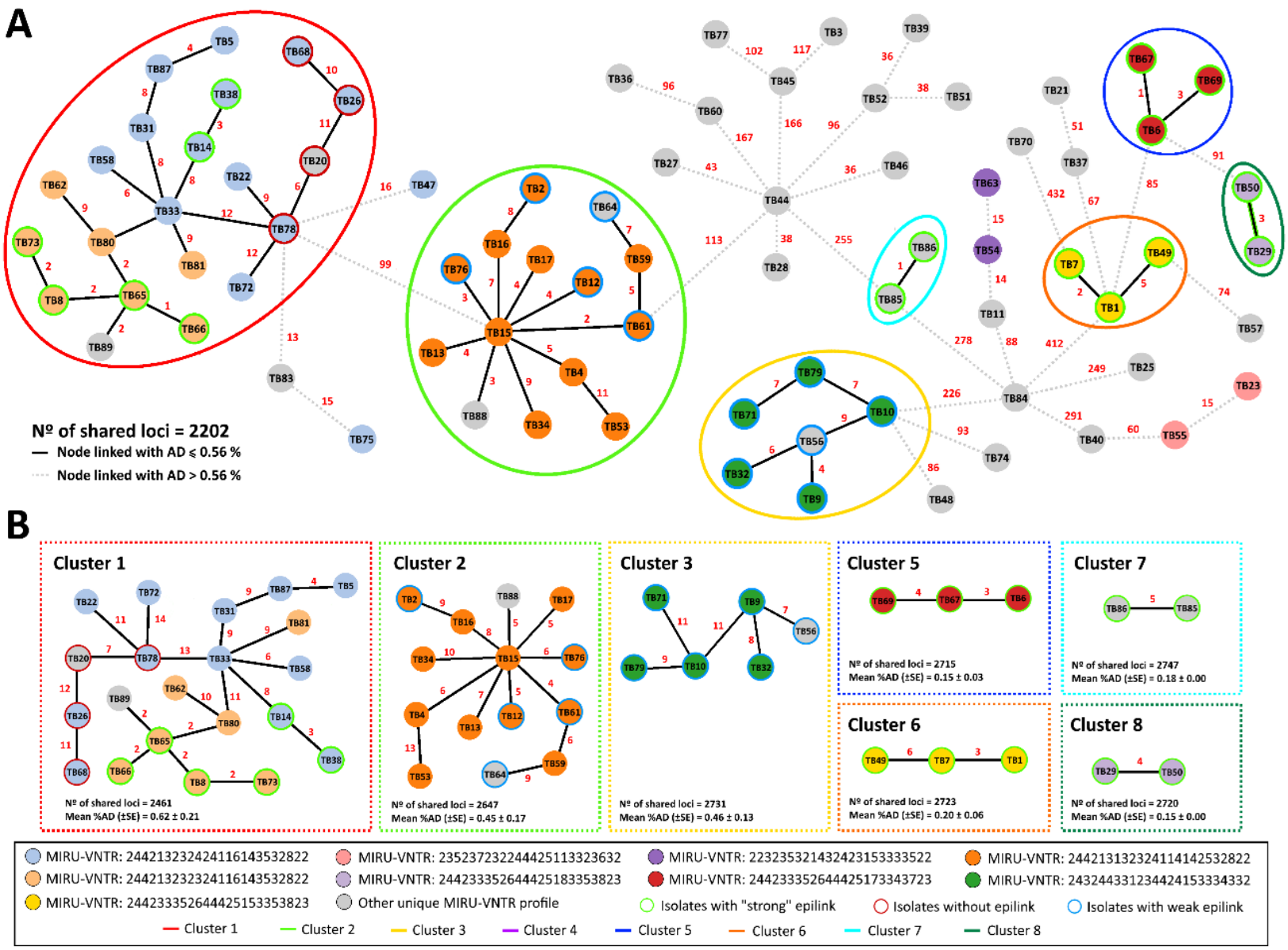
Supplementary material

Supplementary material

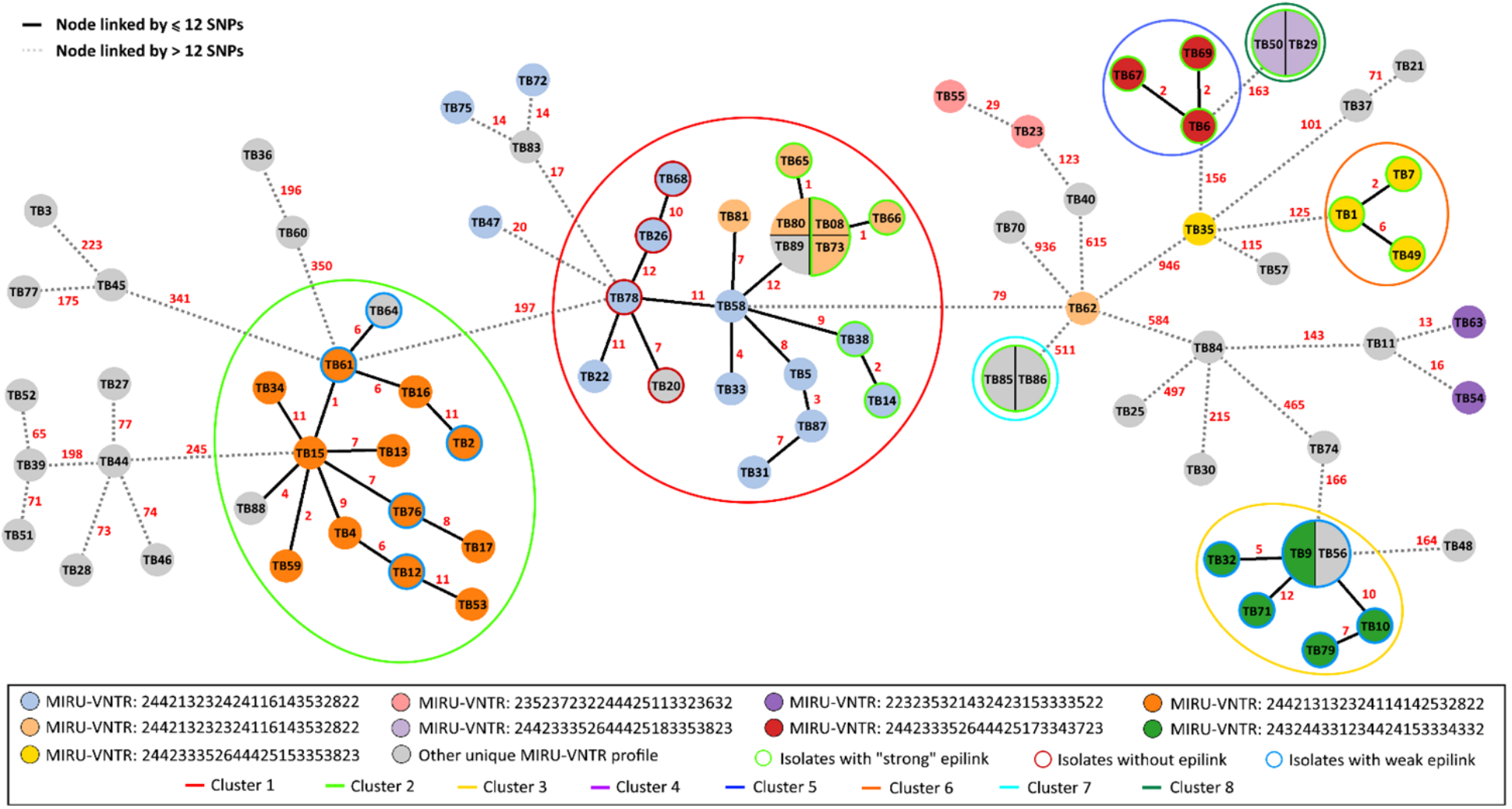
Supplementary Figure 4.1. Performance of the in silico determination of MIRU-VNTR profiles using MIRU-profiler software (243). The number of correctly assigned and mismatched loci refer to the comparison between traditional genotyping and in silico prediction. Not detected loci refer to loci flagged as such by MIRU-profiler. Results are plotted in correlation with (A) genome mean depth of coverage after read quality improvement using Trimmomatic (240), and (B) with the number of contigs generated after genome assembly.



