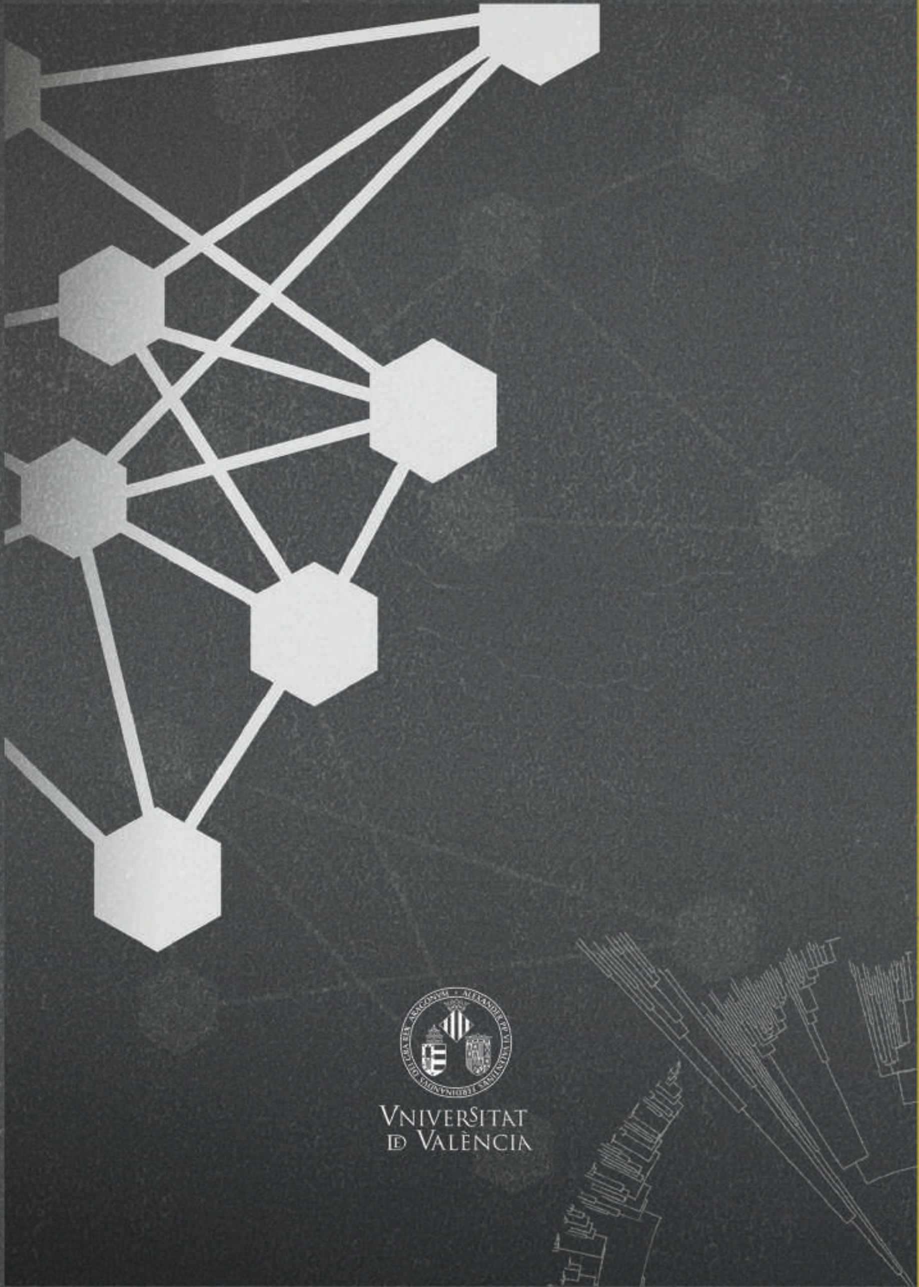


VNIVERSITAT DE VALÈNCIA

VNIVERSITAT DE VALÈNCIA

DECODING TUBERCULOSIS TRANSMISSION AND DRUG RESISTANCE IN VALENCIA REGION USING WHOLE GENOME SEQUENCING

Irving Norberto Cancino Muñoz

VNIVERSITAT DE VALÈNCIA

Irving Norberto Cancino Muñoz
PhD thesis

Iñaki Comas Espadas
Supervisor

DECODING TUBERCULOSIS TRANSMISSION AND DRUG RESISTANCE IN VALENCIA REGION USING WHOLE GENOME SEQUENCING

- DOCTORAL PROGRAMME IN BIOMEDICINE AND BIOTECHNOLOGY, JULY 2020 -

VNIVERSITAT DE VALÈNCIA

Decoding tuberculosis transmission and drug resistance in Valencia Region using whole genome sequencing

Author: Irving Norberto Cancino Muñoz

SUPERVISOR
Iñaki Comas Espadas

PhD thesis
Doctoral Programme in Biomedicine and Biotechnology

Valencia, 2020

D. Iñaki Comas Espadas, científico titular del Instituto de Biomedicina de Valencia (IBV-CSIC), en calidad de Director de la tesis doctoral de D. Irving Norberto Cancino Muñoz, adscrito al Programa de Doctorado en Biomedicina y Biotecnología de la Universitat de València.

CERTIFICA

Que la tesis titulada *"Decoding tuberculosis transmission and drug resistance in Valencia Region using whole genome sequencing"* se ha desarrollado bajo su dirección y supervisión, y que el trabajo de investigación realizado y la memoria del mismo, ha sido elaborada por el doctorado y cumple los requisitos científicos y formales para proceder al acto de defensa de la Tesis Doctoral.

Y para que conste, en el cumplimiento de la legislación presente, firman el presente certificado en València, 2020.

Dr. Iñaki Comas Espadas
Director

Dr. Juan Ferre Manzanero
Tutor Académico

Irving Cancino Muñoz
Doctorando

AGRADECIMIENTOS

Desde pequeño mi familia me ha inculcado a ser agradecido con todas aquellas personas que de una forma u otra me han ayudado y apoyado en algún momento de mi vida. Han sido 5 años en los que he aprendido mucho, y en los que he conocido personas que llegaron para quedarse en mi vida. Por eso, me gustaría agradecerles a cada una de ellas.

En primer lugar quiero agradecer a mi tutor y mentor Iñaki Comas Espadas, quien me dio la oportunidad de realizar este proyecto. Has sido un guía e inspiración para mi, tanto para el desarrollo de esta tesis, como de manera profesional y personal (algo así como un *master Jedi*). Tus comentarios y consejos siempre acertados y críticos, han sido de gran ayuda para adaptarme en esta etapa profesional. De verdad, muchas gracias. Siempre ha sido, es y será un placer trabajar contigo.

En segundo lugar, quiero agradecer a mi familia. A mi abuela, que con su cariño y paciencia, me ha enseñado y formado a ser la persona que soy ahora. A mis padres, que siempre han estado apoyándome incondicionalmente en la distancia. A mi tía Yoreley, que con sus consejos me ha ayudado a adaptarme a vivir fuera de casa. Al resto de mi familia, gracias por todo su apoyo y amor.

De la misma manera, quiero agradecer a mis amigos/as del laboratorio. Por orden cronológico: Luis, Galo, Vicky, Miguel, Manoli, Álvaro, Carla, Mariana y Ana. Ustedes siempre han estado ahí para echarme una mano, de buen humor y con una sonrisa. Sin su ayuda en el laboratorio, esta tesis no estaría terminada. Aprovecho para mencionar a las personas de FISABIO y del IBV: Jorge, Rodrigo, Alicia, Laura y Marina. Gracias a ustedes por los buenos momentos que pasamos en todas las reuniones y salidas fuera del laboratorio.

A mis amigo/as fuera del laboratorio: Fabricio, Félix, Paco, Christian, Khalid, Campo, Mara, Bob, Aldo y Eduardo, por apoyarme y darme ánimos cuando estaba estresado. Gracias por aguantarme noches completas escuchándome hablar de tuberculosis y por meterme presión por terminar la tesis.

Quiero agradecer a lo/as investigadore/as con los que colaboramos para publicar algunos los resultados obtenidos en esta tesis. Por último, quiero agradecer al CONACYT por brindarme una beca para realizar mi doctorado en el extranjero. Sin esa ayuda, no hubiese podido llegar a donde estoy.

Seguramente me olvidé de mencionar a otras personas que me ayudaron y estuvieron presentes en algún momento durante la realización de este trabajo. Lo siento. ¡Muchas gracias a todos y todas!.

”Infectious disease which antedated the emergence of the humankind will last as long as humanity itself, and will surely remain, as it has been hitherto, one of the fundamental parameters and determinants of human history.”

— William H. McNeill, *Plagues and Peoples*

Contents

General introduction

Tuberculosis (TB) is an infectious disease caused by the airborne transmitted bacterial pathogens belonging to the *Mycobacterium tuberculosis* complex (MTBC). Tuberculosis is a curable and preventable disease although it is one of the top ten causes of death worldwide according to the World Health Organization (WHO). TB can be transmitted by air droplets from human to human, so, the most common infection site are the lungs. However, TB infections in different organs are reported in lower frequency. Individuals with TB are treated with a six-month regimen of four antibiotics, nevertheless, this treatment can be extended up to 20 months with less effective drugs in patients infected with a drug resistant strain. It is estimated that about one-quarter of the worldwide population has latent TB, which is characterized by no suffering TB symptoms and presumably non transmissible disease.

The MTBC is a group of highly related, slow-growing mycobacteria which causes tuberculosis in humans and animals [1]. The causative agents of tuberculosis in humans are *Mycobacterium tuberculosis*, and the highly related lineage *Mycobacterium africanum*, both with no know reservoir outside the human host. The MTBC also includes the animal-adapted mycobacteria named *M.bovis*, *M. caprae*, *M. microti*, *M. pinnipedii*, *M. orygis*, *M. suricattae*, and *M.mungi* [2].

The genus *Mycobacterium* comprises more than 170 species that share unique phenotypic and genotypic characteristics [3]. All the involved species have a high content of G/C ranging from 60 to 72%. Although most *Mycobacterium* species are free living organisms that do not cause TB disease [4], some of them can cause infections within human populations [5]. These non-tuberculous mycobacterial infections included cases by *M.avium*, *M.kansasii* and others [6]. A major physiological feature of the *Mycobacterium* genus is the unusual cell wall structure [7, 8]. This cell barrier consisted of a lipid bilayer made of long fatty acids (mycolic acids), and waxy components providing a hydrophobic permeable barrier that confer resistance to harmful compounds and avoid dehydration [9]. Due to this, mycobacterial species are not easy to decolorize and are considered acid fast bacilli. In fact, they need a special staining method called Ziehl-Neelsen staining [8]. Additionally, all mycobacterial species are aerobic, non-spore-forming, nonmotile, and have a rod-shaped form. Their size ranged from 0.2 to 0.6μ by 1.0 to 10μ, forming colonies that morphologically varies in texture and color among species [7]. MTBC strains are slow growing bacteria, with a generation time of 12-24h on commonly media growth.

1.1 Global burden of tuberculosis

The last report of the WHO estimated that 10 millions of people got infected with TB in 2018 [10]. Of these, 1.45 million cases died because of TB disease (484,000 due to drug resistant tuberculosis, 251,000 with HIV coinfection). With these numbers, TB is considered the leading cause of death by a single infectious disease overcoming the HIV/AIDS infection. The most affected regions are South-East Asia and Africa with 44% and 24% of all TB notified cases, respectively. The countries that contributed with two-thirds of all reported cases are India (27%), China (9%), Indonesia (8%), the Philippines (6%), Pakistan (6%), Nigeria (4%), Bangladesh (4%) and South Africa (3%) [10](**Figure 1.1A**).

The increasing prevalence of cases that fail to respond to standard antibiotic treatment is a main threat to global TB control efforts. Multidrug-resistant TB

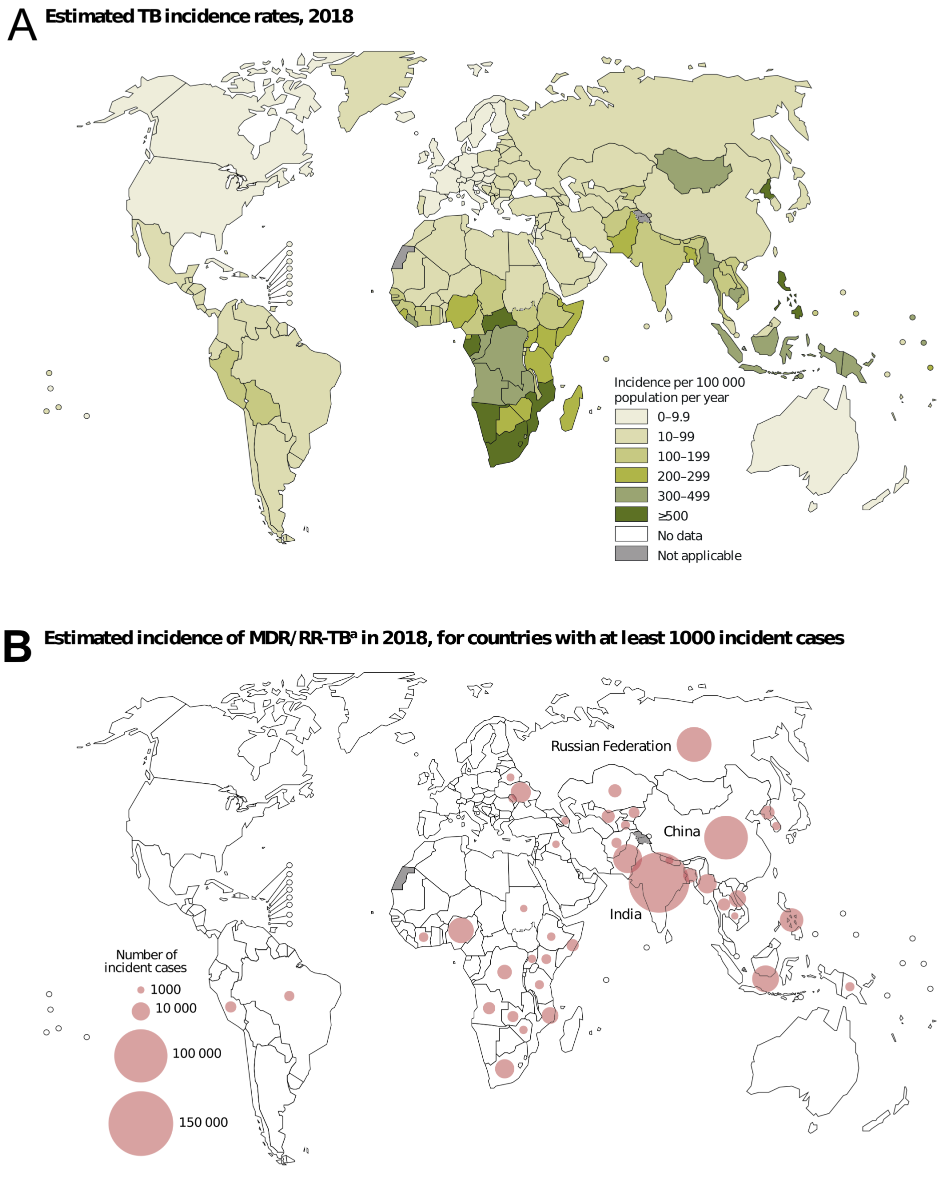
(MDR-TB) are those infections that are resistant to rifampicin and isoniazid simultaneously, the most powerful antibiotics against TB. During 2018 it was estimated that 484,000 of new cases were caused by MDR-TB or rifampicin resistant TB (RR-TB), accounting for 3.4% and 18% of all globally new and previously treated TB cases, respectively. Fifty percent of all MDR-TB/RR-TB cases are concentrated in India (27%), China (14%) and the Russian Federation (9%) [10] (**Figure 1.1B**).