Supplementary Figure 4.2. Phylogeny of 80 M/XDR-TB strains based on a dynamic gene-by-gene approach using a short schema (2891 loci). A – Initial Minimum spanning tree (MST) constructed based on allelic diversity found among the 2202 genes shared by 100% of the strains. Potential clusters defined for fine-tune analysis are highlighted by colored circles. B – Sub-MST reconstruction based on the maximum number of shared loci between strains within a potential cluster. Each circle (node) contains the strain's designation and represents a unique allelic profile. Nodes are colored according to traditional MIRU-VNTR profiles. The numbers in red on the connecting lines represent the allele differences (AD) between strains. MST were constructed using the goeBURST algorithm implemented in the PHYLOViZ Online platform, and are based on allelic profiles relying on distinct number of shared loci (indicated near the tree).



Supplementary Figure 4.3. M/XDR-TB core-SNV-based phylogenetic tree. Core-SNV–based MST of 82 MTBC strains reconstructed by using 5274 variant sites identified when mapping to the reference genome H37Rv (RefSeq NC_000962.3) and filtered out for SNVs falling within known genomic regions with high GC content or repetitive elements, as well as resistance-associated positions. Each circle (node) contains the strain's designation and represents a unique SNP profile. Nodes are colored according to traditional MIRU-VNTR profiles. The numbers in red on the connecting lines represent the SNP differences between strains. MST were constructed using the goeBURST algorithm implemented in the PHYLOViZ Online platform.



Supplementary Table 4.1. Sample dataset characterization. *Previously published under BioProject SRP131205. For better visualization of the information on this table, please refer to the internet link: https://www.sciencedirect.com/science/article/pii/S1472979218304748