1.1.1 Tuberculosis in Europe

In 2018 there were reported 259,000 tuberculosis cases equivalent to a TB incidence of 28 cases per 100,000 inhabitants [10]. This corresponds to about 2.6% of all globally TB cases, being the continent with the lowest TB incidence value. Nevertheless, the TB rate is highly variable among countries, with higher rates in Eastern Europe than in the Western region. The countries with the highest incidence rate were Kyrgyzstan (144 per 100,000 population), followed by the Republic of Moldova (95), Georgia (86), Tajikistan (85) and Ukraine (84) [11].

Spain is in the 18th position of TB incidence of the forty-two European countries. In 2017, the Spanish surveillance system reported 4,483 TB cases (incidence rate of 9.6 per 100,000 inhabitants), being Galicia, Catalunya and Asturias the regions with the highest TB incidence rate with 19.6, 12.9 and 10.8 cases per 100,000 people, respectively [12]. Particularly, in the Comunidad Valenciana, the study region of this thesis, were reported 424 TB cases and an incidence close to the mean of the country (8.6 per 100,000 inhabitants).

Regarding TB drug resistance (DR), the European countries with the major MDR-TB prevalence are the Russian Federation (28.2%), the Republic of Moldova (26.6%), Kyrgyzstan (22.4%) and Ukraine (21.4%). Notably, according to the national public health system, the MDR-TB incidence in Spain is remarkably low, with just 20-30 MDR-TB cases reported every year [11].



A Estimated TB incidence rates, 2018

B Estimated incidence of MDR/RR-TB[a] in 2018, for countries with at least 1000 incident cases

[a] MDR-TB is a subset of RR-TB.

Figure 1.1: Estimated global incidence of TB incidence rates (A) and MDR/RR-TB cases (B) during 2018. Source: WHO [10]

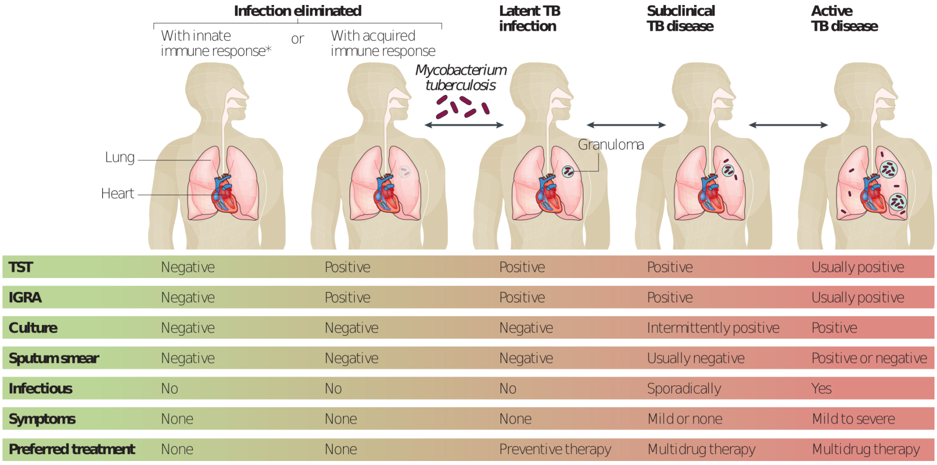
1.1.2 The End TB Strategy

Currently, the WHO has set a series of targets to stop TB around the world. This initiative was called End TB Strategy, approved in 2014 by the World Health Assembly [13]. Its main objectives are to reduce the number of TB deaths by 90% and decrease the TB incidence by 80% by 2030, especially on high-burden countries [10]. Although the number of TB cases have been decreasing since the 1990s, the majority of the countries worldwide still have an incidence rate higher than 10 per 100,000 inhabitants. In order to reach the objectives stated by the End TB Strategy, the WHO has released a report called “Implementing the End TB strategy: the essentials” with the aim to guide actions to help countries on how to implement this plan. It states three main pillars to accomplish the end of the TB epidemic, (1) Integrated, patient-centred care and prevention, (2) Bold policies and supportive systems and (3) Intensification of research and innovation. These guides highlight, among others, the importance of early and rapid TB diagnosis, accurate and fast drug susceptibility testing, application of appropriate treatments and effective control of the transmission. Additionally, it emphasizes the requirement of research and development of new diagnostics, drugs, vaccines and innovative delivery methods [13].

1.2 Tuberculosis infection

TB infection begins when the tubercle bacillus enters the lungs via inhalation, reaches the alveolar space and encounters the resident alveolar macrophages, where the bacteria replicates [14]. Once the infection is established, there are two different scenarios depending on disease severity: 1) the immune system can control the bacteria growth (but do not eliminate the initial infection) and the individual is infected without disease symptoms. This asymptomatic condition is known as latent TB infections (LTBI). Nevertheless, 2) there is a small percentage of people (between 5-15%) in whom the immune system fails and develop active disease, presenting mild, moderate or severe symptoms, or even developing lung cavitations. Active TB cases can transmit the pathogen to other individuals [15]. Additionally, there is approximately a 10% lifetime risk of developing active TB from LTBI [16]. However, the factors that trigger infection progress and the underlying bacterial metabolic changes remain unclear and could involve host's immunological, clinical, and bacterial factors [17]. The most common methods for the detection of LTBI are the positivity of the tuberculin skin test (TST), and the IFN- release assay (IGRA) tests, both based on immunological responses by the host but not able to distinguish between latent and active tuberculosis.

For many years the dichotomy between active and latent TB clinical stages has been used to differentiate between those that are infected and can remain asymptomatic during years or lifelong and those that develop typical TB symptoms (continuous coughing, high fever, night sweats, fatigue, chest pain). However, there is a growing consensus that TB infection is better reflected by the existence of a wide spectrum of different infection status [18, 19], which are mediated by the heterogeneous nature of the bacterial dynamics and the host immune responses in the granulomas [15] (**Figure 1.2**). The heterogeneity associated with LTBI outcomes has been mainly shown by analyzing the whole blood transcriptome signature of latent and active TB patients [20], and by positron emission tomography combined with computed tomography technology (PET/CT scan) within lung lesions [21]. Although a reduced host gene-set of transcriptome signature is promising, the classification of patients into these different infection status is still difficult [22]. In clinical settings, a combination of clinical disease manifestations, microbiological evidence, and immunological tests (TST and IGRA) results can help to define different LTBI status [19]. For instance, it has been described a group of contacts called "resisters", which are patients that never developed TB diseases even if they are constantly exposed to an active case. These individuals seemed to be resistant to *M.tuberculosis* infection due to the fact that they do not present any TB symptoms and negative TST/IGRA outcomes even though they are constantly exposed to TB infection. In addition, these individuals are presumed to not transmit the disease. A recent study has identified household contacts with no evidence of infection even two years after index cases exposures [23]. Another interesting and recently discovered infection status is subclinical TB. This TB infection is characterized by having no TB symptoms but with positive results of sputum and/or culture and immunological assays, and sometimes positive chest radiographs [15]. These TB cases are difficult to detect mainly because people without clear symptoms do not go to a medical care facility for further examination. Instead, the diagnosis of these cases are found by household contact investigations, representing major challenges for public health systems. A recent study for example has identified presence of subclinical TB in patients otherwise healthy by examining the bacterial content of exhaled secretions in masks [24]. It is important to differentiate subclinical TB infection from classic LTBI, mainly because a percentage of these latent individuals are preventelly treated with isoniazid once identified by epidemiological interventions. Thus, there is the danger that those individuals with subclinical disease may acquire isoniazid resistance even before showing symptoms [25]. Whether these patients can transmit the disease is still in debate but will be the focus of the present dissertation.



	Infection eliminated: With innate immune response*	Infection eliminated: With acquired immune response	Latent TB infection	Subclinical TB disease	Active TB disease
TST	Negative	Positive	Positive	Positive	Usually positive
IGRA	Negative	Positive	Positive	Positive	Usually positive
Culture	Negative	Negative	Negative	Intermittently positive	Positive
Sputum smear	Negative	Negative	Negative	Usually negative	Positive or negative
Infectious	No	No	No	Sporadically	Yes
Symptoms	None	None	None	Mild or none	Mild to severe
Preferred treatment	None	None	Preventive therapy	Multidrug therapy	Multidrug therapy

Figure 1.2: The spectrum of tuberculosis, from *Mycobacterium tuberculosis* infection to active (pulmonary) tuberculosis disease. Taken and adapted from [14].

1.3 Tuberculosis transmission

TB can be transmitted in aerosolized droplets by coughing from a person with active disease. Thus, the most common infection site is the respiratory system but any organ can be infected. In addition, many other non-tuberculous mycobacteria can cause disease in humans, however, these infections tend to be associated with extra-pulmonary disease and their transmissibility among individuals are unlikely [4]. Recent transmission (that occurring within a window of a few years) is the main contributor to TB cases in high burden countries [26] and is also a major contributor in low burden countries [27, 28]. Interrupting transmission is essential in order to reduce the TB incidence and to advance towards eradication.

1.3.1 Factors associated with transmission

There are many factors associated with TB transmission, these include host and bacterial features, but also social and behavioral factors (**Figure 1.3**). Host factors are likely related with disease infectiousness, thus, individuals with more severe pulmonary disease (for example, with larger lung cavities) are more likely to transmit TB [29]. This may be related to the coughing frequency and the bacillary load. However, it has been shown a low correlation between bacillary load in the sputum and infectiousness [30] and it is estimated that around 30% of index cases are sputum negative [31]. Other determinants that affect progression to active disease involve individual social behaviours such as smoking [32] and alcohol abuse [33]. In addition, factors like malnutrition [34] and comorbidities such as HIV infections and diabetes are related with disease susceptibility [35, 36]. Closer interactions with these more infectious individuals affect disease transmissibility. In addition to this, delay in TB diagnosis as well as the initiation of TB treatment, increase the probability of disease transmission through population [37]. Associated factors to this delayed status include patient awareness of the symptoms and the efficiency to detect a TB infection by healthcare systems [38]. Population-level social patterns are influenced by age and demographic structure, cultural behaviors and migration patterns. For example, in high-burden settings, TB transmission occurred more frequently in outdoors scenarios rather than indoors [26]. On the contrary, TB incidence in older populations is related with LTBI [18].

Regarding bacterial factors, it has been described that there are differences of transmission among MTBC lineages. For instance, in a cosmopolitan setting like San Francisco, MTBC lineages tend to transmit better in specific human population [39]. In fact, it is described that some MTBC lineages are "generalists" or "specialists" according to their geographic spread [40], whether this uneven distribution is due to historical contingency or to bacterial biological factors is still under debate.

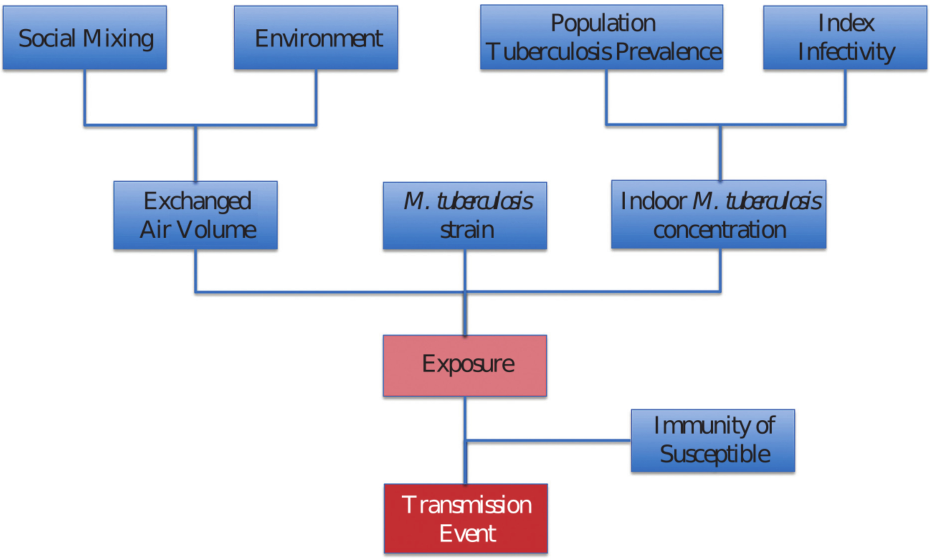


Figure 1.3: Factors influencing tuberculosis transmission. Taken from [41]

1.3.2 How to measure transmission

Cutting TB transmission is a cornerstone of TB control. Nevertheless, measuring tuberculosis transmission is complex due to the natural history of the bacteria, resulting in a few patients developing the disease [41]. Also there is not a unique tool to evaluate transmission as some measure infection (like TST and IGRA), while others are limited to active TB cases as they are based on the similarity of the infecting bacteria (molecular genotyping) [26]. Gobally, epidemiological interventions to detect and control TB transmission mainly focus on the identification of active TB cases coupled with contract tracing. This passive case finding strategy assumes that a person with TB symptoms will seek a healthcare facility to be diagnosed and treated [42, 43], while the main goal of contact tracing is to reduce time required to detect and treat a case by identifying secondary cases among active TB patients, thus cutting downstream transmission [44]. Contact tracing combines epidemiological surveys to close contacts of active TB cases with examination of clinical evidence such as chest radiographies, bacili detection in sputum and TST/IGRA testing [45]. Contact questionnaires include information regarding people that likely had contact when symptoms started, and their respective social-behavior habits and thus heavily relies on the assumption that prolonged contact is necessary for transmission. Contact tracing has been shown to improve successfully the detection of TB cases within close contacts cases [45, 43]. A meta-analysis described that contact tracing investigations have proven to better estimate the prevalence of active TB and LTBI within contacts in high-, middle and low-income countries [46]. For many years, contact tracing has been the standard practice used as a control intervention in high-income settings [47, 48], in fact, it is recommended practice in Spain [12]. Nevertheless, its implementation in middle- and low-income countries is poor [49]. Moreover, its cost-effectiveness value in National TB programmes is still unknown.

Since the early 1990's, molecular approaches to investigate TB transmission have been developed. These genotyping tools have improved our understanding of TB transmission dynamics by revealing cases belonging to a transmission cluster. In addition, genotyping approaches can estimate transmission at population level, instead of individual level assessments from classic contract tracing [50]. Due to their fast and replicable application in TB surveillance, molecular techniques combined with epidemiological strategies have helped to resolve TB investigations as well as to identify and establish risk factors for transmission [51, 52]. Furthermore, molecular-based approaches such as MIRU-VNTR have been widely used globally to track specific MTBC strains in different populations [53]. However, their implementation at the population-based level has been scarced, particularly because many links could not be corroborated by epidemiological investigations. Now we know that these techniques overestimate the number of cases that are part of TB transmission clusters [54, 55]. Due to this, they are not implemented in public health systems, especially in low- and middle-income countries.

More recently, WGS has started to be used as a tool to detect TB transmission. In principle, WGS provides a greater resolution than traditional molecular approaches to trace infection sources and delineate transmission networks [56]. In addition, WGS is getting cheaper and offers a cost-effective alternative for investigating TB transmission as the agreement with epidemiological data is greater than previous tools [144]. A further discussion is detailed below and in **chapters 3 and 4** of this dissertation.

1.4 Tuberculosis treatment

Treatment of TB aims to cure all the patients that had active or LTBI to stop the transmission of the disease or at least minimize it. Thus, the objectives of TB therapy are to reduce the number of growing bacilli within the patient; to eradicate infecting bacteria populations in order to prevent a relapse episode and the development of MDR-TB during therapy [57].

1.4.1 Drug-susceptible TB

The standard treatment for patients with drug-susceptible TB lasts at least 6 months. WHO recommendations consist of an initial 2-months intensive phase (rifampicin, isoniazid, pyrazinamide, and ethambutol every day), followed by 4-months of isoniazid and rifampicin. This effective anti-TB-drugs are known as first-line drugs. This first-line regime costs around US$20 and its clinical success is approximately 85% of all newly diagnosed TB cases [58]. However, it is a long therapy and not well tolerated by some patients. Some studies described side effects and an increased risk of developing a relapse in individuals who had large pulmonary cavities [59, 60] and slow response to first-line treatment [61, **?**]. Additionally, many of these cases end up developing resistance to one or more drugs. In 2017, It was estimated that 4.1% of all globally new TB cases had multidrug-resistant or rifampicin-resistant TB. This proportion increased up to 19% in those cases that were previously treated with TB, also known as relapse cases [10]. In order to reduce these adverse reactions as well as the number of relapses, some short-term treatments have been proposed [58]. However, the majority of these studies are still in the clinical trial phase.

1.4.2 Rifampicin

The rifampicin is considered one of the most powerful first-line drugs against TB. The mechanism of action of rifampicin is to inhibit bacterial transcription by targeting RNA polymerase subunit [62] that is encoding by the *rpoB* gene. Unfortunately, rifampicin drug-resistant isolates were reported shortly after the drug introduction as primary treatment in 1966, especially when rifampicin was the only active drug administered [63, 64]. Around 95% of all mutations conferring rifampicin resistance are located in a 81bp *rpoB* region called rifampicin resistance determining region (RRDR) [65]. Nevertheless, less common mutations outside this region had been described elsewhere [66]. This topic is discussed in-depth in **chapter 5** of this thesis.

1.4.3 Isoniazid

Isoniazid is a first-line pro-drug that blocks the synthesis of mycolic acids, which are one of the main and most important components of the *M. tuberculosis* cell wall [67]. Isoniazid is activated by the catalase-peroxidase *katG* gene [68]. Isoniazid resistant isolates commonly lack catalase and peroxidase activity. Thus, between 83-96% of globally isoniazid related mutations occurred within *katG* gene [69]. There are reported another isoniazid resistant clinical strains that harbour mutations in other regions such as *inhA*,*KasA* and *ahpC* promoter and coding genes, all related with the synthesis of mycolic acid and likely playing a compensatory function for the loss of catalase-peroxidase activity [70].

1.4.4 Ethambutol

Ethambutol targets the bacterial cell wall by inhibiting the synthesis of cell wall arabinan [71]. Ethambutol resistant strains typically harboured mutations within the *embCAB* operon, which is involved in the synthesis of the cell wall arabinan. Thus, the majority of the ethambutol resistant strains have mutations in the corresponding coding genes, especially in the *embB* gene. However, other target genes related to ethambutol drug resistance had been described. For example, *Rv3806c* and *Rv3792* genes seemed to affect the synthesis or utilization of DPA (decaprenylphosphoryl--d-arabinose) pathway, resulting in high-level resistance [72].

1.4.5 Pyrazinamide

Pyrazinamide is another pro-drug that is converted to pyrazionic acid by the enzyme pyrazinamidase, which is encoded by the well-known *pncA* gene [73]. This first-line antibiotic is important because it inhibits the bacterial growth in acidic environments (i.e. inside the macrophages). Moreover, pyrazinamide is widely used to treat MDR-TB cases, improving the success rates as well as to shorten the drug regimen period [74]. Contrary to other genes related to drug resistance, mutations in all over the *pncA* gene had been reported to cause resistance to pyrazinamide [75].

1.4.6 Multidrug-resistant and extensively drug-resistant TB

With the first-line treatment regime the majority of TB cases are cured. However, drug resistant strains are globally reported more frequently [58]. Recently, the WHO recommends an standardized drug regimen consisting of the administration of all-oral drugs for 6-9 months. These therapy included bedaquiline, fluoroquinolones, ethionamide, clofazimine, in combination with effective first-line drugs [76]. In contrast, MDR patients with fluoroquinolones resistance have to take a longer (up to 20 months) MDR-TB treatment that involve a combination of WHO endorsed second-line drugs. In addition to this, longer MDR-TB treatment is related with adverse side effects to the patient, and it is more expensive compared with the first-line treatment [77, 78, 79]. Importantly, during 30 years, and during the development of this thesis, the standard treatment for MDR-TB was the administration of fluoroquinolones and one injectable aminoglycoside agent (kanamycin, capreomycin and amikacin) for 18 months. This regimen is still in use in Spain as there is no routine access to bedaquiline, a cornerstone of the all-oral regimen.

In addition to MDR-TB, extensively drug-resistant TB (XDR-TB) cases are reported worldwide. XDR-TB strains involve resistance to isoniazid, rifampicin, any fluoroquinolones, and at least one injectable agent. In 2018, the WHO reported that 6.2% of all MDR-TB cases were XDR-TB [10]. Just like MDR treatment, the WHO has recently endorsed a shorten drug regimen to treat these cases. This therapy consisted in using all-oral bedaquiline, pretomanid, and linezolid drugs for 6-9 months. In all cases, treatment success depends on the extent of drug resistance, the severity of the disease and the patient's immune system state. Drug resistance monitoring and patient follow-up are recommended during all the therapy time.

1.4.7 Diagnosis of drug resistant tuberculosis

Shortly after the use of anti TB drugs in the 40's, drug resistant strains emerged and were transmitted within the population [80]. As a consequence, in the 60's, George Canetti described a phenotypic-based method called proportion method for detecting resistant *M. tuberculosis* bacteria populations [81]. Since then, this TB drug susceptibility testing (DST) method has not changed, being the reference approach to identify drug resistance for decades, especially in low-income regions, where the TB incidence is high [82]. The proportion method uses different serial critical drug concentrations and compared the bacterial growth between susceptible and resistant strains. It uses an inoculum size prepared by shaking and adjusted to a specific opacity, then, the colony-forming-units are counted. Moreover, it can be performed in both liquid-based and agar-based growth media. The most common liquid-based proportion method is the automated BACTEC MGIT 960 system (Becton Dickinson, USA). Nevertheless, evidence pointed out that this assay introduces errors in susceptibility testing due to the bacterial inoculum size and, thus, the results should be used carefully [83]. In addition to this, it has been shown that some mutations related with rifampicin resistance (also known as disputed mutations) are not detected by automated liquid-based media growth [84]. Some of these disputed variants, particularly the *rpob* I491F, has been described to drive a MDR epidemic in Africa as they passed undetected by countries surveillance systems [85, 86]. For this reason, the European Committee on Antimicrobial Testing recommended a revision of the distribution of minimal inhibitory concentrations (MICs) for some anti TB drugs [82], in order to define the clinical breakpoints of each drug. For instance, it has been reported that susceptible and resistant strains have similar MICs for some first-line [87], and second-line drugs [88].

The steady increase of clinical MDR or XDR-TB cases, together with the monomorphic and clonal features of *M.tuberculosis* [89], offered the opportunity to identify the molecular basis of many resistances. Thus, specific point mutations (such as single nucleotide polymorphisms [SNP], deletions and insertions) were initially described by sequencing different drug target genes [73]. After identification of drug targets and associated mutations, molecular assays were implemented into routine clinical diagnostics. These assays are faster than phenotypic-based methods, being able to detect one particular resistance in a couple of hours or days. Moreover, they require less technical expertise, as well as less biosecurity laboratory facilities than phenotypic-based DST [82]. However, they tend to have lower sensitivity and/or specificity values than cultured-based systems [90]. Among the available commercial detection kits based on Nucleic Acids Amplification Tests (NAATs tests), the GeneXpert MTB/RIF assay (Cepheid, USA) as well GenoType MTBDRplus (Hain Lifescience, Germany) have been approved by the WHO since 2008 [91]. The Xpert MTB/RIF assay consists of a real-time PCR-based methodology for the detection of *M. tuberculosis* DNA as well as rifampicin related mutations. On the other hand, GenoType MTBDRplus is a line probe assay that detects MDR- and XDR-TB cases by screening specific mutations related with first-line (except pyrazinamide) and second-line drug resistance. Using the cultured-based methodologies as resistance reference, both molecular assays give sensitivity and specificity values of $>$98%, demonstrating that they are reliable diagnostic tests for TB patients as well as for MDR-TB individuals [92, 93]. The major difference is that Xpert MTB/RIF can be used on diagnostic samples (eg, sputum samples), while GenoType MTBDRplus is usually performed on cultured isolates. Recently, the Xpert Ultra assay (an updated version of the Xpert MTB/RIF) has been tested and demonstrated that improves the MTBC diagnostic accuracy and can be used as an initial test to diagnose pulmonary TB. In addition, the latest version of Truenat® MTB and MTB Plus system (Molbio Diagnostics, India) have also become an alternative MTBC diagnostic tool [94]. Despite massive deployment of Xpert assays (both versions), only a few low- and middle-income countries have access to these techniques. Instead, they use the sputum smear microscopy as primary diagnostic method for MTBC identification and phenotypic-based DST for drug resistance detection. Although sputum smear microscopy has lower sensitivity and specificity values compared to Xpert MTB/RIF [95], it is still cheaper and needs lower technical requirements.

Despite genotypic-based assays show a high agreement percentage compared with those phenotypic-based methods, there is still a significant percentage of resistant strains that are classified as "susceptible" by these genotypic-based approaches, especially those related with resistance to second-line drugs [96]. This is because the genotypic probes are limited in the number of mutations detected, only the most common variants conferring phenotypic resistance are tested. A solution to this is to increase the number of validated mutations by sequencing large collections of MTBC resistant and susceptible strains and compare to phenotypic results [97]. The application of new technologies like WGS could help to resolve this limitation; since it is possible to identify and annotate all the related mutations present in the genome [97] as well as to identify novel mutations related with drug resistance. A further discussion is detailed below and in **chapter 5** of this thesis.

1.5 Whole genome sequencing in infectious diseases and tuberculosis

The development and application of high-throughput technologies such as Next Generation Sequencing (NGS) in the study of infectious diseases have improved clinical detection methods and public health surveillance systems [98]. Most NGS applications offer faster, more comprehensive, and more accurate than traditional microbiological techniques. For instance, in the area of food-borne diseases, especially with *Listeria monocytogenes* bacteria, WGS detects more bacterial outbreaks in two years than those identified by classic pulse-field gel electrophoresis (known as PFGE technique), and thus, with epidemiological investigations, the outbreaks were solved, and eventually, the number of cases decreased [99]. In addition, NGS is proving to be a higher discriminatory molecular tool for identifying and studying pathogens outbreaks such as *Legionella* infections [100]. Regarding drug resistance detection, WGS-based studies have demonstrated its similar sensitivity and specificity values compared with routine phenotypic DST and/or genotypic probes references, resulting in a reliable and alternative technique for resistance prediction. Some bacterial examples include *Neisseria gonorrhoeae* [101], *Klebsiella pneumoniae* [102], *Staphylococcus aureus* [103], and *Pseudomonas aeruginosa* [104].

WGS also offers crucial insights into other infectious diseases, such as viral and fungal infections, malaria or neglected tropical diseases. WGS provides a better picture of the viruses diversity and its evolving patterns to gain resistance to antiviral agents [105]. Also, WGS improves estimations regarding high likely origin and date of certain outbreaks [106], and helped in the development of vaccines such as seasonal influenza [107]. Moreover, WGS has improved the knowledge about *Candida auris*, providing novel insights into treatment and epidemiology of the fungal pathogen [108]. Another NGS direct application is the rapid diagnosis of parasitic diseases. The Centers for Disease Control and Prevention (CDC) of the United States is working on the development of diagnostic tools for identifying drug resistance in malaria parasites [109], and an effective genotyping method for *Cyclospora cayetanensis* pathogen [110].

Although the NGS prices are cheaper (it is estimated that costs between 150-250US dollars per bacteria isolate [111]), only the United Kingdom [112] and the United States [113] have implemented this technique as a part of their routine diagnostic surveillance public health system. The use of NGS in routine clinical laboratories requires the initial investment in sequencing equipment as well as optimal infrastructure to work and store sequencing data. In addition, specialized personnel are needed, including bioinformaticians to create, handle and maintain the pipelines for analysis.

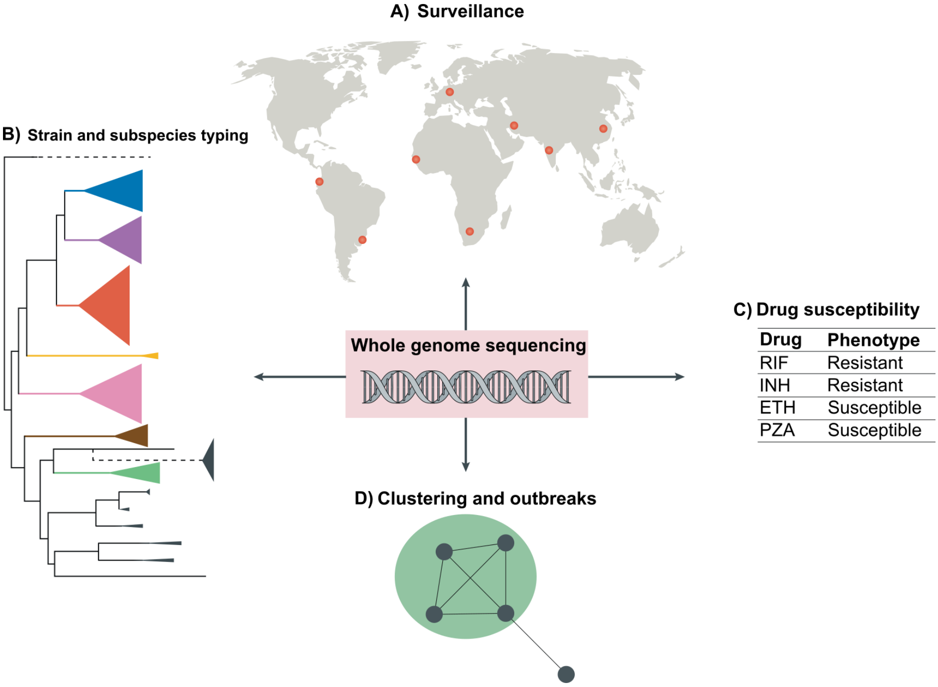
Recently, a single-molecule NGS portable instrument has been available. The MinION (Oxford Nanopore Technologies) has become an affordable and easy-to-use sequencing alternative within public health surveillance. This technology offers the capacity of sequencing in real-time at the point-of-care. For example, MinION helped public health surveillance during the Ebola [114] and Zika [115] outbreaks. Nevertheless, the current version of this technology has high sequencing error rates, especially in G/C genomic regions [116], which is common in some pathogens such as *M. tuberculosis*.