Id	Lab. ID	Year	Gender	Age	Origin of isolation	TB-M/XDR	STR	INH	RMP	EMB	PZA1	AMI	CAP	ETI	MOX	OFL	LIN	KAN	CIC	PAS1	SRA accession #*	ENA ID	ENA accession #	MIRU-VNTR Profile	MIRU-Cluster	Assembled Genome size	read size	mean depth of coverage	# of contigs	# of loci called in the short schema	% of loci called in the short schema	# of loci called in the extended schema	% of loci called in the extended schema	Cluster with confirmed epilink
TB01	TB25429-14	2014	M	41	Lisbon and Tagus Valley	MR	R	R	R	R	R	S	S	R	S	S	S	S	S	S	SRR6511596	PT_TB0001	ERR2864235	244233352644425153353823	9	4352623	250	169.8	132	2726	94.29	3364	92.27	6
TB02	1595-14	2014	M	37	Lisbon and Tagus Valley	MR	R	R	R	S	R	S	S	R	S	S	S	S	S	S	SRR6511595	PT_TB0002	ERR2864288	24421313232411411412532822	1	4300244	250	119.3	119	2748	95.05	3394	93.09	2
TB03	1279-14	2014	F	31	Lisbon and Tagus Valley	MR	R	R	R	R	R	S	S	R	S	S	S	R	S	S	SRR6511594	PT_TB0003	ERR2864247	1322543222241251533226 22	---	4339519	250	75.8	169	2751	95.16	3397	93.17	---
TB04	1428-14	2014	M	48	Lisbon and Tagus Valley	MR	R	R	R	S	R	S	S	R	S	S	S	S	S	S	SRR6511593	PT_TB0004	ERR2864239	24421313232411411412532822	1	4335856	250	89.7	68	2751	95.16	3396	93.14	---
TB05	1378-14	2014	M	50	Lisbon and Tagus Valley	XDR	R	R	R	R	R	R	S	S	R	R	R	S	R	S	SRR6511592	PT_TB0005	ERR2864298	24421323242411611413532822	4	4290760	150	33.3	165	2744	94.92	3387	92.9	---
TB06	TB24818-14	2014	F	32	Lisbon and Tagus Valley	MR	R	R	R	R	S	S	S	S	S	S	S	S	S	S	SRR6511591	PT_TB0006	ERR2864238	24423333526444251733343723	10	4373724	250	30.7	93	2723	94.19	3360	92.16	5
TB07	1599-14	2014	M	43	Lisbon and Tagus Valley	MR	R	R	R	R	S	S	S	S	S	S	S	S	S	S	SRR6511590	PT_TB0007	ERR2864296	24423335526444251533533823	9	4356270	150	65.8	100	2725	94.26	3363	92.24	6
TB08	187-14	2014	M	58	Lisbon and Tagus Valley	MR	R	R	R	R	R	S	S	R	S	S	S	S	S	S	SRR6511589	PT_TB0008	ERR2864263	24421323232411611413532822	3	4336178	250	81.4	88	2747	95.02	3393	93.06	1b
TB09	88-14	2014	M	63	North region	MR	R	R	R	S	S	S	S	S	S	S	S	S	S	S	SRR6511588	PT_TB0009	ERR2864291	24324443312344241533344332	6	4364003	250	139.3	115	2740	94.78	3393	93.06	3
TB10	1536-14	2014	M	23	North region	MR	R	R	R	S	S	S	S	S	S	S	S	S	S	S	SRR6511587	PT_TB0010	ERR2864265	24324443312344241533344332	6	4379658	250	47.6	69	2739	94.74	3394	93.09	3
TB11	729-14	2014	M	47	North region	MR	R	R	R	S	S	S	S	S	S	S	S	S	S	S	SRR6511635	PT_TB0011	ERR2864233	2232353521432423153332522	---	4381013	250	57.1	77	2747	95.02	3392	93.03	---
TB12	1928-15	2015	M	61	Lisbon and Tagus Valley	MR	R	R	R	R	R	S	S	R	S	S	S	S	S	S	SRR6511634	PT_TB0012	ERR2864218	24421313232411411412532822	1	4333002	250	36.3	70	2751	95.16	3396	93.14	2
TB13	1229-15	2015	M	39	Lisbon and Tagus Valley	XDR	R	R	R	R	R	R	R	R	R	R	R	S	R	S	SRR6511637	PT_TB0013	ERR2864225	24421313232411411412532822	1	4328223	250	138.6	104	2748	95.05	3392	93.03	---
TB14	1876-15	2015	M	55	Lisbon and Tagus Valley	XDR	R	R	R	R	R	R	R	R	R	R	R	S	R	S	SRR6511636	PT_TB0014	ERR2864234	24421323242411611413532822	4	4332370	250	39.2	82	2751	95.16	3397	93.17	1a
TB15	1926-15	2015	M	52	Lisbon and Tagus Valley	XDR	R	R	R	R	R	R	R	R	R	R	R	S	R	S	SRR6511631	PT_TB0015	ERR2864240	24421313232411411412532822	1	4336664	250	141.1	110	2749	95.09	3393	93.06	---
TB16	2829-15	2015	M	44	Lisbon and Tagus Valley	MR	S	R	R	S	S	S	S	R	S	S	S	S	S	S	SRR6511630	PT_TB0016	ERR2864264	24421313232411411412532822	1	4290704	250	101.9	131	2749	95.09	3395	93.12	---
TB17	2184-15	2015	M	75	Lisbon and Tagus Valley	MR	R	R	R	R	R	R	R	R	S	S	S	R	S	S	SRR6511633	PT_TB0017	ERR2864250	24421313232411411412532822	1	4306787	250	89.6	104	2748	95.05	3395	93.12	---
TB20	2058-15	2015	M	57	Lisbon and Tagus Valley	MR	R	R	R	S	R	S	S	R	S	S	S	S	S	S	SRR6511642	PT_TB0020	ERR2864292	24421323242411611413532622	---	4294971	250	54.1	180	2746	94.98	3387	92.9	---
TB21	TB24737-15	2015	M	44	Lisbon and Tagus Valley	XDR	R	R	R	R	R	R	R	R	S	R	R	S	R	S	R	SRR6511610	PT_TB0021	ERR2864261	24422335526344251533533823	---	4364988	250	17.0	139	2708	93.67	3341	91.63	---
TB22	1585-15	2015	F	42	Lisbon and Tagus Valley	MR	R	R	R	R	S	S	S	R	S	S	S	S	S	S	SRR6511611	PT_TB0022	ERR2864256	24421323242411611413532822	4	4292967	250	175.4	158	2745	94.95	3387	92.9	---
TB23	2354-15	2015	F	27	Lisbon and Tagus Valley	MR	R	R	R	R	R	R	R	R	S	S	S	R	S	S	SRR6511612	PT_TB0023	ERR2864242	2352373232244425133233632	5	4360251	250	96.3	243	2720	94.09	3368	92.38	---
TB25	2579-15	2015	M	35	Lisbon and Tagus Valley	MR	R	R	R	R	R	R	S	S	R	S	S	S	S	S	SRR6511614	PT_TB0025	ERR2864270	24421323242411611413632822	---	4388702	250	22.4	93	2746	94.98	3403	93.34	---
TB26	2452-15	2015	M	57	North region	MR	R	R	R	S	S	S	S	R	S	S	S	S	S	S	SRR6511615	PT_TB0026	ERR2864257	24421323242411611413532822	4	4329104	250	45.5	111	2749	95.09	3394	93.09	---
TB27	982-15	2015	M	37	North region	MR	R	R	R	R	R	S	S	S	S	S	S	S	S	S	SRR6511616	PT_TB0027	ERR2864254	2142132322251261533323222	---	4334560	250	125.1	102	2740	94.78	3391	93.01	---
TB28	1880-15	2015	M	58	North region	MR	R	R	R	R	S	S	S	S	S	S	S	R	S	S	SRR6511617	PT_TB0028	ERR2864295	2142132322251261533323222	---	4260761	150	42.4	287	2714	93.88	3354	91.99	---
TB29	2353-15	2015	M	21	Island of Madeira	MR	S	R	R	R	S	S	S	S	S	S	S	S	S	S	SRR6511608	PT_TB0029	ERR2864231	24423335264442511533823	8	4353529	250	101.4	177	2722	94.15	3360	92.16	8
TB30	27-15	2015	M	44	Algarve	MR	S	R	R	R	S	S	S	S	R	S	S	S	S	S	SRR6511609	PT_TB0030	ERR2864278	2232353532534425153333632	---	4221968	250	137.3	476	2498	86.41	3081	84.5	---
TB31	308144-16	2016	M	41	Lisbon and Tagus Valley	XDR	R	R	R	R	R	R	S	R	R	R	S	R	S	S	SRR6511645	PT_TB0031	ERR2864267	24421323242411611413532822	4	4224372	250	137.0	310	2646	91.53	3252	89.19	---
TB32	309968-16	2016	F	41	North region	MR	R	R	R	S	S	S	S	S	S	S	S	S	S	S	SRR6511644	PT_TB0032	ERR2864221	24324443312344241533344332	6	4378927	250	44.4	72	2742	94.85	3397	93.17	3
TB33	310368-16	2016	M	42	North region	XDR	R	R	R	R	R	R	S	S	R	R	R	S	R	S	SRR6511601	PT_TB0033	ERR2864276	24421323242411611413532822	4	4210994	250	118.2	375	2604	90.07	3201	87.79	---
TB34	312205-16	2016	F	63	Lisbon and Tagus Valley	MR	S	R	R	S	R	R	R	R	S	S	S	R	S	S	SRR6511626	PT_TB0034	ERR2864289	24421313232411411412532822	1	4234612	250	162.4	246	2677	92.6	3290	90.24	---
TB35	314371-16	2016	M	15	Lisbon and Tagus Valley	MR	R	R	R	R	S	S	S	R	S	S	S	S	S	S	SRR6511641	PT_TB0035	ERR2864244	24423335526444251533533823	9	4246999	250	116.0	391	2571	88.93	3159	86.64	---
TB36	324134-16	2016	M	61	Centre region	MR	R	R	R	R	R	S	S	S	S	S	S	R	S	S	SRR6511640	PT_TB0036	ERR2864224	2324443224242261331317 22	---	4385993	250	83.0	94	2752	95.19	3412	93.58	---
TB37	326551-16	2016	M	40	Lisbon and Tagus Valley	MR	R	R	R	R	R	S	S	S	S	S	S	S	S	S	SRR6511639	PT_TB0037	ERR2864290	24423335526334251533533823	---	4367138	250	34.5	87	2726	94.29	3365	92.29	---
TB38	327889-16	2016	F	22	Lisbon and Tagus Valley	XDR	R	R	R	R	R	S	S	R	R	R	S	R	S	S	SRR6511638	PT_TB0038	ERR2864216	24421323242411611413532822	4	4335738	250	80.2	84	2752	95.19	3397	93.17	1a
TB39	320857-16	2016	M	43	Lisbon and Tagus Valley	MR	R	R	R	R	R	S	S	S	S	S	S	S	S	S	SRR6511619	PT_TB0039	ERR2864217	2542s122322241251533326222	---	4339057	250	35.3	75	2746	94.98	3397	93.17	---
TB40	316569-16	2016	M	54	North region	MR	R	R	R	S	S	R	R	S	S	S	S	R	S	S	SRR6511618	PT_TB0040	ERR2864248	2352371322244425133233532	---	4308129	250	145.1	287	2672	92.42	3297	90.43	---
TB44	TB31393-16	2016	F	40	Lisbon and Tagus Valley	MR	R	R	R	S	S	S	S	S	S	S	S	S	S	S	SRR6511620	PT_TB0044	ERR2864219	2142132322251261533324222	---	4341106	250	84.9	73	2745	94.95	3395	93.12	---
TB45	332846-16	2016	M	70	Centre region	MR	R	R	R	S	R	S	S	S	S	S	S	S	S	S	SRR6511606	PT_TB0045	ERR2864279	1342623322241261533227 22	---	4371371	250	78.4	82	2743	94.88	3399	93.23	---
TB46	347401-16	2016	F	20	Lisbon and Tagus Valley	MR	R	R	R	R	S	S	S	R	S	S	S	S	S	S	SRR6511607	PT_TB0046	ERR2864246	2142132322251261633324222	---	4327468	250	24.6	128	2690	93.05	3333	91.42	---