1.6 Applications of whole genome sequencing in tuberculosis

In 1998 the first *M. tuberculosis* complete genome was available. The reference H37Rv strain was sequenced by a combination of large-insert clones (cosmids and BACs) and small-insert clones approaches (shotgun sequencing libraries). Thus, the complete genome consisted of a single-chromosome sequence of 4,441,529 bp (with a G/C content of 65.6%), encoding around 4,000 genes [117]. Since then, we have improved our understanding regarding genetic determinants of drug resistance, adaptation to host immunological responses and the evolution of the pathogen. In addition, it helped in the development of molecular typing methods with epidemiological implications. Despite all the advantages, it took around two years to obtain a single genome.

Nowadays, WGS is becoming an essential tool in the TB field, not only in basic research areas but also in diagnostics and public health. Currently, we have the ability to whole-genome sequence from dozens to thousand MTBC strains at the same time depending on the instrument used. In fact, MTCB has become the most whole-genome sequenced pathogen bacteria. the main WGS applications in TB focused on disease control by: 1) improving drug susceptibility prediction; 2) rapid detection of transmission clusters; 3) strain genotype surveillance across country borders and 4) diagnosis of MTBC strains and lineage identification [56] **Figure 1.4**.

MTBC is genetically monomorphic with very little diversity among strains even when comparing animal- and human-adapted lineages [118]. In fact, a maximum genetic distance of 2,200 SNPs (which corresponds to 0.05% of the genome) between strains of different lineages has been detected using Illumina-NGS technology in a global collection of MTBC strains [119]. This low diversity is also contributed by the lack of significant ongoing recombination or horizontal gene transfer events, major contributors to diversity in other bacterial pathogens. WGS analyses are mainly based on the detection of specific SNPs and/or small genomic deletions or insertions (INDELS) using customized bioinformatics pipelines [56]. These pipelines consisted of three main steps. Briefly, raw sequences (those obtained from sequencing device) are trimmed according to a quality control value followed by mapping to a reference genome, and finally the detection of SNPs and INDELS. An additional step is to exclude genetic elements that cause mapping errors (approximately 10% of the MTBC genome) such as large gene families, as well as some mobile genetic elements [120]. Despite a well defined step-by-step protocol, there does not exist a gold standard WGS MTBC pipeline. A more detailed pipeline as well as all the softwares used will be described in the next chapters.



Drug	Phenotype
RIF	Resistant
INH	Resistant
ETH	Susceptible
PZA	Susceptible

Figure 1.4: The primary applications for whole genome sequencing of *M.tuberculosis* in public health include international surveillance of prevalence and drug resistance (panel A), determination of the species or subspecies of M. tuberculosis complex isolates (panel B), determination of drug resistance patterns on the basis of the presence of specific SNPs (panel C) and identification of transmission clusters and outbreaks (panel D). ETH, ethambutol; INH, isoniazid; PZA, pyrazinamide; RIF, rifampicin. Taken and adapted from [56]

1.6.1 *Mycobacterium tuberculosis* diversity

TB cases are caused by the members of the MTBC. They are a group of highly related bacilli sharing the 99.9% of nucleotide content of the whole genome [121]. The MTBC includes the two well known human-adapted mycobacterial groups; *M. tuberculosis* sensu stricto [lineages (L) 1 to 4 and L7], and the lineages traditionally referred to as *M. africanum* (L5 and L6). Moreover, at least nine MTBC phylogenetic groups mainly infecting wild and domestic mammalian hosts are part of MTBC (including *M.bovis*, *M. caprae*, *M. microti*, *M. pinnipedii*, *M. orygis*, *M. suricattae*, and *M.mungi*) [2]. Recently, a new lineage (defined as L8) has been described as part of MTBC involving strains causing human tuberculosis [122]. The seven phylogenetic lineages of human-adapted MTBC are geographically spread [123] (**Figure 1.5**). Overall, L2 and L4 are the most globally distributed. L2 is the most common in East Asia, whilst isolates belonging to L4 are frequently found all over the globe although is particularly frequent in Western Europe, The Americas and Africa. This distribution is in agreement with the hypothesis that L4 originated in Europe and then was carried to America during the colonization period in the XVI century [124]. On the other hand, L1 dominates the Indian Ocean region and parts of East Africa and L3 is restringed to Central and South Asia. Contrarily, L5 and L6 are restricted to West Africa regions, L7 can be only found in Ethiopia and L8 has only been found in two cases from is limited to the African Great Lakes region. This fact suggests that lineages could be adapted to specific human populations [125]. In fact, sub-lineages somewhat reproduce the geographic restrictions observed at the lineage level. This has led to the hypothesis that some lineages and sub-lineages are “specialists” human-adapted genotypes with a narrow niche to a specific human population, while globally distributed lineages are considered “generalists” genotypes infecting a wider range of human populations [40]. Data shows that the differential success of genotypes is likely due to a combination of historical contingency and pathogen biology [126].

There is substantial evidence supporting that the most likely geographic origin of MTBC was in Africa. First, *M. canetti*, the mycobacterial species closest to the common ancestor of the MTBC, is restricted to the Horn of Africa. In addition, Africa has the largest diversity of MTBC lineages. More recently, a phylogeographical analysis using 259 whole genome sequences from a global MTBC isolates collection, postulated Africa as the most likely origin region [127]. The result has been recently corroborated by the recent discovery of L8 in the Great Lakes branching before the diversification of the seven known lineages. Contrary to place of origin, the time of the most common ancestor of the complex (tMRCA) is still controversial; one of work dating the ancestor with bayesian analysis estimated that the most common ancestor existed around 70,000 years ago [127]. On the contrary, another recent research inferred a high likely origin around 6,000 years ago by including ancient MTBC DNA from 1,000-years peruvian mummies [128].

In addition to ecological adapted features, some MTBC lineages show genetic differences that have an impact in clinical as well as epidemiological features, resulting in a more virulent MTBC phenotype. This virulent phenotype is related with disease severity and its transmission rate. A study performed in Tanzania demonstrated that patients infected with L4 strains induced more level of acute-phase reactants, which are proteins involved in the inflammatory response, than patients infected with L1 strains, suggesting an increased virulence among L4 isolates [129]. Similar to this, Jong et al[130] showed that individuals harbouring MTBC infections with L2 and L4 had shorter latency periods compared with those infected with L6 strains. The most strong evidence a clinical role of MTBC lineages is the repeated association of L2 to drug resistance. This has been observed in many parts of the world with supportive evidence from in-vitro experiments [131]. Regarding transmissibility features, there is an overall view that strains from L2 and L4 are more transmissible than other MTCB lineages, mainly because there has been an increase in frequency of these genotypes over the time [132]. However, this evidence is based on clustering rates obtained from culture positive cases and culture bias due to MTBC variability can affect the growth and metabolism of some lineages. For instance, It has been described that lineages 5 and 6 grow slowly *in vitro* due to a specific mutation in *pykA* gene [133], suggesting that the prevalence of these MTBC lineages could be underestimated in culture [134]. Furthermore, while L2 strains have expanded in South Africa [135] there is no such expansion observed in other regions like Southern Europe where L2 strains are constantly imported from Asia and Eastern Europe. Thus, the success of MTBC genotypes is highly dependent on the socioeconomic context of the country, the presence of other successful genotypes and the human genetic backgrounds present in each specific region [136].

1.6.2 Drug susceptibility prediction

WGS drug susceptibility prediction is based on the presence or absence of specific mutations (SNPs and INDELS) related with drug resistance. In this sense, full catalogues of well curated mutations are required. In the last few years, high-confidence DR mutations have been described using genotype-phenotype statistical associations obtained by a sequencing a large number of isolates from multiple clinical datasets [97]. For example, the Comprehensive Resistance Prediction for Tuberculosis (CRyPTIC) project aims to replace the phenotypic testing by predicting the drug susceptibility profile given specific DR mutations, especially those related with first-line treatment [137]. Another international consortium is the Relational Sequencing Tuberculosis Data Platform (ReSeqTB), whose main objective is to expand our current catalogue of high-confidence DR list by using a public TB database, where researches can contribute with their WGS data (including genotypic, phenotypic and clinical outcome data) [138].

With these international initiatives, currently we are able to predict first-line DR strains profiles in the absence of phenotypic data [139, 137]. A WGS-based meta-analysis reported that mean sensitivity and specificity values for drug resistance detection were 0.98 (95% CI 0.93–0.98) and 0.98 (95% CI 0.98–1.00) for rifampicin and 0.97 (95% CI 0.94–0.99) and 0.93 (95% CI 0.91–0.96) for isoniazid, respectively. On the contrary, the remaining first-line

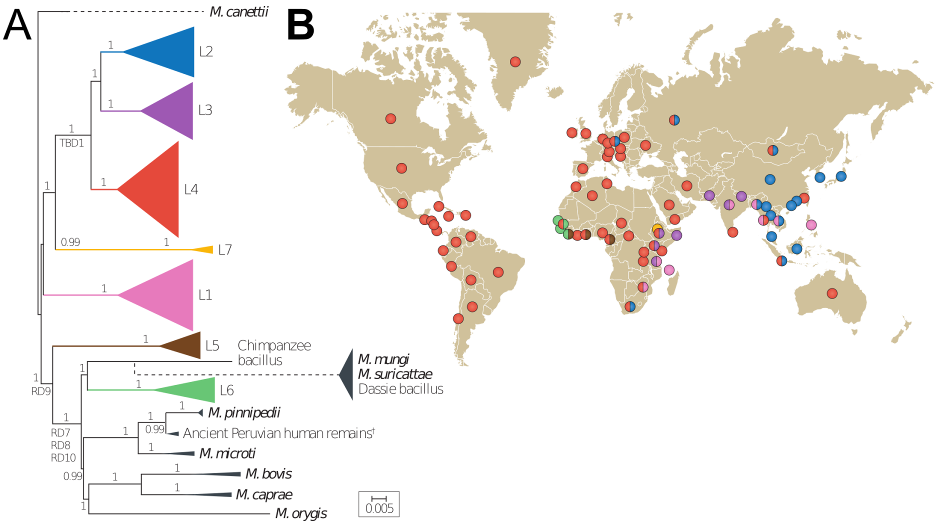


Figure 1.5: Global phylogeography of the human-adapted MTBC. A. Genome-based phylogeny of the *Mycobacterium tuberculosis* complex (MTBC). The MTBC comprises seven human-adapted lineages (in colour) and several lineages adapted to various wild and domestic animals (in grey). Branches of the main lineages are collapsed to improve clarity (indicated by triangles). M. tuberculosis-specific deletion 1 (TBD1) indicates that all lineage 2 (L2), L3 and L4 strains share this genomic deletion. Similarly, the deletion of the region of difference 7 (RD7), RD8, RD9 and RD10 is indicated under the respective branches. The grey dotted line leading to *Mycobacterium mungi*, *Mycobacterium suricattae* and the dassie bacillus indicates the most likely phylogenetic relationship of these animal-adapted ecotypes with the other members of the MTBC. The dagger indicates genomes generated from 1,000-year-old MTBC DNA that was recovered from archaeological human remains in Peru. Bootstrap confidence intervals are indicated. Scale bar represents the number of nucleotide substitutions per site. B. The global distribution of the seven main human-adapted MTBC lineages. Taken and adapted from [1]

drugs (ethambutol, pyrazinamide and streptomycin) reported more varied performance values. Thus, sensitivity values ranged from 0.71-1.00, 0.43-1.00, and 0.57-1.00 for ethambutol, pyrazinamide and streptomycin, respectively. While specificity values ranged from 0.15-95.8, 0.67-1.00, and 0.4-1.00, respectively [140]. These results showed that WGS can be considered an accurate alternative for drug susceptibility prediction and as such it has now replaced routine microbiological culture in the United Kingdom and The

Netherlands.

Unfortunately, a small percentage of first-line DR isolates harbouring uncommon mutations will be classified as susceptible, giving a false negative result and, as a consequence, compromising standard treatment. For instance, disputed mutations have been described to confer low-level rifampicin resistance (eg, *rpoB*, I491F) and, thus, some automated phenotypic liquid-based instruments may not detect it, resulting in an erroneous resistance classification [66, 84, 86]. A similar situation happened with XDR-TB cases, in which most of DR variants related with some second-line and/or new antibiotics such as delamanid are still unknown [141]. This is because available XDR phenotypic tests are not well standardized, resulting in discrepancies. Development of alternative or standardization of current phenotypic methods is needed in order to validate these variants as well as to increase the number of high-confidence DR mutations global list, especially those related with second-line drugs. A further discussion is detailed in **chapter 5** of this thesis.

1.6.3 Whole genome sequencing as an epidemiological marker

In the last two decades, TB molecular methods to classify MTBC strains and detect transmission have evolved from complicated techniques such as the restriction-fragment-length-polymorphism (RFLPs) to more reproducible methods like Spoligotyping and the Mycobacterial Interspersed Repetitive Units-Variable Number of Tandem Repeats (MIRU-VNTR) typing. These latter methods are based on the amplification of MTBC repetitive sequences, and the presence/absence of the different regions to determine the distinct genotypes. In addition, global online databases have been created to classify strains according to their respective Spoligotyping and MIRU-VNT profile [136]. Despite these protocols are well standardized and have been widely used worldwide, it is not clear their added value to epidemiological interventions compared to contact tracing which targets close contacts and is not dependent on culture [142].

More recently, WGS has been demonstrating a higher discriminatory power to trace and disentangle transmission networks [143, 144]. The first reported study was in 2005, in which the authors used WGS to resolve three identical isolates by MIRU-VNTR approach [145]. Since then, several studies have implemented the use of WGS as a supportive marker to help epidemiological investigations [146, 147, 148]. However, very few are population-based and mostly based on retrospective collections (**Additional Tables 10.1-10.2**), despite the fact that they are extremely helpful to understand transmission dynamics in a population as well to track the impact of TB control interventions [27, 28, 147, 149].

Using epidemiological links information and WGS data, it has been proposed that a maximum genetic distance of 12 SNPs between two different MTBC isolates indicated recent transmission [150], and a threshold of 5 SNPs for very recent events [151, 152]. The estimated time of infection is based on the low mutation rate of the MTBC (0.04-2.2 SNPs per genome, per year [153]). Although these SNPs cut-offs can resolve TB outbreaks, they were calibrated in low-burden TB settings. Whether these genetic distances can be applied in other TB settings is still unknown and will be discussed in **chapter 3** of this dissertation.

However delimiting transmission clusters is not the only application of WGS in transmission studies [144]. Given the higher resolution it can be used to infer individual links of transmission, in other words, identify who infected whom within a transmission cluster. Many modelling-based algorithms, as well as bayesian-based inferences, have been developed [154, 155], however, these were created focused in other pathogens with higher mutation rate [156], and its application in MTBC epidemiology has been scared and usually limited to large outbreak rather than to entire population. This topic will be discussed further in **chapter 4** of this thesis.

1.6.4 Other whole genome sequencing applications

Despite decreasing costs, WGS is still an expensive technique difficult to implement in low- and middle-income countries [56]. However, thanks to WGS of representative MTBC isolates, we can extract SNP markers that can characterize strains without additional whole genome sequencing. For example, Coll et al. [157] used a global collection of MTBC strain genomes to define a set of SNPs that can help to type major MTBC lineages and sub-lineages. MTBC strain classification is important to describe the bacteria diversity and to understand its population structure at local and global scales. Also, there is increasing evidence that there are biological differences between strains that may play a role in disease outcome between the MTBC lineages [158]. These variations include clinical outcomes and drug resistance acquisition [119]. For example, isolates belonging to L2 have a higher basal mutation rate and more likely develop drug resistance than L4 strains, at least in *in vitro* experiments [159]. In an epidemiological context, strain classification is important because we can track specific MTBC lineages or sub-lineages outbreaks that are spreading in local settings as well as between countries [148, 160]. Due to the very low rate of homoplasy, SNPs are the most robust markers for phylogenetic and epidemiological purposes [136]. Thus, different PCR-based molecular approaches to identify phylogenetic SNPs simultaneously in one reaction, have been developed as an alternative and affordable tool for MTBC genotyping [161, 162].

Although the SNP-typing technique “per-se” does not have the required resolution for defining and resolving transmission clusters, some SNP-based alternatives have been developed for identifying TB outbreaks in specific settings after applying WGS of representative isolates. Examples included the design and application of targeted SNP-based PCRs to track different MTBC clusters in a local settings [163, 164, 165]. More recently, a study has now been published to detect in almost “real-time” transcontinental spreading of a specific MTBC strain by SNP-typing [166]. Other SNP-typing applications include the detection of specific mutations related with drug resistance [167, 168]. Due to their multiple applications, technically easy and fast to perform, SNP-typing methodologies will continue to play an important role in TB research and control, especially in low- middle-income laboratories, where WGS is still far from being an essential tool for TB.

1.7 Implementation of tuberculosis whole genome sequencing into the health systems

Despite the great advance of using WGS in TB research, there are only few countries that have implemented it as part of routine diagnosis. There are a number of technical and economical issues (for example, market research) regarding the implementation of WGS as a routine practice. These issues involved the required laboratory and computing infrastructure [56]. In addition to this, standardized and easy-to-use pipelines need to be created by specialized bioinformaticians. That is why just high-income countries such as the United Kingdom (Public Health England) and The Netherlands (National Institute for Public Health and the Environment, RIVM) have implemented WGS as part of their routine TB diagnostic tools [113]. It would be desirable that following this tendency, other high-income countries will introduce WGS into their public health systems.

In low- and middle-income countries, this implementation seems farther. Short-term solutions include web-based WGS analysis pipelines. For example, PhyResSE [169] and TBprofiler [139] databases may help to cover this issue. Both pipelines are optimized for a rapid and complete MTBC-WGS analysis with a DR prediction as well as phylogenetic classification. However, the main limitation to this is the need of the sequencing files as an input, and importantly, the poor internet connection that some places have. International supportive and political commitments will be necessary for a sustainable implementation of WGS in TB diagnostic pipelines.

Objectives and outline

2.1 Aim the thesis

Although the use of WGS in the TB field is increasingly common, its use as an epidemiological and diagnostic markers is still scarce. Even in high-income regions like Spain, there is very little integration in public health systems. Valencia Region (Comunidad Valenciana) is a low-burden TB incidence area in which the current TB diagnostics and epidemiological interventions are enough to maintain a low TB incidence rate through time, but with a slow pace of incidence decline. It is also a region where most TB cases are contributed by local-born individuals, in contrasting differences with other low-burden countries, such as the United Kingdom or The Netherlands. Furthermore, there is limited data about the amount of ongoing transmission of the disease in different settings, and the very few published studies are mostly based on molecular markers with low resolution. We hypothesized that applying WGS will improve our understanding and knowledge regarding the TB disease clinical and epidemiological characteristics in the region. Additionally, we reason that lessons and/or methods learned in this thesis can then be extrapolated to other TB settings.

In this dissertation, we used WGS to genomically characterize a large proportion of MTBC clinical isolates collected during three years (2014-2016) in Valencia Region. First, we performed an epidemiological study, estimated the genomic transmission rate and identified risk factors associated (**chapters 3 and 4**). Second, we also evaluated the use of WGS to predict drug resistance in the studied population (**chapter 3**) and to identify novel resistance determinants and assist on personalizing the treatment of a challenging TB patient (**chapter 5**). Finally, we studied the global genomic diversity of the bacteria to propose a new, efficient and rapid methodology for strain genotyping (**chapter 6**). Our results were compared with those of the local health system. To our knowledge, this is the first regional and likely national project of this kind. We hope that this population-based study may serve as a precursor to using WGS as a routine tool for TB in the public health surveillance system and could also be extrapolated to the national level as well as to low- and middle-income countries.

Finally, this thesis has been possible thanks to the invaluable contribution of all the health personnel of the hospitals that participated during its realization. In total, 18 health entities from Valencian Region were involved (**Additional Table 10.3**)

2.2 Objectives

The main objectives of the thesis are focused on the use WGS applied to tuberculosis surveillance, thus, the specific objectives were:

- To characterize by WGS the MTBC clinical isolates collected during the study period. Specifically, to estimate and predict drug resistance and transmission burden based on sequencing data (**chapter 3**).

- To identify clinical and epidemiological features associated with genomic transmission (**chapter 3**).

- To compare the TB transmission detected by WGS against those detected by routine contact tracing investigations performed by local health system (**chapter 3**).

- To evaluate transmission dynamics within genomic clusters by inferring high likely transmitters (index cases included) as well as identifying risk factors associated with them (**chapter 4**).

- To apply WGS data to personalize TB treatment, especially in those TB patients with uncertain DR phenotypic profile (**chapter 5**).

- To develop two rapid and affordable PCR-based techniques to characterize MTBC isolates from specific phylogenetic SNPs derived from WGS data (**chapter 6**).

- To perform and validate our SNP typing assays in two different TB burden settings (**chapter 6)**.

2.3 Outline

This thesis is composed of 4 main chapters, 3 of them have already been published in high-impact scientific journals (chapters 4 to 6). For editing purposes, the final accepted version of them is included in this dissertation. At the beginning of each chapter, there is a link to each manuscript journal page.

In chapter 3, we used the WGS data obtained from 785 MTBC clinical isolates to describe the TB population. More specifically, we evaluated risk factors associated with TB and disease transmission. We also classified and predicted DR profiles from all the TB cases for which a WGS of the corresponding isolate was available. In addition, we detected the genomic transmission and compared it with the local surveillance system. Finally, we estimated WGS sensitivity and specificity values using the routine TB diagnostic methods as a reference.

In chapter 4, we combined mathematical modelling with WGS data to infer whether a most likely index case is sampled or not within a genomic transmission cluster. Moreover, we estimated when these transmitters infected other individuals. In other words, when high likely transmission events occured.

Once transmitters were identified, we searched for risk factors specifically associated with them.

In chapter 5, we used WGS data to identify and report a misidentified MDR-TB case within a TB patient with a presumed “fully susceptible” infection for 9 years. We first identified that uncommon and novel DR mutations were responsible for the MDR status. Also, those DR variants were not detected by routine clinical methods, explaining why the strain infecting the patient was identified as susceptible. In this chapter, we highlight the importance of WGS DR prediction to provide appropriate drug treatment, this method is more accurate and also faster than culture based DR methods.

Finally, in chapter 6, we developed two cheap and fast PCR-based SNP-typing assays to classify clinical isolates into the main phylogenetic lineages of MTBC as well as lineage 4 sublinages. After validation, we applied our molecular approaches within a clinical MTBC collection of 491 samples, demonstrating high sensitivity and specificity values.

Using whole genome sequencing to determine and identify tuberculosis transmission: a population-based study in Valencia Region, Spain.

3.1 Introduction

Controlling tuberculosis (TB) transmission is the key to reducing the number of cases and to eradicate the TB worldwide. Recent transmission is a major contributor to global TB burden, notably in high-burden incidence settings [151]. On the contrary, in some low-burden areas, most TB prevalence is related with infections derived by reactivations of migrants from high-burden countries [170, 149]. Although TB incidence within local-born patients in these settings is low, recent transmission rate among them is elevated. For instance, a population-based study in the United Kingdom (UK) evidenced that 67% of TB cases were due to foreign-born people from high-burden countries, however, recent transmission within UK borders mostly involved local-born individuals [27].

Tuberculosis incidence is contributed by both recently infected TB cases (usually defined as less than 5 years after exposure), and long-term latent TB infections (LTBI) [170]. It is estimated that the prevalence of LTBI in close contacts ranges between 24.2-32.4% in low-burden settings [46]. Current TB control efforts are mainly based on a passive case finding strategy, which focuses on a rapid detection of active TB individuals once they report to a primary healthcare facility, followed by close contacts investigations [43]. By contrast, active case finding intervention aims the disease eradication by targeting people in high-risk groups to find LTBI (eg, crowded settings) and, thus, potentially reduced transmission [171]. Passive case finding followed by contact tracing also focuses on sputum positive cases as the assumptions is that those are the most contagious [172]. However, in several studies it has been observed that between 10%-25% of index cases were sputum negative [31], suggesting that an important fraction of transmission cases are missed by our current strategy of passive case finding and contact tracing. A similar conclusion has been reached when applying whole genome sequencing (WGS), as very likely transmission cases by WGS are not identified by epidemiological investigations [149]. Thus, there is an emergent view that active case finding can play an important role to halt tuberculosis transmission. There are different strategies to implement active case finding [173]. Community-wide active case finding has been tested in Vietnam and has shown a significant decrease of TB incidence as compared to passive case finding [174]. However, given the prevalence of LTBI most of these strategies are targeted, focusing on specific communities or high-risk groups.

Measuring TB transmission is not straightforward. Approaches to estimate it range from epidemiological investigations (eg, contact tracing method) to molecular tools to detect TB outbreaks (eg, MIRU-VNTR genotyping technique) [41]. These latter approaches have the potential to reveal genetic links that are not identified by contact tracing. However, traditional molecular genotyping has its limitations. For instance, there is reported that many genotyping links identified by molecular tools have not epidemiological evidence associated, resulting in an overestimation of TB transmission rates [55], resulting in no cost-effective added value to the public health surveillance systems [142]. As a result, the adoption of this technique as a universal genotyping method is scarce, especially in low-income countries.

This situation has started to change with the use of whole genomes as epidemiological markers [175, 150]. Whole genomes show a higher degree of agreement with epidemiological investigations , and avoid overclustering detected by conventional methods. The most common approach to delineate this genomic transmission is by using different single nucleotide polymorphisms (SNP) thresholds [150]. Although the number of SNPs is still under debate [56], there is no doubt that a 5 SNP threshold identifies cases of recent transmission while some additional cases can be identified with higher cut-offs at the expense of including false clustered cases [143, 144]. At the same time, using whole genome sequencing (WGS), we can model the evolution of the pathogen sequences and, hence, define how likely two cases are in the same transmission cluster even without epidemiological data support [147, 176]. However, the majority of publications are focused on transmission previously detected by conventional genotyping methods while population-based studies are scarce [144].

WGS is also useful to predict drug resistance (DR). Recently, it has been proved that DR prevalence at country level can be accurately estimated by genomic surveillance [177], and could replace the classic phenotypic drug susceptibility testing (DST), especially DR predictions related with first-line drugs [178, 137, 179] . The sensitivity and specificity values of this WGS-DST prediction largely depends on the use of well-curated catalogues of DR mutations robustly linked to individuals drug susceptibility profiles. Thus, international consortiums have been publishing catalogues of well-established mutations [169, 97]. Additionally, our increased knowledge of these causative genetic variants demonstrates that first-line drug susceptibility profiles can be precisely predicted and used at patient level [137].

Here, we reported a TB population-based study in Valencia Region, a low-incidence setting from 2014 to 2016. We carried out a WGS-based analysis of 785 patients corresponding to 77% of all MTBC culture positive cases. During the study period the mean TB incidence was 8.4 cases per 100,000 population, including local- and foreign-born inhabitants. Contrary to the tenet that most cases in high-income regions are due to foreign-born individuals, most cases were Spanish-born patients (63%) and were more likely involved in recent transmission events. At the population level, we identified that between 35-41.3% of the TB patients were genomically related within a transmission cluster (47.4% of them were local-born patients). Our results show that transmission is still a major contributor to TB burden even in low-burden settings like Valencia Region. Compared to the United Kingdom where most cases are imported and genomic clustering rate is low, our results suggests that halting transmission among the local-born population can accelerate TB elimination in the region.

3.2 Methods

3.2.1 Study Population Setting

Valencia Region, including Alicante; Castellón and Valencia provinces, is the fourth largest populated region in Spain with 4,963,703 million inhabitants, of which 15% corresponds to foreign-born people. During the study period there were reported a total of 1281 TB cases, giving an incidence rate of 8.4 per 100,000 population (https://www.sp.san.gva.es/).

Tuberculosis local surveillance is managed by the Regional Public Health Agency (DGSP: Dirección General de Salud Pública); they use contact tracing method as a gold standard approach to detect TB transmission within the population. As a part of a mandatory routine, each positive diagnosed TB patient fills up a local-standardized questionnaire in order to identify epidemiologically related cases and highly likely new infected individuals. In addition, this survey also collected clinical, microbiological and demographic data. Contact investigation was done in 72% of all TB notified individuals through 2014 and 2016. A total of 9,312 close contacts were investigated by

DGSP and found that 23.7% of these individuals were infected contacts, which included active TB and LTBI cases. Of these, just 13.3% of patients developed the disease. Standard phenotypic drug susceptibility testing (DST) using liquid-based (BACTEC-MGIT 960 system), and/or solid-based (Löwenstein-Jensen) growth media is used for routine drug resistance (DR) inspection.

Approval for the study was given by the Ethics Committee for Clinical Research from the Valencia Regional Public Health Agency (Comité Ético de Investigación Clínica de la Dirección General de Salud Pública y Centro Superior de Investigación en Salud Pública). Informed consents were waived on the basis that tuberculosis is part of the regional compulsory surveillance program of communicable diseases. All personal information was anonymized and no data allowing individual identification was retained.

3.2.2 Isolate Collection and DNA extraction

A total of 785 single-patient cultures were retrieved from 18 Hospitals from Valencia Region over the 2014-2016 period. This number corresponded to 77% (785/1019) of all culture-positive cases reported by the DGSP. Hospitals previously tested TB positivity on liquid Mycobacteria Growth Indicator Tube (MGIT) or solid-media Löwenstein-Jensen (LJ). In order to obtain sufficient DNA amount all clinical isolates were cultured again in Middlebrook 7H11 agar plates supplemented with 10% OADC (Becton-Dickinson) for 3 weeks at 37℃. DNA was extracted with CTAB method from a representative population sample (four times plate scraping). A TB biological library was constructed by storing all scrapped samples in 1ml of glycerol (20%) at -80C°. All procedures were performed at FISABIOS's BSL3 facility (Valencia city), under WHO protocols recommendations.

3.2.3 Whole Genome Sequencing and bioinformatics analysis

Sequencing libraries were constructed with Nextera XT DNA Library Preparation Kit (Illumina, San Diego, CA) following the manufacturer's instructions. WGS procedure was performed on the Illumina MiSeq platform.