TB47	TB31250-16	2016	M	62	North region	MR	R	R	R	S	S	S	S	R	S	S	S	S	S	S	SRR6511604	PT_TB0047	ERR2864249	24421323242411611413532822	4	4329390	250	21.2	88	2747	95.02	3393	93.06	---
TB48	348387-16	2016	M	UNK	Centre region	MR	S	R	R	S	S	S	S	S	S	S	S	S	S	S	SRR6511605	PT_TB0048	ERR2864259	24324433233442515333473 2	---	4384558	250	78.8	74	2744	94.92	3402	93.31	---
TB49	352139-17	2017	F	22	Centre region	MR	R	R	R	R	R	S	S	S	S	S	S	S	S	S	SRR6511598	PT_TB0049	ERR2864252	24423335526444251533533823	9	4363470	250	54.0	95	2724	94.22	3363	92.24	6
TB50	375001-17	2017	F	22	Centre region	MR	S	R	R	R	S	S	S	R	S	S	S	S	S	S	SRR6511599	PT_TB0050	ERR2864272	24423335526444251833533823	8	4376598	250	49.0	127	2723	94.19	3362	92.21	8
TB51	374686-17	2017	F	34	Lisbon and Tagus Valley	MR	S	R	R	R	R	S	S	R	S	S	S	S	S	S	SRR6511623	PT_TB0051	ERR2864236	2542142322241251633333622	---	4342099	250	52.5	118	2747	95.02	3400	93.25	---
TB52	381277-17	2017	M	30	North region	MR	R	R	R	R	S	S	S	S	S	S	S	S	S	S	SRR6511622	PT_TB0052	ERR2864260	2542142322241251633332622	---	4346252	250	77.9	94	2749	95.09	3403	93.31	---
TB53	TB33470LB-17	2017	M	52	Lisbon and Tagus Valley	MR	S	R	R	R	R	S	S	R	S	S	S	S	S	S	SRR6511625	PT_TB0053	ERR2864294	24421313232411411412532822	1	4336120	250	89.8	99	2751	95.16	3396	93.14	---
TB54	3898865-17	2017	F	59	North region	MR	R	R	R	R	R	S	S	S	S	S	S	S	S	S	SRR6511624	PT_TB0054	ERR2864223	2232353221432423153333522	7	4380694	250	54.3	132	2747	95.02	3389	92.95	---
TB55	157108-13	2013	M	24	Lisbon and Tagus Valley	MR	R	R	R	S	S	S	S	R	S	S	S	R	S	S	---	PT_TB0055	ERR2864220	2352373232244425133233632	5	4405717	250	85.7	139	2743	94.88	3401	93.28	---
TB56	157637-13	2013	F	16	North region	XDR	R	R	R	R	R	R	R	R	R	S	S	S	S	S	---	PT_TB0056	ERR2864243	24324423123442415333432	---	4342056	250	100.6	166	2734	94.57	3386	92.87	3
TB57	159528-13	2013	M	42	North region	MR	R	R	R	R	S	R	S	S	S	S	S	R	S	S	---	PT_TB0057	ERR2864286	23423335526444251533533623	---	4348454	250	98.7	165	2728	94.36	3363	92.24	---
TB58	163-14	2014	M	53	Lisbon and Tagus Valley	XDR	R	R	R	R	R	R	S	R	R	R	S	R	S	S	---	PT_TB0058	ERR2864293	24421323242411611413532822	4	4308144	250	95.3	165	2746	94.98	3388	92.92	---
TB59	163761-13	2013	F	32	Lisbon and Tagus Valley	XDR	R	R	R	R	R	R	R	R	R	R	S	R	S	S	---	PT_TB0059	ERR2864283	24421313232411411412532822	1	4312712	250	126.3	143	2748	95.05	3394	93.09	---
TB60	171591-13	2013	M	47	Lisbon and Tagus Valley	XDR	R	R	R	R	R	R	R	S	R	S	S	S	S	S	---	PT_TB0060	ERR2864227	24423433222242261534141512	---	4348211	250	132.1	228	2736	94.64	3392	93.03	---
TB61	178958-13	2013	F	24	Lisbon and Tagus Valley	XDR	R	R	R	R	R	R	R	R	R	S	S	S	S	S	---	PT_TB0061	ERR2864271	24421313232411411412532822	1	4279065	250	126.2	218	2742	94.85	3377	92.62	2
TB62	180161-13	2013	M	50	North region	MR	R	R	R	R	R	R	S	S	R	S	S	S	S	S	---	PT_TB0062	ERR2864229	24421323232411611413532822	3	4303856	250	83.0	196	2737	94.67	3378	92.65	---
TB63	180198-13	2013	M	34	North region	MR	R	R	R	R	R	S	S	S	S	S	S	S	S	S	---	PT_TB0063	ERR2864275	2232353521432423153333522	7	4327660	250	97.2	292	2719	94.05	3341	91.63	---
TB64	199228-14	2014	M	42	Lisbon and Tagus Valley	XDR	R	R	R	R	R	R	R	R	R	R	S	S	S	S	---	PT_TB0064	ERR2864253	24421313232411411412532722	2	4267296	250	97.3	261	2732	94.5	3365	92.29	2
TB65	1994-15	2015	M	41	Lisbon and Tagus Valley	MR	R	R	R	R	S	R	S	S	R	S	S	S	S	S	---	PT_TB0065	ERR2864268	24421323232411611413532822	3	4329740	250	97.3	141	2745	94.95	3391	93.01	1b
TB66	207797-14	2014	M	53	Lisbon and Tagus Valley	MR	R	R	R	R	S	R	S	S	R	S	S	S	S	S	---	PT_TB0066	ERR2864226	24421323232411611413532822	3	4228239	250	48.2	454	2609	90.25	3205	87.9	1b
TB67	211891-14	2014	M	36	Lisbon and Tagus Valley	MR	R	R	R	R	S	S	S	R	S	S	S	S	S	S	---	PT_TB0067	ERR2864255	24423335526444251733433723	10	4362169	250	124.3	173	2723	94.19	3358	92.1	5
TB68	2471-15	2015	M	34	Lisbon and Tagus Valley	MR	R	R	R	R	R	S	S	S	R	S	S	S	S	S	---	PT_TB0068	ERR2864245	24421323242411611413532822	4	4337289	250	99.3	118	2752	95.19	3398	93.2	---
TB69	2624-15	2015	M	28	Lisbon and Tagus Valley	MR	R	R	R	R	S	S	S	S	R	S	S	S	S	S	---	PT_TB0069	ERR2864228	24423335526444251733433723	10	4351551	250	111.3	207	2716	93.95	3349	91.85	5
TB70	2911-14	2014	M	28	Lisbon and Tagus Valley	MR	R	R	R	R	R	S	S	S	S	S	S	S	S	S	---	PT_TB0070	ERR2864232	24223644277422516335 3843	---	4356216	250	103.7	186	2730	94.43	3374	92.54	---
TB71	292-14	2014	F	41	North region	MR	R	R	R	R	S	S	S	S	S	S	S	S	S	S	---	PT_TB0071	ERR2864258	24324443312344241533344332	6	4372065	250	87.1	148	2740	94.78	3394	93.09	3
TB72	340-14	2014	M	48	Centre region	XDR	R	R	R	R	S	R	S	R	S	R	S	R	S	S	---	PT_TB0072	ERR2864277	24421323242411611413532822	4	4306675	250	24.3	127	2743	94.88	3383	92.79	---
TB73	340248-16	2016	M	20	Lisbon and Tagus Valley	MR	R	R	R	S	R	S	S	S	R	S	S	S	S	S	---	PT_TB0073	ERR2864273	24421323232411611413532622	3	4327253	250	138.3	146	2746	94.98	3390	92.98	1b
TB74	356-14	2014	M	42	North region	MR	S	R	R	R	S	S	S	S	S	S	S	S	S	S	---	PT_TB0074	ERR2864285	2231543424344251533533812	---	4383915	250	121.6	130	2745	94.95	3404	93.36	---
TB75	387683-17	2017	M	56	Lisbon and Tagus Valley	XDR	R	R	R	R	R	R	S	R	R	R	S	R	S	S	---	PT_TB0075	ERR2864241	24421323242411611413532822	4	4324855	250	149.3	130	2748	95.05	3392	93.03	---
TB76	396397-17	2017	M	51	Lisbon and Tagus Valley	XDR	R	R	R	R	R	R	R	R	R	R	S	S	R	S	---	PT_TB0076	ERR2864269	24421313232411411412532822	1	4330853	250	243.1	126	2751	95.16	3398	93.2	2
TB77	399045-17	2017	F	41	North region	MR	R	R	R	S	S	S	S	R	R	R	S	S	S	S	---	PT_TB0077	ERR2864237	13322433222412775332622	---	4376744	250	197.9	132	2748	95.05	3404	93.36	---
TB78	399986-17	2017	M	58	Lisbon and Tagus Valley	MR	R	R	R	R	S	S	S	R	R	R	S	S	S	S	---	PT_TB0078	ERR2864262	24421323242411611413532822	4	4327232	250	224.8	154	2749	95.09	3394	93.09	---
TB79	423-14	2014	F	75	North region	MR	R	R	R	R	S	S	S	S	S	S	S	S	S	S	---	PT_TB0079	ERR2864222	24324443312344241533344332	6	4366225	250	114.6	130	2740	94.78	3395	93.12	3
TB80	884-15	2015	M	61	Lisbon and Tagus Valley	MR	R	R	R	R	S	R	S	S	R	S	S	S	S	S	---	PT_TB0080	ERR2864230	24421323232411611413532822	3	4287079	250	107.0	226	2732	94.5	3368	92.38	---
TB81	92-14	2014	M	59	Centre region	MR	R	R	R	R	S	S	S	S	R	S	S	S	S	S	---	PT_TB0081	ERR2864251	24421323232411611413532822	3	4323098	250	124.2	159	2749	95.09	3392	93.03	---
TB82	TB22229-13	2013	F	38	Lisbon and Tagus Valley	XDR	R	R	R	R	R	R	R	R	R	R	S	S	S	S	---	PT_TB0082	ERR2864297	24421313232411411412532722	2	4091146	150	22.9	1313	2077	71.84	2561	70.24	---
TB83	TB22634-13	2013	M	50	Lisbon and Tagus Valley	XDR	R	R	R	R	R	R	R	R	R	R	S	S	R	S	---	PT_TB0083	ERR2864280	24421323242411541413532822	---	4248079	250	48.8	396	2646	91.53	3254	89.25	---
TB84	TB23204-13	2013	M	35	Lisbon and Tagus Valley	MR	R	R	R	R	S	S	S	S	S	S	S	S	S	S	---	PT_TB0084	ERR2864266	2232353324342423163331732	---	4361159	250	80.4	132	2741	94.81	3380	92.7	---
TB85	TB23392-13	2013	M	41	Lisbon and Tagus Valley	MR	R	R	R	R	S	S	S	S	S	S	S	S	S	S	---	PT_TB0085	ERR2864282	234334-2243423614343522	---	4352831	250	115.4	160	2748	95.05	3402	93.31	7
TB86	TB24007-13	2013	M	37	Lisbon and Tagus Valley	MR	R	R	R	R	S	S	S	S	S	S	S	S	S	S	---	PT_TB0086	ERR2864274	2343333224342361433333522	---	4359274	250	119.0	140	2749	95.09	3406	93.42	7
TB87	TB24207-13	2013	F	29	Lisbon and Tagus Valley	XDR	R	R	R	R	R	R	S	R	R	R	S	R	S	S	---	PT_TB0087	ERR2864287	24421323242411611413532822	4	4336319	250	117.4	125	2752	95.19	3398	93.2	---
TB88	TB25380-13	2013	M	42	Lisbon and Tagus Valley	XDR	R	R	R	R	R	R	R	R	R	R	S	R	S	S	---	PT_TB0088	ERR2864284	24423316232411411412532822	---	4282031	250	56.4	203	2741	94.81	3377	92.62	---
TB89	TB34192-17	2017	M	62	Lisbon and Tagus Valley	MR	R	R	R	S	S	S	S	R	S	S	S	S	S	S	---	PT_TB0089	ERR2864281	24421323222411611413532822	---	4325767	250	98.0	146	2745	94.95	3391	93.01	---

Supplementary Table 4.2. List of loci masked from core-SNV-based analysis. As currently done by several studies, regions with high GC content or repetitive elements are excluded for SNV-based phylogenetic analysis. The list used in the present study included a recently compiled list available in a pipeline under construction (MTBSeq), as well as additional loci described by Merker 2013 (*). Locus tags, gene names, products and positions refer to the H37Rv reference genome (NC_000962.3).

Locus tag	Gene name	Product	Position
Rv0031	---	Possible remnant of a transposase	33582 - 33794
Rv0096	PPE1	PPE family protein PPE1	105324 - 106715
Rv0109	PE_PGRS1	PE-PGRS family protein PE PGRS1	131382 - 132872
Rv0124	PE_PGRS2	PE-PGRS family protein PE PGRS2	149533 - 150996
Rv0151c	PE1	PE family protein PE1	179309 - 177543
Rv0152c	PE2	PE family protein PE2	180896 - 179319
Rv0159c	PE3	PE family protein PE3	188839 - 187433
Rv0160c	PE4	PE family protein PE4	190439 - 188931
Rv0256c	PPE2	PPE family protein PPE2	309547 - 307877
Rv0278c	PE_PGRS3	PE-PGRS family protein PE PGRS3	336310 - 333437
Rv0279c	PE_PGRS4	PE-PGRS family protein PE PGRS4	339073 - 336560
Rv0280	PPE3	PPE family protein PPE3	339364 - 340974
Rv0285	PE5	PE family protein PE5	349624 - 349932
Rv0286	PPE4	PPE family protein PPE4	349935 - 351476
Rv0297	PE_PGRS5	PE-PGRS family protein PE PGRS5	361334 - 363109
Rv0304c	PPE5	PPE family protein PPE5	372764 - 366150
Rv0305c	PPE6	PPE family protein PPE6	375711 - 372820
Rv0335c	PE6	PE family protein PE6	400050 - 399535
Rv0354c	PPE7	PPE family protein PPE7	424694 - 424269
Rv0355c	PPE8	PPE family protein PPE8	434679 - 424777
Rv0387c	---	Conserved hypothetical protein	467406 - 466672
Rv0388c	PPE9	PPE family protein PPE9	468001 - 467459
Rv0442c	PPE10	PPE family protein PPE10	532214 - 530751
Rv0453	PPE11	PPE family protein PPE11	543174 - 544730
Rv0532	PE_PGRS6	PE-PGRS family protein PE PGRS6	622793 - 624577
Rv0578c	PE_PGRS7	PE-PGRS family protein PE PGRS7	675916 - 671996
Rv0741	---	Probable transposase (fragment)	832534 - 832848
Rv0742	PE_PGRS8	PE-PGRS family protein PE PGRS8	832981 - 833508
Rv0746	PE_PGRS9	PE-PGRS family protein PE PGRS9	835701 - 838052
Rv0747	PE_PGRS10	PE-PGRS family protein PE PGRS10	838451 - 840856
Rv0754	PE_PGRS11	PE-PGRS family protein PE PGRS11	846159 - 847913
Rv0755c	PPE12	PPE family protein PPE12	850040 - 848103
Rv0755A	---	Putative transposase (fragment)	850527 - 850342
Rv0795	---	Putative transposase for insertion sequence element IS6110 (fragment)	889072 - 889398
Rv0796	---	Putative transposase for insertion sequence element IS6110	889347 - 890333
Rv0797	---	Putative transposase for insertion sequence element IS1547	890388 - 891482
Rv0832	PE_PGRS12	PE-PGRS family protein PE PGRS12	924951 - 925364
Rv0833	PE_PGRS13	PE-PGRS family protein PE PGRS13	925361 - 927610
Rv0834c	PE_PGRS14	PE-PGRS family protein PE PGRS14	930485 - 927837
Rv0850	---	Putative transposase (fragment)	947312 - 947644
Rv0872c	PE_PGRS15	PE-PGRS family protein PE PGRS15	970244 - 968424
Rv0878c	PPE13	PPE family protein PPE13	978203 - 976872
Rv0915c	PPE14	PPE family protein PPE14	1021329 - 1020058
Rv0916c	PE7	PE family protein PE7	1021643 - 1021344
Rv0920c	---	Probable transposase	1026816 - 1025497
Rv0922	---	Possible transposase	1027685 - 1029337
Rv0977	PE_PGRS16	PE-PGRS family protein PE PGRS16	1090373 - 1093144
Rv0978c	PE_PGRS17	PE-PGRS family protein PE PGRS17	1094356 - 1093361
Rv0980c	PE_PGRS18	PE-PGRS family protein PE PGRS18	1096451 - 1095078
Rv1034c	---	Probable transposase (fragment)	1159307 - 1158918
Rv1035c	---	Probable transposase (fragment)	1160061 - 1159375
Rv1036c	---	Probable IS1560 transposase (fragment)	1160433 - 1160095
Rv1039c	PPE15	PPE family protein PPE15	1162472 - 1161297