Data analysis was carried out following a validated pipeline previously described [180] (http://tgu.ibv.csic.es/?page_id=1794). Briefly, high quality sequencing reads not belonging to the *Mycobacterium tuberculosis* Complex (MTBC) were filtered out. Then, resulting reads were mapped to an inferred MTBC most likely common ancestor genome (https://zenodo.org/record/3497110). SNPs with a minimum 10X depth coverage and a quality score of 20 were kept for further analysis. We separated SNPs according to their frequency. Fixed-SNPs (fSNP), those called with at least 90% of allele frequency, were used for epidemiological analysis (e.g transmission detection); while variable-SNPs (vSNP), those detected with frequency between 10-89%, were also included for DR prediction. Indels were considered when the mutation was present in a minimum of 10 reads and 10% of frequency. We used the H37Rv as reference (NCBI, AL123456.2) for variant annotation. Variants detected in regions difficult to map such as repetitive sequences and PPE/PE-PGRS genes were removed from the analysis as well as those detected within higher density regions ($>$ 3 SNPs in 10 bp).

3.2.4 Drug resistance prediction and MTBC classification based on WGS data

Drug resistance profile for every sample was predicted by identification of well-known mutations conferring resistance. Briefly, all mutations (fSNPs and vSNPs included) were compared with ReSeqTB [181] and PhyResSE [169] resistance databases, well-validated catalogues of variants associated with first and second line drugs resistance. Isolates were classified as susceptible, resistant, multidrug resistant (MDR) or extensively drug-resistant (XDR), according to their predicted resistance profile. MTBC isolates classification was determined using a combination of lineage (L) and sub-lineage specific markers previously described [157, 40]. Additionally, we searched for phylogenetic allele's frequencies in vSNPs, if we found two different lineage or sub-lineage SNPs in one sample or those variants frequencies were less than 95%, the isolate was considered a likely mixed infection. Despite Valencia region being considered a low-burden TB setting, five samples were identified as mixed infection and removed from the analysis.

3.2.5 Genomic transmission based on genetic distances and phylogeny

Genomic-based estimation of transmission was evaluated by performing clustering analysis based on pairwise distance of whole genome data. A multisequence alignment (MSA) file with the fSNPs from all samples was constructed discarding drug resistance position. Then, pairwise distances between each sample was computed with R ape package. A genomic cluster was defined whether a group of at least two MTBC strains shared a defined genetic distance, we evaluated genomic clusters using three different thresholds (0, 5 and 12 SNPs). Clusters were defined using a customized script. To confirm that all clusters were monophyletic, a maximum likelihood tree was constructed with the MSA (50,184 SNPs) file using RAxML v8 [182], GTRGAMMA model and 1,000 bootstrap replicates. *Mycobacterium canetti* was included as an outgroup.

3.2.6 Statistical analysis

In order to evaluate association between risk factors and genomic clustering (clustered cases vs unique cases), we computed Fisher's exact test in samples with epidemiological data available (n=724, 92.2% of all sequenced individuals). Univariate analysis identified that Spanish-born was related to genomic transmission (see **Results**). Next, we carried out univariate comparisons on Spanish-born individuals that were in genomic transmission (clustered cases, n=299) against all MTBC lineages and sub-lineages.

We made a comparison between the transmission rate detected by WGS with that identified by local Public Health surveillance system to correlate SNP distances with epidemiological evidence. We applied three different SNPs thresholds (0,5 and 12 SNPs). In addition, we estimated the number of transmission events detected by routine contact tracing method as well as WGS. The number of potential transmission events was calculated by the number of individuals within a cluster minus one case, which could be an index case (n-1 method). All statistical analyses and graphics were performed in R.

3.2.7 Performance of WGS for drug resistance prediction and epidemiology

We evaluate the performance of the WGS technique for DR prediction as well as for detecting transmission. Sensitivity, specificity, accuracy, positive predictive values (PPV), and negative predictive values (NPV) were calculated for both comparisons [183]. Sensitivity was calculated as the probability that a given result will be a positive result (true positive value). Specificity was calculated as the probability that a given result will be a negative result (true negative value). PPV was calculated as the probability that a given positive result was a true positive result. NPV was calculated as the probability that a given negative result was a true negative result.

In the case of DR prediction, we compared our genomic DR profile against the DST results obtained by the hospitals (n=684). Routine phenotypic DST for first-line drugs was carried out using the BACTEC MGIT 960 system (Becton Dickinson, USA). DST was considered the reference method. Comparisons were made for the first-line drugs (rifampicin, isoniazid, ethambutol, pyrazinamide and streptomycin). In the case of transmission, we tested the genomic clustered cases detected (considering 0, 5 and 12 SNPs threshold), against the reference method, which was the epidemiologically linked cases identified by the standard contact tracing investigations.

3.3 Results

3.3.1 Study population

Between 2014 and 2016, 1281 new cases of tuberculosis were diagnosed and notified to the DGSP, of which 1019 (79.5%) were culture-positive. We performed WGS on 785 individuals, which corresponds to 77% of all culture-positive TB patients (785/1019); epidemiological information was available for 724 individuals, which constitute a representative subset from the local TB population. **Table** 3.1 presents an overview of the demographic and clinical characteristics of culture-positive TB cases. Further analysis was performed with this dataset.

The majority of the individuals were Spanish-born (63%, n=456), men (62.3%, n=451) and with a median age of 44 years (SD±19.27). Most of the cases (82%,n=594) were related to pulmonary tuberculosis and only 57.7% (n=418) were sputum smear positive. A small proportion of patients displayed HIV positive status (7%, n=51) or alcohol abuse (19%, n=138). The mean TB incidence rate in Spanish-born people was 6.7 per 100,000 inhabitants compared with 20 per 100,000 in foreign population.

Table 3.1: Demographic characteristics of all culture-positive TB individuals analyzed in the study. All patients column is showing the values obtained by the local surveillance system, while WGS patients define the TB cases that we were able to sequence. Abbreviations; WGS, Whole Genome Sequencing.

Characteristic	**All Patients (n= 1019)**	**WGS Patients (n=724)**
Age (years)		
≤15	23 (2.2%)	20 (2.8%)
16-24	70 (6.8%)	57 (7.9%)
25-44	413 (40.5%)	287 (39.6%)
45-65	338 (33.1%)	238 (32.9%)
≥66	175 (17.1%)	122 (16.8%)
Sex		
Female	381 (37.4%)	273 (37.7%)
Male	638 (62.6%)	451 (62.3%)
Place of birth		
Spanish-born	682 (67%)	456 (63%)
Foreign-born	337 (33%)	268 (37%)
Sputum smear		
Positive	575 (56.4%)	418 (57.7%)
Negative	439 (43.1%)	302 (41.7%)
Disease type		
Pulmonary	832 (81.6%)	594 (82%)
Extrapulmonary	187 (18.3%)	130 (17.9%)
Alcoholism[a]		
Yes	181 (17.7%)	138 (19%)
No	777 (76.2%)	543 (75%)
Diabetes[b]		
Yes	110 (10.7%)	74 (10.2%)
No	887 (87%)	639 (88.2%)
HIV infected[c]		

continued

Characteristic	**All Patients (n= 1019)**	**WGS Patients (n=724)**
Yes	61 (5.9%)	51 (7%)
No	862 (84.6%)	609 (84.1%)
Social exclusion[d]		
Yes	99 (9.7%)	90 (12.4%)
No	882 (86.5%)	606 (83.7%)
Health care workers[e]		
Yes	24 (2.3%)	14 (1.9%)
No	986 (96.7%)	705 (97.4%)
Imprisonment[f]		
Yes	45 (4.4%)	39 (5.3%)
No	952 (93.4%)	668 (92.3%)
Diagnostic delay[g]		
Yes	264 (25.9%)	180 (24.8%)
No	712 (69.9%)	520 (71.8%)
Contact tracing investigation[h]	139 (13.6%)	97 (13.4%)

[a] Unknown data in 61 individuals.
[b] Unknown data in 22 individuals.
[c] Unknown data in 96 individuals.
[d] Unknown data in 38 individuals.
[e] Unknown data in 9 individuals.
[f] Unknown data in 22 individuals.
[g] Diagnostic delay was considered whether there were at least 90 days between symptom onset and diagnosis date. Unknown data in 43 individuals.
[h] TB individuals found by routine epidemiological contact tracing method used by local surveillance system.

3.3.2 MTBC genotyping

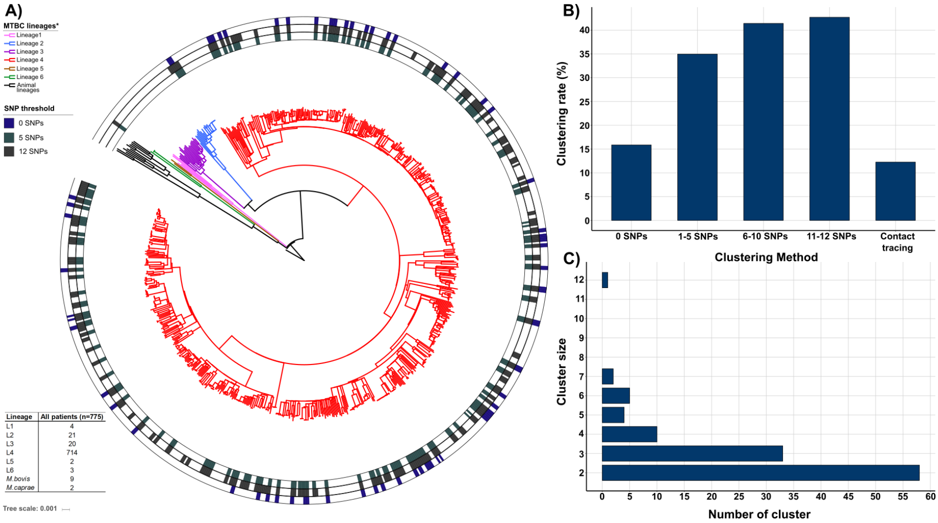
Of the 785 MTCB clinical samples analyzed, 10 isolates were removed due to either non-MTBC samples (5/785) and likely mixed infections (5/785). Using specific SNPs to determine MTBC lineages and sub-lineages circulating in the region (Coll et al. 2014; Stucki et al. 2016), we identified six different lineages, including animal genotypes. The most frequent was L4 in 714 isolates (92.1%), followed by L2 and L3, with 21 (2.7%) and 20 (2.6%) strains, respectively. Lineages 1, 6 and 5, were the least frequent with 4, 3 and 2 cases, respectively. In contrast, we detected 11 cases within animal lineages, 9 of these belonged to *M. bovis*, and 2 were assigned to *M. caprae* (**Figure 3.1A**).

Because L4 was the most frequent in our study, we further inspected its main sub-lineages. We identified that 89.9% (n=644) of all L4 isolates corresponded to the three main L4 generalists genotypes described by (Stucki et al. 2016); 33.3% L4.1.2 (Haarlem family, n=240); 30.2% L4.3 (LAM family, n=216); and 26.3% L4.10 (PGG3 family, n=187). In contrast, specialist sub-lineages were found in a lower proportion; 1.4% L4.6.2 (Cameroon family, n=10); L4.1.3 and L4.5 with two cases each. In addition, isolates belonging to L4.4, L4.1.1 ("X" genotype) and L4.2 were detected in 3.2% (n=23), 2.6% (n=19), and 1.4% (n=10), respectively. Moreover, we couldn't identify any of the known specific sub-lineage SNPs in five L4 strains, thus they remain as L4.

3.3.3 TB transmission rate in Valencia Region is higher than evidenced by contact tracing

Transmission clustering rate was estimated using all sequenced data (n=775). Using a genomic clustering approach to evaluate transmission (12 SNPs), we detected that 331 (42,7%) samples were within a transmission group (**Figure 3.1B**). They were included in 112 different clusters with variable size (ranging from 2-12 cases per cluster (**Figure 3.1C**). In order to distinguish transmission due to recent or old infection, we evaluated different SNP cut-offs. Using a 5 SNP threshold, the clustering rate was 35%, which corresponded to

271 isolates within 97 genomic clusters. A 5 SNP threshold should roughly correspond to an infection event 5-8 years ago (assuming a molecular clock of 0.3-0.5 SNPs/year). Furthermore, the clustering rate by applying a 0 SNP cut-off was 15.9%, which involved 123 individuals within 51 groups and is indicative of very recent transmission. In contrast, the percentage of linked cases revealed by contact tracing was considerably lower, 12.5% (n=97), than the proportion estimated by WGS clustering method (**Figure 3.1B**).



Lineage	All patients (n=775)
L1	4
L2	21
L3	20
L4	714
L5	2
L6	3
M.bovis	9
M.caprae	2

Figure 3.1: Genomic characterization of the study region.A) Phylogeny of 775 TB isolates collected during 2014-2016. Each ring represents different genomic clusters detected by different SNP threshold (0, 5, 10 and 12 SNPs). textitM.canneti was used as an outgroup. B) Clustering percentage obtained by using different SNP threshold. C) Number of genomic clusters by different cluster size. 12 SNP threshold was used as a standard. Clusters sizes from 8 to 11 samples were not detected. * Nomenclature proposed by Comas et al. [127]

By contact tracing, 97 out of the 775 samples with WGS data, had been identified to be epidemiologically linked (12.5%, (**Figure 3.1B**), involving 66 transmission clusters. From these 97 samples, we observed that only 74 (76.3%, 74/97) and 65 cases (67%, 65/97) were genomically related by applying 12 and 5 SNPs threshold, respectively. The link between the rest of the cases (23/97) was incompatible with the genetic distance threshold used to estimate transmission since they are separated by more than 100 SNPs which roughly corresponds to a transmission event around 100 years ago (**Figure** 3.2).

Using the standard 12 SNP threshold, and the available epidemiological data from patients (n=724), we calculated 199 transmission events which involve 288 TB cases belonging to 111 genomic clusters. Of these genomic transmission events, only 15,5% (31/199) were epidemiologically detected by the local surveillance system." It is notably that all transmission events identified by epidemiological investigations happened between individuals that had a genetic distance of 5 SNPs, suggesting that contact tracing method only identifies very recent transmission.

We then benchmarked WGS as an epidemiological tool for transmission by using 12, 5, and 0 SNPs thresholds against the epidemiologically linked cases detected by contact tracing. For this comparison we used the samples with contact epidemiological data associated (n=724) and declared as a cluster by public health officials. The number of confirmed linked cases by contact tracing in our dataset was 59 and we used them as a gold-standard to identify the SNP thresholds that maximizes sensitivity and specificity. When using different genetic thresholds to define transmission the sensitivity values were 93.22% (95% CI, 83.54-98.12%) for 12 SNP cutoff, and 91.53% (95% CI, 81.32-97.19%) for 5 SNP threshold and decreased to 49.15% (95% CI, 35-89-62.50%) when using a 0 SNP cutoff. Specificity values were lower with 63.31% (95% CI, 59.52-66.98%), 71.73% (68.14-75.12%), and 88.27% (95% CI, 85.58-87.60%) using 12, 5, and 0 SNP thresholds, respectively. In addition, PPV was 30% in all transmission cutoffs (**Table** 3.2). In general, the accuracy of WGS were 65.75% (95% CI, 62.16-69.20%) using 12 SNPs, 73.34% (95% CI, 69.96-76.53%) using 5 SNPs, and 85.08% (95% CI, 82.28-87.60%) by 0 SNPs.