Supplementary Table 4.2. (cont.)

Rv1040c	PE8	PE family protein PE8	1163376 - 1162549
Rv1041c	---	Probable is like-2 transposase	1165435 - 1164572
Rv1042c	---	Probable is like-2 transposase	1165499 - 1165092
Rv1047	---	Probable transposase	1169423 - 1170670
Rv1054	---	Probable integrase (fragment)	1176928 - 1177242
Rv1067c	PE_PGRS19	PE-PGRS family protein PE PGRS19	1190424 - 1188421
Rv1068c	PE_PGRS20	PE-PGRS family protein PE PGRS20	1192148 - 1190757
Rv1087	PE_PGRS21	PE-PGRS family protein PE PGRS21	1211560 - 1213863
Rv1088	PE9	PE family protein PE9	1214513 - 1214947
Rv1089	PE10	PE family protein PE10	1214769 - 1215131
Rv1091	PE_PGRS22	PE-PGRS family protein PE PGRS22	1216469 - 1219030
Rv1135c	PPE16	PPE family protein PPE16	1264128 - 1262272
Rv1149	---	Possible transposase	1277893 - 1278300
Rv1168c	PPE17	PPE family protein PPE17	1299804 - 1298764
Rv1169c	lipX	PE family protein. Possible lipase LipX.	1300124 - 1299822
Rv1172c	PE12	PE family protein PE12	1302681 - 1301755
Rv1195	PE13	PE family protein PE13	1339003 - 1339302
Rv1196	PPE18	PPE family protein PPE18	1339349 - 1340524
Rv1199c	---	Possible transposase	1342605 - 1341358
Rv1214c	PE14	PE family protein PE14	1357625 - 1357293
Rv1243c	PE_PGRS23	PE-PGRS family protein PE PGRS23	1386677 - 1384989
Rv1313c	---	Possible transposase	1469505 - 1468171
Rv1325c	PE_PGRS24	PE-PGRS family protein PE PGRS24	1489965 - 1488154
Rv1361c	PPE19	PPE family protein PPE19	1533633 - 1532443
Rv1369c	---	Probable transposase	1542980 - 1541994
Rv1370c	---	Putative transposase for insertion sequence element IS6110 (fragment)	1543255 - 1542929
Rv1386	PE15	PE family protein PE15	1561464 - 1561772
Rv1387	PPE20	PPE family protein PPE20	1561769 - 1563388
Rv1396c	PE_PGRS25	PE-PGRS family protein PE PGRS25	1573857 - 1572127
Rv1430	PE16	PE family protein PE16	1606386 - 1607972
Rv1441c	PE_PGRS26	PE-PGRS family protein PE PGRS26	1619684 - 1618209
Rv1450c	PE_PGRS27	PE-PGRS family protein PE PGRS27	1634627 - 1630638
Rv1452c	PE_PGRS28	PE-PGRS family protein PE PGRS28	1638229 - 1636004
Rv1468c	PE_PGRS29	PE-PGRS family protein PE PGRS29	1656721 - 1655609
Rv1548c	PPE21	PPE family protein PPE21	1753333 - 1751297
Rv1573	---	Probable PhiRv1 phage protein	1779314 - 1779724
Rv1574	---	Probable PhiRv1 phage related protein	1779930 - 1780241
Rv1575	---	Probable PhiRv1 phage protein	1780199 - 1780699
Rv1576c	---	Probable PhiRv1 phage protein	1782064 - 1780643
Rv1577c	---	Probable PhiRv1 phage protein	1782584 - 1782072
Rv1578c	---	Probable PhiRv1 phage protein	1783228 - 1782758
Rv1579c	---	Probable PhiRv1 phage protein	1783623 - 1783309
Rv1580c	---	Probable PhiRv1 phage protein	1783892 - 1783620
Rv1581c	---	Probable PhiRv1 phage protein	1784301 - 1783906
Rv1582c	---	Probable PhiRv1 phage protein	1785912 - 1784497
Rv1583c	---	Probable PhiRv1 phage protein	1786310 - 1785912
Rv1584c	---	Possible PhiRv1 phage protein	1786528 - 1786307
Rv1585c	---	Possible phage PhiRv1 protein	1787099 - 1786584
Rv1586c	---	Probable PhiRv1 integrase	1788505 - 1787096
Rv1646	PE17	PE family protein PE17	1855764 - 1856696
Rv1651c	PE_PGRS30	PE-PGRS family protein PE PGRS30	1865382 - 1862347
Rv1705c	PPE22	PPE family protein PPE22	1932654 - 1931497
Rv1706c	PPE23	PPE family protein PPE23	1933878 - 1932694
Rv1753c	PPE24	PPE family protein PPE24	1984775 - 1981614
Rv1756c	---	Putative transposase	1988731 - 1987745
Rv1757c	---	Putative transposase for insertion sequence element IS6110 (fragment)	1989006 - 1988680
Rv1763	---	Putative transposase for insertion sequence element IS6110 (fragment)	1996152 - 1996478
Rv1764	---	Putative transposase	1996427 - 1997413
Rv1765A	---	Putative transposase (fragment)	1999357 - 1999142
Rv1768	PE_PGRS31	PE-PGRS family protein PE PGRS31	2000614 - 2002470
Rv1787	PPE25	PPE family protein PPE25	2025301 - 2026398
Rv1788	PE18	PE family protein PE18	2026477 - 2026776
Rv1789	PPE26	PPE family protein PPE26	2026790 - 2027971
Rv1790	PPE27	PPE family protein PPE27	2028425 - 2029477
Rv1791	PE19	PE family protein PE19	2029904 - 2030203

Supplementary Table 4.2. (cont.)