SNP threshold	Number of transmission clusters	Number of clustered cases	Sensitivity (95% CI)	Specificity (95% CI)	Accuracy (95% CI)	PPV (95% CI)	NPV (95% CI)
12 SNPs	11	288	93.22% (83.54-98.12)	63.31% (59.52-66.98)	65.75% (62.16-69.20)	18.39% (16.64-20.28)	99.06% (97.61-99.63)
5 SNPs	96	242	91.53% (81.32-97.19)	71.73% (68.14-75.12)	73.34% (69.96-76.53)	22.31% (19.92-24.91)	98.96% (97.63-99.55)
0 SNPs	50	107	49.15% (35.89-62.50)	88.27% (85.58-90.62)	85.08% (82.28-87.60)	27.10% (21.04-34.15)	95.14% (93.83-96.18)

Table 3.2: Performance of WGS method for transmission. These values were extracted from 724 clinical TB isolates and assuming epidemiological links from contact tracing as gold-standard. Abbreviations: SNP, Single nucleotide polymorphism; PPV, Positive predictive value; NPV, Negative predictive value.

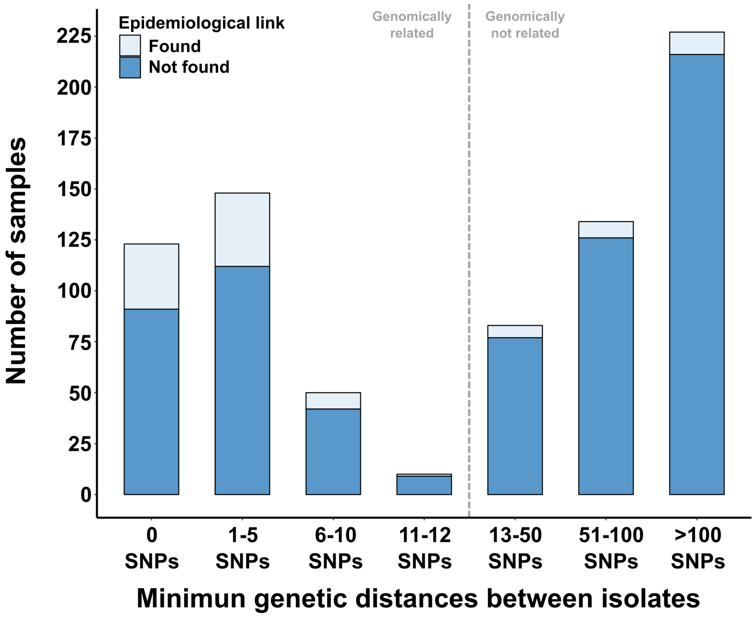


Figure 3.2: Clustered samples using different clustering methods. The graph shows the numbers of samples with their minimum genetic distance between isolates. The grey dashed line separates the genomically related samples (clustered cases) from those that are not (unique cases.

3.3.4 TB transmission is associated with Spanish-born population

Clustering rate resulted considerably high for a low-burden region (42.71%, using a 12 SNP threshold), hence, we next evaluated whether there were any risk factors associated with transmission that could explain the elevated value observed. As expected, transmission was associated with the youngest patients (median age, 4 years) (85% vs 15%, p = 0.001), on the contrary, patients with 65 years were associated with being unique cases (29.5% vs 70.5%, p = 0.001). Additionally, pulmonary TB was associated with clustered cases (45% vs 75.4%, p = 0.001). No risk factors such as alcohol abuse and HIV positive status were associated with transmission (48.5% vs 51.5%, p= 0.123; and 41.2% vs 58.8%, p = 1, respectively). Strikingly, we found that the Spanish-born people were more likely associated with transmission when compared to foreigners (47.4% vs 31%, p= 0.001) (**Table 3.3**). Knowing this, we investigated whether there was any particular risk factor within Spanish-born patients associated with transmission. We obtained the same results as for the total population (**Table 3.3**).

Table 3.3: Risk factors associated with transmission among the whole population and Spanish-born population. Abbreviations; REF, Reference; WGS, Whole Genome Sequencing.

	All Patients (n=724)				Spanish Population (n=456)			
Variable	**Clustered cases (n=299)**	**Unique cases (n=425)**	**Odd Ratio (95% CI)**	**p-value**	**Clustered cases (n=216)**	**Unique cases (n=240)**	**Odd Ratio (95%CI)**	**p-value**
Transmission method								
WGS	299 (41.3%)	425 (58.7%)	4.54 (3.48-5.96)	<0.001	216 (47.4%)	240 (52.6%)	5.04 (3.64-7.02)	<0.001
Contact tracing	97 (13.4%)	627 (86.6.%)			69 (15.1%)	387 (84.9%)		
Age (years)								
≤15	17 (85%)	3 (15%)	6.53 (1.83 - 35.51)	<0.001	16 (88.9%)	2 (11.1%)	5.45 (1.21-50.67)	0.018
16-24	20 (35.1%)	37 (64.9%)	0.63 (0.33 - 1.17)	0.144	12 (57.1%)	9 (42.9%)	0.97 (0.33-2.64)	1
25-44	133 (46.3%)	154 (53.7%)	**REF**	**REF**	83 (59.3%)	57 (40.7%)	**REF**	**REF**
45-64	93 (39%)	145 (61%)	0.74 (0.52 - 1.07)	0.111	74 (45.4%)	89 (54.6%)	0.57 (0.35-0.92)	0.020
≥65	36 (29.5%)	86 (70.5%)	0.49 (0.30 - 0.78)	<0.001	31 (27.2%)	83 (72.8%)	0.26 (0.14-0.45)	<0.001
Sex								
Male	197 (43.6%)	254 (56.4%)	1.3 (0.94 - 1.79)	0.102	141 (49.8%)	142 (50.2%)	1.07 (0.72-1.59)	0.775
Female	102 (37.3%)	171 (62.7%)	**REF**	**REF**	75 (43.3%)	98 (56.7%)	**REF**	**REF**
Place of birth								
Spanish-born	216 (47.4%)	240 (52.6%)	2 (1.44 - 2.79)	<0.001	-	-	-	-
Foreign-born	83 (31%)	185 (69%)	**REF**	**REF**	-	-	-	-
Sputum smear								
Positive	184 (44%)	234 (56%)	1.32 (0.96-1.80)	0.078	129 (50%)	129 (50%)	1.29 (0.87-1.91)	0.185

continued

	All Patients (n=724)				Spanish Population (n=456)			
Variable	**Clustered cases (n=299)**	**Unique cases (n=425)**	**Odd Ratio (95% CI)**	**p-value**	**Clustered cases (n=216)**	**Unique cases (n=240)**	**Odd Ratio (95%CI)**	**p-value**
Negative	113 (37.4%)	189 (62.6%)	**REF**	**REF**	86 (43.6%)	111 (56.4%)	**REF**	**REF**
Disease type								
Pulmonary	267 (45%)	327 (55%)	2.5 (1.60-3.98)	<0.001	196 (51%)	188 (49%)	2.70 (1.52-4.97)	<0.001
Other	32 (24.6%)	98 (75.4%)	**REF**	**REF**	20 (27.8%)	52 (72.2%)	**REF**	**REF**
Alcoholism								
Yes	67 (48.5%)	71 (51.5%)	1.35 (0.91 - 2.00)	0.123	47 (54%)	40 (46%)	1.32 (0.80-1.72)	0.281
No	223 (42.9%)	320 (58.9%)	**REF**	**REF**	165 (47.1%)	185 (52.9%)	**REF**	**REF**
Diabetes								
Yes	31 (41.8%)	43 (58.1%)	1.02 (0.61 - 1.71)	1	25 (41.6%)	35 (58.4%)	0.76 (0.42-1.37)	0.404
No	264 (41.3%)	375 (58.7%)	**REF**	**REF**	188 (48.3%)	201 (51.7%)	**REF**	**REF**
HIV infected								
Yes	21 (41.2%)	30 (58.8%)	0.97 (0.52 - 1.80)	1	15 (48.4%)	16 (51.6%)	1.02 (0.48-2.28)	1
No	255 (41.9%)	354 (58.1%)	**REF**	**REF**	186 (47.8%)	203 (52.2%)	**REF**	**REF**
Social exclusion								
Yes	34 (37.7%)	56 (62.3%)	0.85 (0.52-1.36)	0.493	16 (44.4%)	20 (55.6%)	0.89 (0.42-1.87)	0.862
No	253 (41.7%)	353 (58.3%)	**REF**	**REF**	194 (47.3%)	216 (52.7%)	**REF**	**REF**
Health care workers								
Yes	9 (64.3%)	5 (35.7%)	2.60 (0.77-9.99)	0.100	9 (64.3%)	5 (35.7%)	2.03 (0.60-7.85)	0.277
No	288 (40.9%)	417 (59.1%)	**REF**	**REF**	206 (46.9%)	233 (53.1%)	**REF**	**REF**
Imprisonment								
Yes	20 (51.3%)	19 (48.7%)	1.53 (0.76-3.09)	0.241	16 (59.3%)	11 (40.7%)	1.64 (0.70-4.02)	0.237
No	272 (40.7%)	396 (59.3%)	**REF**	**REF**	198 (46.9%)	224 (53.1%)	**REF**	**REF**
Diagnostic delay								

continued

	All Patients (n=724)				Spanish Population (n=456)			
Variable	Clustered cases (n=299)	Unique cases (n=425)	Odd Ratio (95% CI)	p-value	Clustered cases (n=216)	Unique cases (n=240)	Odd Ratio (95%CI)	p-value
Yes	66 (36.6%)	114 (63.4%)	0.78 (0.54-1.12)	0.161	46 (39%)	72 (61%)	0.64 (0.41-1.01)	0.050
No	222 (42.7%)	298 (57.3%)	**REF**	**REF**	162 (50.5%)	159 (49.5%)	**REF**	**REF**

[a] Unknown data in 43 individuals.

[b] Unknown data in 11 individuals.

[c] Unknown data in 64 individuals.

[d] Unknown data in 28 individuals.

[e] Unknown data in 5 individuals.

[f] Unknown data in 17 individuals.

[g] Diagnostic delay was considered whether there were at least 90 days between symptom onset and diagnosis date. Unknown data in 24 individuals.

As no risk factor is associated with local-born increased transmission we wondered if transmission was driven by the bacterial genotype. We analyzed all the genotypes (lineages and sub-lineages included) that had at least 10 cases infecting Spanish-born and performed a multivariate analysis. We identified no association between transmission among Spanish-born and a specific MTBC genetic lineage, except for sub-lineage L4.1.1, for which the proportion of clustered and unique cases in Spanish-born people is very different (14.3% vs 85.7%, p-value 0.011). In this case, this statistically significance result is associated to be a unique case rather than being a clustered case. These results suggest that there is no specific MTBC lineage involved in transmission among local population studied, but this does not preclude that more specific genotypes are, further analysis identifying specific genotypes must be conducted (**Table 3.4**).

Lineage	**Patients (n=456)**	**Clustered cases (n=216)**	**Unique cases (n=240)**	**Odds ratio (95% CI)**	**p-value**
L4.1.1	14	2 (14.3%)	12 (85.7%)	0.17 (0.02-0.81)	0.011
L4.1.2	141	70 (49.6%)	71 (50.4%)	**reference**	**reference**
L4.3	145	68 (46.9%)	77 (53.1%)	0.90 (0.55-1.46)	0.722
L4.4	11	5 (45.5%)	6 (54.5%)	0.85 (0.19-3.50)	1
L4.10	126	62 (49.2%)	64 (50.8%)	0.98 (0.59-1.63)	1

Table 3.4: MTBC lineage distribution among Spanish-born population separated by clustered and unique cases. Only genotypes that had at least 10 Spanish-born individuals are represented.

Finally, we also inspected the distribution of genetic distances between Spanish-born beyond the SNP thresholds we used for recent transmission. In that way, we can identify older transmission events. We noticed a continuous distribution of genomic distances between strains without a clear cut-off at 12 SNPs (**Figure 3.3**), contrary to what was observed in other low-burden TB settings, such as the United Kingdom [27] where transmission are very recent and no case is found between 12-50 SNPs (see **General Discussion**). This results suggests that transmission in the Valencia Region has been maintained over the last decades, contrary to what is observed in the UK where only very recent transmission events are detected.

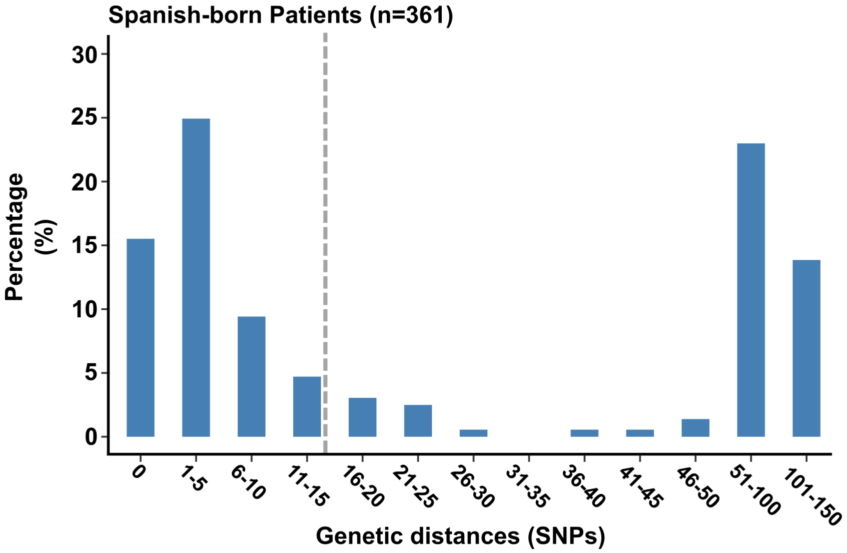


Figure 3.3: Distribution of genetic distances between Spanish-born isolates. The grey dashed line separates the genomically related samples (clustered cases) from those that are not (unique cases).

3.3.5 Using WGS as a routine tool to predict drug resistance

WGS allowed the prediction to first- and second-line drugs. Using our catalogue of confident mutations we predicted the DR profile in all the 775 MTBC sequenced isolates. The majority of these were predicted as susceptible to all TB drugs (691/775, 89.2%). Nevertheless, 85 samples harboured at least one mutation related with drug resistance (10.9%). Fifteen isolates were predicted as MDR-TB and one as an XDR-TB case. Based on genetic data, isoniazid resistance was the most frequent resistance found with 57.8% (n=44) of the cases, followed by rifampicin and fluoroquinolones resistance with 25% (n=19) and 23.6% (n=18), respectively. Pyrazinamide resistance was predicted in 19.7% (n=15) individuals. Finally, ethambutol and streptomycin resistance phenotypes were predicted in 17.1% (n=13) each. All the *M. bovis* strains harboured the phylogenetic mutation H57D in *pncA* that confers resistance to pyrazinamide.