Rv1800	PPE28	PPE family protein PPE28	2039453 - 2041420
Rv1801	PPE29	PPE family protein PPE29	2042001 - 2043272
Rv1802	PPE30	PPE family protein PPE30	2043384 - 2044775
Rv1803c	PE_PGRS32	PE-PGRS family protein PE PGRS32	2046842 - 2044923
Rv1806	PE20	PE family protein PE20	2048072 - 2048371
Rv1807	PPE31	PPE family protein PPE31	2048398 - 2049597
Rv1808	PPE32	PPE family protein PPE32	2049921 - 2051150
Rv1809	PPE33	PPE family protein PPE33	2051282 - 2052688
Rv1818c	PE_PGRS33	PE-PGRS family protein PE PGRS33	2062674 - 2061178
Rv1840c	PE_PGRS34	PE-PGRS family protein PE PGRS34	2089518 - 2087971
Rv1917c	PPE34	PPE family protein PPE34	2167311 - 2162932
Rv1918c	PPE35	PPE family protein PPE35	2170612 - 2167649
Rv1983	PE_PGRS35	PE-PGRS family protein PE PGRS35	2226244 - 2227920
Rv2013	---	Transposase	2260665 - 2261144
Rv2014	---	Transposase	2261098 - 2261688
Rv2105	---	Putative transposase for insertion sequence element IS6110 (fragment)	2365465 - 2365791
Rv2106	---	Probable transposase	2365740 - 2366726
Rv2107	PE22	PE family protein PE22	2367359 - 2367655
Rv2108	PPE36	PPE family protein PPE36	2367711 - 2368442
Rv2123	PPE37	PPE family protein PPE37	2381071 - 2382492
Rv2126c	PE_PGRS37	PE-PGRS family protein PE PGRS37	2387972 - 2387202
Rv2162c	PE_PGRS38	PE-PGRS family protein PE PGRS38	2424838 - 2423240
Rv2167c	---	Probable transposase	2431145 - 2430159
Rv2168c	---	Putative transposase for insertion sequence element IS6110 (fragment)	2431420 - 2431094
Rv2177c	---	Possible transposase	2439947 - 2439282
Rv2278	---	Putative transposase for insertion sequence element IS6110 (fragment)	2550065 - 2550391
Rv2279	---	Probable transposase	2550340 - 2551326
Rv2328	PE23	PE family protein PE23	2600731 - 2601879
Rv2340c	PE_PGRS39	PE-PGRS family protein PE PGRS39	2618908 - 2617667
Rv2352c	PPE38	PPE family protein PPE38	2634098 - 2632923
Rv2353c	PPE39	PPE family protein PPE39	2635592 - 2634528
Rv2354	---	Probable transposase for insertion sequence element IS6110 (fragment)	2635628 - 2635954
Rv2355	---	Probable transposase	2635903 - 2636889
Rv2356c	PPE40	PPE family protein PPE40	2639535 - 2637688
Rv2371	PE_PGRS40	PE-PGRS family protein PE PGRS40	2651753 - 2651938
Rv2396	PE_PGRS41	PE-PGRS family protein PE PGRS41	2692799 - 2693884
Rv2408	PE24	Possible PE family-related protein PE24	2706017 - 2706736
Rv2424c	---	Probable transposase	2721777 - 2720776
Rv2430c	PPE41	PPE family protein PPE41	2727920 - 2727336
Rv2431c	PE25	PE family protein PE25	2728266 - 2727967
Rv2479c	---	Probable transposase	2785643 - 2784657
Rv2480c	---	Possible transposase for insertion sequence element IS6110 (fragment)	2785918 - 2785592
Rv2487c	PE_PGRS42	PE-PGRS family protein PE PGRS42	2797385 - 2795301
Rv2490c	PE_PGRS43	PE-PGRS family protein PE PGRS43	2806236 - 2801254
Rv2512c	---	Transposase for insertion sequence element IS1081	2829803 - 2828556
Rv2519	PE26	PE family protein PE26	2835785 - 2837263
Rv2591	PE_PGRS44	PE-PGRS family protein PE PGRS44	2921551 - 2923182
Rv2608	PPE42	PPE family protein PPE42	2935046 - 2936788
Rv2615c	PE_PGRS45	PE-PGRS family protein PE PGRS45	2944985 - 2943600
Rv2634c	PE_PGRS46	PE-PGRS family protein PE PGRS46	2962441 - 2960105
Rv2646	---	Probable integrase	2970551 - 2971549
Rv2648	---	Probable transposase for insertion sequence element IS6110 (fragment)	2972160 - 2972486
Rv2649	---	Probable transposase for insertion sequence element IS6110	2972435 - 2973421
Rv2650c	---	Possible PhiRv2 prophage protein	2975234 - 2973795
Rv2651c	---	Possible PhiRv2 prophage protease	2975775 - 2975242
Rv2652c	---	Probable PhiRv2 prophage protein	2976554 - 2975928
Rv2653c	---	Possible PhiRv2 prophage protein	2976909 - 2976586
Rv2654c	---	Possible PhiRv2 prophage protein	2977234 - 2976989
Rv2655c	---	Possible PhiRv2 prophage protein	2978658 - 2977231
Rv2656c	---	Possible PhiRv2 prophage protein	2979052 - 2978660

Supplementary Table 4.2. (cont.)

Rv2657c	---	Probable PhiRv2 prophage protein	2979309 - 2979049
Rv2659c	---	Probable PhiRv2 prophage integrase	2980818 - 2979691
Rv2666	---	Probable transposase for insertion sequence element IS1081 (fragment)	2983071 - 2983874
Rv2741	PE_PGRS47	PE-PGRS family protein PE PGRS47	3053914 - 3055491
Rv2768c	PPE43	PPE family protein PPE43	3078078 - 3076894
Rv2769c	PE27	PE family protein PE27	3078985 - 3078158
Rv2770c	PPE44	PPE family protein PPE44	3080457 - 3079309
Rv2791c	---	Probable transposase	3101581 - 3100202
Rv2810c	---	Probable transposase	3116142 - 3115741
Rv2812	---	Probable transposase	3116818 - 3118227
Rv2814c	---	Probable transposase	3121552 - 3120566
Rv2815c	---	Probable transposase	3121827 - 3121501
Rv2853	PE_PGRS48	PE-PGRS family protein PE PGRS48	3162268 - 3164115
Rv2885c	---	Probable transposase	3195548 - 3194166
Rv2892c	PPE45	PPE family protein PPE45	3202020 - 3200794
Rv2943	---	Probable transposase for insertion sequence element IS1533	3288464 - 3289705
Rv2943A	---	Possible transposase	3289705 - 3290235
Rv2944	---	Possible transposase for insertion sequence element IS1533	3289790 - 3290506
Rv2961	---	Probable transposase	3313283 - 3313672
Rv2978c	---	Probable transposase	3335164 - 3333785
Rv3018c	PPE46	PPE family protein PPE46	3378243 - 3376939
Rv3018A	PE27A	PE family protein PE27A	3378415 - 3378329
Rv3021c	PPE47	PPE family protein PPE47	3380452 - 3379376
Rv3022c	PPE48	PPE family protein PPE48	3380682 - 3380440
Rv3022A	PE29	PE family protein PE29	3380993 - 3380679
Rv3023c	---	Probable transposase	3382622 - 3381375
Rv3115	---	Probable transposase	3481451 - 3482698
Rv3125c	PPE49	PPE family protein PPE49	3491651 - 3490476
Rv3135	PPE50	PPE family protein PPE50	3501334 - 3501732
Rv3136	PPE51	PPE family protein PPE51	3501794 - 3502936
Rv3144c	PPE52	PPE family protein PPE52	3511317 - 3510088
Rv3159c	PPE53	PPE family protein PPE53	3529163 - 3527391
Rv3184	---	Probable transposase for insertion sequence element IS6110 (fragment)	3551281 - 3551607
Rv3185	---	Probable transposase	3551556 - 3552542
Rv3186	---	Probable transposase for insertion sequence element IS6110 (fragment)	3552764 - 3553090
Rv3187	---	Probable transposase	3553039 - 3554025
Rv3191c	---	Probable transposase	3558345 - 3557311
Rv3325	---	Probable transposase for insertion sequence element IS6110 (fragment)	3710433 - 3710759
Rv3326	---	Probable transposase	3710708 - 3711694
Rv3327	---	Probable transposase fusion protein	3711749 - 3713461
Rv3343c	PPE54	PPE family protein PPE54	3736935 - 3729364
Rv3344c	PE_PGRS49	PE-PGRS family protein PE PGRS49	3738438 - 3736984
Rv3345c	PE_PGRS50	PE-PGRS family protein PE PGRS50	3742774 - 3738158
Rv3347c	PPE55	PPE family protein PPE55	3753184 - 3743711
Rv3348	---	Probable transposase	3753765 - 3754256
Rv3349c	---	Probable transposase	3755033 - 3754293
Rv3350c	PPE56	PPE family protein PPE56	3767102 - 3755952
Rv3367	PE_PGRS51	PE-PGRS family protein PE PGRS51	3778568 - 3780334
Rv3380c	---	Probable transposase	3796086 - 3795100
Rv3381c	---	Probable transposase for insertion sequence element IS6110 (fragment)	3796361 - 3796035
Rv3386	---	Possible transposase	3800092 - 3800796
Rv3387	---	Possible transposase	3800786 - 3801463
Rv3388	PE_PGRS52	PE-PGRS family protein PE PGRS52	3801653 - 3803848
Rv3425	PPE57	PPE family protein PPE57	3842239 - 3842769
Rv3426	PPE58	PPE family protein PPE58	3843036 - 3843734
Rv3427c	---	Possible transposase	3844640 - 3843885
Rv3428c	---	Possible transposase	3845970 - 3844738
Rv3429	PPE59	PPE family protein PPE59	3847165 - 3847701
Rv3430c	---	Possible transposase	3848805 - 3847642
Rv3474	---	Possible transposase for insertion element IS6110 (fragment)	3890830 - 3891156

Supplementary Table 4.2. (cont.)

Rv3475	---	Possible transposase for insertion element IS6110 [second part]	3891105 - 3892091
Rv3477	PE31	PE family protein PE31	3894093 - 3894389
Rv3478	PPE60	PE family protein PPE60	3894426 - 3895607
Rv3507	PE_PGRS53	PE-PGRS family protein PE PGRS53	3926569 - 3930714
Rv3508	PE_PGRS54	PE-PGRS family protein PE PGRS54	3931005 - 3936710
Rv3511	PE_PGRS55	PE-PGRS family protein PE PGRS55	3939617 - 3941761
Rv3512	PE_PGRS56	PE-PGRS family protein PE PGRS56	3941724 - 3944963
Rv3514	PE_PGRS57	PE-PGRS family protein PE PGRS57	3945794 - 3950263
Rv3532	PPE61	PPE family protein PPE61	3969343 - 3970563
Rv3533c	PPE62	PPE family protein PPE62	3972453 - 3970705
Rv3539	PPE63	PPE family protein PPE63	3978059 - 3979498
Rv3558	PPE64	PPE family protein PPE64	3997980 - 3999638
Rv3590c	PE_PGRS58	PE-PGRS family protein PE PGRS58	4033158 - 4031404
Rv3595c	PE_PGRS59	PE-PGRS family protein PE PGRS59	4038050 - 4036731
Rv3621c	PPE65	PPE family protein PPE65	4061889 - 4060648
Rv3622c	PE32	PE family protein PE32	4062198 - 4061899
Rv3636	---	Possible transposase	4075752 - 4076099
Rv3637	---	Possible transposase	4076484 - 4076984
Rv3638	---	Possible transposase	4076984 - 4077730
Rv3640c	---	Probable transposase	4079749 - 4078520
Rv3650	PE33	PE family protein PE33	4091233 - 4091517
Rv3652	PE_PGRS60	PE-PGRS family-related protein PE PGRS60	4093632 - 4093946
Rv3653	PE_PGRS61	PE-PGRS family-related protein PE PGRS61	4093940 - 4094527
Rv3738c	PPE66	PPE family protein PPE66	4190232 - 4189285
Rv3739c	PPE67	PPE family protein PPE67	4190517 - 4190284
Rv3746c	PE34	Probable PE family protein PE34 (PE family-related protein)	4196506 - 4196171
Rv3751	---	Probable integrase (fragment)	4198874 - 4199089
Rv3798	---	Probable transposase	4252993 - 4254327
Rv3812	PE_PGRS62	PE-PGRS family protein PE PGRS62	4276571 - 4278085
Rv3827c	---	Possible transposase	4302789 - 4301563
Rv3844	---	Possible transposase	4318775 - 4319266
Rv3872	PE35	PE family-related protein PE35	4350745 - 4351044
Rv3873	PPE68	PPE family protein PPE68	4351075 - 4352181
Rv3892c	PPE69	PPE family protein PPE69	4375683 - 4374484
Rv3893c	PE36	PE family protein PE36	4375995 - 4375762
Rv0287*	esxG	ESAT-6 like protein EsxG (conserved protein TB9.8)	351525 - 351818
Rv1037c*	esxI	Putative ESAT-6 like protein EsxI (ESAT-6 like protein 1)	1160828 - 1160544
Rv1038c*	esxJ	ESAT-6 like protein EsxJ (ESAT-6 like protein 2)	1161151 - 1160855
Rv1197*	esxK	ESAT-6 like protein EsxK (ESAT-6 like protein 3)	1340659 - 1340955
Rv1198*	esxL	Putative ESAT-6 like protein EsxL (ESAT-6 like protein 4)	1341006 - 1341290
Rv1792*	esxM	ESAT-6 like protein EsxM	2030347 - 2030643
Rv1793*	esxN	Putative ESAT-6 like protein EsxN (ESAT-6 like protein 5)	2030694 - 2030978
Rv2346c*	esxO	Putative ESAT-6 like protein EsxO (ESAT-6 like protein 6)	2626172 - 2625888
Rv2347c*	esxP	Putative ESAT-6 like protein EsxP (ESAT-6 like protein 7)	2626519 - 2626223
Rv3017c*	esxQ	ESAT-6 like protein EsxQ (TB12.9) (ESAT-6 like protein 8)	3376852 - 3376490
Rv3019c*	esxR	Secreted ESAT-6 like protein EsxR (TB10.3) (ESAT-6 like protein 9)	3379001 - 3378711
Rv3020c*	esxS	ESAT-6 like protein EsxS	3379329 - 3379036
Rv3444c*	esxT	Putative ESAT-6 like protein EsxT	3862926 - 3862624
Rv3445c*	esxU	ESAT-6 like protein EsxU	3863264 - 3862947
Rv3619c*	esxV	Putative ESAT-6 like protein EsxV (ESAT-6 like protein 1)	4060268 - 4059984
Rv3620c*	esxW	Putative ESAT-6 like protein EsxW (ESAT-6 like protein 10)	4060591 - 4060295
Rv3875*	esxA	6 kDa early secretory antigenic target EsxA (ESAT-6)	4352609 - 4352896
Rv3890c*	esxC	ESAT-6 like protein EsxC (ESAT-6 like protein 11)	4374013 - 4373726
Rv3891c*	esxD	Possible ESAT-6 like protein EsxD	4374372 - 4374049
Rv3904c*	esxE	Putative ESAT-6 like protein EsxE (hypothetical alanine rich protein) (ESAT-6 like protein 12)	4390709 - 4390437
Rv3905c*	esxF	Putative ESAT-6 like protein EsxF (hypothetical alanine and glycine rich protein) (ESAT-6 like protein 13)	4391031 - 4390720
Rv2543*	lppA	Probable conserved lipoprotein LppA	2866468 - 2867127
Rv2544*	lppB	Probable conserved lipoprotein LppB	2867124 - 2867786
Rv2048c*	pks12	Polyketide synthase Pks12	2306986 - 2294531