In order to test whether WGS can be used as a tool for first-line drug resistance prediction in our setting, we calculated their performance values. We used culture-based DST as the reference method as it is the routine method for determining first-line DR in the Valencia region. Out of the 775 samples studied, 702 had a complete DST profile available, 79 of them (11.2%, 79/702) were classified as DR resistant to at least one first-line antibiotic. WGS specificity values for all first-line drugs ranged between 98-100%, being pyrazinamide the lowest with 98.81%. In contrast, sensitivity and PPV values were low, especially for some drugs such as ethambutol, pyrazinamide and streptomycin, whose values were around 50% or lower (50%, 52.38% and 39.13%. respectively). Overall, the accuracy values were $>$97% for all the drugs (**Table 3.5**), suggesting that WGS could be used as a diagnostic tool for DR prediction in the region. The discrepancies between WGS and DST might be explained by the presence of MTBC strains harbouring uncommon mutations not present in available catalogues (see **chapter 5** for an example), labelling errors in the genomics laboratory or false negative/positive DST results.

Drug	Resistant isolates by DST	Resistant isolates by WGS	Sensitivity (95% CI)	Specificity (95% CI)	PPV (95% CI)	NPV (95% CI)	Accuracy (95% CI)
Rifampicin	15	14	73.33% (44.90-92.21)	99.56% (98.72-99.91)	78.57% (53.24-92.19)	99.42% (98.66-99.75)	99.00% (97.95-99.60)
Isoniazid	40	39	85.00% (70.16-94.29)	99.39% (98.45-99.83)	89.47% (76.03-95.79)	99.09% (98.11-99.56)	98.57% (97.38-99.31)
Ethambutol	12	10	50.00% (21.09-78.91)	99.41% (98.50-99.84)	60.00% (32.66-82.27)	99.12% (98.46-99.50)	98.56% (97.36-99.31)
Pyrazinamide	21	20	52.38% (29.78-74.29)	98.81% (97.66-99.48)	57.89% (38.18-75.38)	98.51% (97.69-99.04)	97.40% (95.91-98.45)
Streptomycin	23	12	39.13% (19.71-61.46)	99.55% (98.69-99.91)	75.00% (46.50-91.19)	97.94% (97.16-98.51)	97.54% (96.09-98.56)

Table 3.5: Performance of WGS as a diagnostic tool for drug resistance. These values were extracted from 724 clinical TB isolates. DST was used as a reference method. Abbreviations: DST, Drug susceptibility testing; PPV, Positive predictive value; NPV, Negative predictive value; WGS, Whole genome sequencing.

3.4 Discussion

This is the first national population study-based using WGS to genomic characterize the MTBC population circulating in Valencia region and more importantly, to delineate its transmission and compared with local epidemiological investigations. We whole-genome sequenced 77% of all MTBC culture-positive cases during three years, which is a representative proportion of all the TB notified cases of the region, and give us a real picture of the bacteria population that are infecting the population as well as some epidemiological factors that are associated with disease prevalence, especially local-born feature.

Valencia Region is a TB low-burden area with an incidence of 8.4 (per 100,000 inhabitants), however, we found multiple evidence that suggests transmission is still playing an important role in the setting. First, the proportion of Spanish-born population with TB was 63%, this percentage is in agreement with the reported in other Spanish cities such as Barcelona [131] and Madrid [184] with 65.3%, and 63.4%, respectively. On the contrary, the proportion of local-born TB cases in other low-burden countries such as Canada [185], the UK [27], and The Netherlands [149] are lower than 30% of all TB individuals. Second, we detected that almost half of all TB cases were clustered (42.7%), using the standard of 12 SNP threshold to delineate genomic transmission, and one third (35%) when applying a 5 SNP threshold (35%). In both cases, this clustering rate is higher compared to those reported in other low-burden countries, where genomic transmission ranged between 14-16% [27, 149], and somewhat closer to that reported in high-burden TB countries, where the detected transmission ranged from 39 to 66% [151, 186]. However, the sampling period of these studies ranged from one to ten years, and thus the molecular clustering rate may not be directly comparable (but see **General Discussion**). Finally, we identified that this elevated transmission rate was associated with local-born TB cases as opposed to recent immigrants. A similar situation has been reported in the UK, where foreign-born TB incidence is contributed by LTBI acquired abroad and, hence, they are less associated with local transmission clusters [27]. Although the TB local-born incidence rate in Valencia Region was low (6.7 per 100,000 inhabitants), we did not identify a clear SNP cut-off to delineate genomic transmission (from 0 to 35 SNPs, **Figure 3**), demonstrating a continuum distribution of genetic distances in our setting involving very recent but also older transmission events that had occurred among local-born individuals. By contrast, while in the UK transmission is also higher among local-born only only very recent transmission is observed (see **General Discussion**).

Despite the availability of a few additional population-based studies that use WGS as a primary tool for detecting transmission, our results are not fully comparable with them, due to the specifics of the TB setting. For example, there are two studies that genomically characterized endemic regions within Canada [28, 176]. In both reports, the genomic clustering percentage is higher than 70%. However, those studies focused on specific and low-migration populations and, hence, the TB incidence rate is low but the majority of the cases are genomically related.

The fact that epidemiological investigations only detected 12.5% clustered cases against 42.71% identified by WGS denotes the limitations of the contact tracing method. A recent meta-analysis has demonstrated that WGS helped to improve the current contact tracing method revealing additional members of transmission clusters [144]. In high-burden countries, epidemiological links were identified only in the 18% of all clustered cases [147]. In contrast, agreement percentage raises up to 42.3% in low-burden countries [27]. The disagreement observed in Valencia region might be explained by the fact that we observed more transmission than other low-burden countries, and this transmission likely happened during sporadic contacts outside the boundaries of contact tracing investigations. As a consequence, contact tracing likely missed a number of epidemiological links, similar to what happens in high burden countries. This also explains the low sensitivity, specificity and PPV obtained when comparing contact tracing with WGS (**Table 3.2**). We noticed that the higher agreement values were obtained with lower SNP thresholds (0 and 5 SNPs) but also that many links below those thresholds are not detected by contact tracing, supporting the idea that routine contact tracing method in Valencia Region only detects a percentage of very recent transmission.

In general, Spain is a low-burden resistant TB country (incidence of 0.5 per 100,000 population). By traditional phenotypic DST, 11.2% of cases were identified to have a drug resistant profile. Using WGS, we predicted that 10.9% (n=85) of all the isolates were at least resistant to one antibiotic and 1.9% (n=15) were predicted as MDR-TB. Using DST as a reference method for detecting phenotypic resistance, we calculated performance values for WGS. Overall, the specificity, NPV and accuracy values were higher than 97% (**Table 3.5**), which is comparable with other worldwide studies [137], and other low-burden TB countries such as The Netherlands [179]. Nevertheless, the fact that some first-line drugs (eg, rifampicin) had a lower sensitivity values than other studies, could suggest the presence of uncommon low-level resistance mutations that are missed by routine phenotypic DST [84, 187]. This phenomenon is further discussed in **chapter 5**.

In conclusion, the use of WGS for the genomic characterization of MTBC isolates, prediction of drug resistances profiles, and most importantly, to detect genomic transmission, provides a perfect complement to the information obtained by routine methods used in this local setting. Our results show that transmission is still a major contributor to local TB prevalence, suggesting that strategies for cutting this transmission are required in order to accelerate TB elimination. In this sense, due to WGS is able to detect genetic links that are missed by contact tracing, the combination of both methods will improve the conventional epidemiological investigations, for example, by designing specific interventions that target high transmission TB risk groups or foci (eg, geographical hotspots of TB). Our data suggests that WGS can add additional actionable information at the patient and the population level and progress must be made to incorporate WGS routinely in the Valencia Region health system. In fact, as a consequence of this thesis, we have generated 10 reports to public health authorities and to microbiological units of the hospitals. Our reports have helped to confirm or to oriented epidemiological investigations. At the individual level they have helped to confirm or discard drug resistance and in some cases to help to treat the patient as in **chapter 5**.

High-resolution mapping of tuberculosis transmission: Whole genome sequencing and phylogenetic modelling of a cohort from Valencia Region, Spain.

The work described in the present chapter has been published as a Research Article in *PLoS Medicine* Journal:

Xu Y, Cancino-Muñoz I, Torres-Puente M, Villamayor LM, Borrás R, Borrás-Máñez M, et al. High-resolution mapping of tuberculosis transmission: Whole genome sequencing and phylogenetic modelling of a cohort from Valencia Region, Spain. Murray MB, editor. PLoS Med. 2019;16: e1002961. https://doi.org/10.1371/journal.pmed.1002961

4.1 Abstract

Background: Whole genome sequencing provides better delineation of transmission clusters in *Mycobacterium tuberculosis* than traditional methods. However, its ability to reveal individual transmission links within clusters is limited. Here, we used a 2-step approach based on Bayesian transmission reconstruction to (1) identify likely index and missing cases, (2) determine risk factors associated with transmitters, and (3) estimate when transmission happened.

Methods and findings: We developed our transmission reconstruction method using genomic and epidemiological data from a population-based study from Valencia Region, Spain. Tuberculosis (TB) incidence during the study period was 8.4 cases per 100,000 people. While the study is ongoing, the sampling frame for this work includes notified TB cases between 1 January 2014 and 31 December 2016. We identified a total of 21 transmission clusters that fulfilled the criteria for analysis. These contained a total of 117 individuals diagnosed with active TB (109 with epidemiological data). Demographic characteristics of the study population were as follows: 80/109 (73%) individuals were Spanish-born, 76/109 (70%) individuals were men, and the mean age was 42.51 years (SD 18.46). We found that 66/109 (61%) TB patients were sputum positive at diagnosis, and 10/109 (9%) were HIV positive. We used the data to reveal individual transmission links, and to identify index cases, missing cases, likely transmitters, and associated transmission risk factors. Our Bayesian inference approach suggests that at least 60% of index cases are likely misidentified by local public health. Our data also suggest that factors associated with likely transmitters are different to those of simply being in a transmission cluster, highlighting the importance of differentiating between these 2 phenomena. Our data suggest that type 2 diabetes mellitus is a risk factor associated with being a transmitter (odds ratio 0.19 [95% CI 0.02–1.10], p ¡ 0.003). Finally, we used the most likely timing for transmission events to study when TB transmission occurred; we identified that 5/14 (35.7%) cases likely transmitted TB well before symptom onset, and these were largely sputum negative at diagnosis. Limited within-cluster diversity does not allow us to extrapolate our findings to the whole TB population in Valencia Region.

Conclusions: In this study, we found that index cases are often misidentified, with downstream consequences for epidemiological investigations because likely transmitters can be missed. Our findings regarding inferred transmission timing suggest that TB transmission can occur before patient symptom onset, suggesting also that TB transmits during sub-clinical disease. This result has direct implications for diagnosing TB and reducing transmission. Overall, we show that a transition to individual-based genomic epidemiology will likely close some of the knowledge gaps in TB transmission and may redirect efforts towards cost-effective contact investigations for improved TB control.

4.2 Introduction

Better understanding of tuberculosis (TB) transmission is key for TB control in the 21st century. Economic resources are very limited in many high-burden countries, while in low-burden countries, TB control is jeopardized by diminishing resources, as TB is not perceived as a public health issue [188]. The limited funding is spent on tracing contacts of individuals diagnosed with TB; many of these contacts test negative for TB infection, whereas other contacts that had substantial exposure may not be screened. Historically a dichotomy between active and latent disease has been used at the epidemiological level to differentiate those TB cases that can transmit (active TB disease) versus those that do not (latent). However, more recent evidence suggests that the transition between these different states is fuzzy, and that TB development may be better represented as a spectrum of clinical and sub-clinical states [18]. The degree to which sub-clinical disease contributes to transmission is largely unknown, particularly because tools to detect sub-clinical disease have only recently become available [19, 15].

Whole genome sequencing (WGS) of patient isolates shows a higher agreement with contact investigations than previous markers [55]. Importantly WGS is also a superior tool to delineate transmission clusters and can be used to estimate the burden of transmission [189]. But only very limited approaches have been developed using WGS to identify individual transmission links. Phylodynamic and transmission network analyses based on the combined use of WGS and epidemiological data have been primarily confined to the analysis of large outbreaks [175, 54, 164, 190]. However, transmission clusters spanning decades are more an exception than a rule in TB epidemiology [191, 151]. For most epidemiological scenarios, 2 key limitations prevent the use of phylodynamic and network models to predict transmission links: the diversity of the bacteria is extremely low, and the time span does not allow a good correlation between time and the accumulation of variation. Population-based analyses where dozens or hundreds of transmission clusters can be identified typically involve cluster sizes of 1–15 TB cases and sampling times of 2–5 years. In high-burden countries, cluster sizes may be larger but time frames are still short. We thus developed an approach that allowed us to simultaneously analyze small clusters from a 3-year population-based study in the Valencia Region of Spain. Our approach infers index cases as well as estimating transmission times.

4.3 Methods

Our overall analysis proceeded as follows: isolate collection, sequencing analysis, identification of transmission clusters meeting certain criteria, phylogenetic tree reconstruction, calculation of tree timing with several choices of molecular clock rate, and, finally, Bayesian transmission analysis.

4.3.1 Case definitions

- **Clustered case**. A clustered case is a case that is genomically close to another case in the population according to a genetic threshold. Typically, for recent transmission, 12 or 5 SNPs are used but see below.

- **Index case**. The index case is the first documented individual in a TB outbreak, usually the one that generates an epidemiological investigation. In most epidemiological investigations in TB, this coincides with (or it is assumed to be) the first diagnosed individual.
- **Most likely ancestral genotype (MLAG)**. The MLAG is the reconstructed genotype of a hypothetical ancestral case of an outbreak. It may coincide or not with the index case from the epidemiological investigation. A match of the MLAG with any sampled genotype suggests that the sampled genotype is likely an index case.

4.3.2 Ethics statement

This study was approved by the Ethics Committee for Clinical Research of the Valencia Regional Public Health Agency (Comité Ético de Investigación Clínica de la Dirección General de Salud Pública y Centro Superior de Investigación en Salud Pública). Informed consent was waived on the basis that TB is part of the regional compulsory surveillance program of communicable diseases. All personal information was anonymized, and no data allowing individual identification was retained.

4.3.3 Study population and isolate collection

Valencia Region has 4,974,475 million inhabitants and is composed of 3 provinces, Castellón, Valencia, and Alicante. In 2018, there were 315 reported individuals with TB in the entire region (incidence rate of 6.4/100,000 inhabitants); Valencia is considered a low-TB-burden region. Contact tracing investigation is the gold standard procedure to detect transmission clusters and is done in 74.1% of all notified TB cases.

We performed a population-based genomic study involving 785 TB culture positive cases in Valencia Region, Spain, during 2014–2016 as a part of an ongoing local genomic epidemiology study. Using WGS data to delineate transmission (based on SNP distances, cutoff of $\leq$15 SNPs; see below), we identified 121 clusters, most of which involved 2 cases per cluster (n = 325 clustered cases; see **Supplementary Methods**). For the present analysis we included all transmission clusters that involved at least 4 TB cases and had more than 1 SNP (variant) between the strains. Based on a reviewer's feedback, we performed a chi-squared test to corroborate that the clusters selected for this study were a good representation of the total number of clustered cases in the population.

A total of 21 clusters met the criteria, involving a total 117 people with TB. For 115 of these we had epidemiological data including date of diagnosis and diagnostic symptom onset as well as other clinical and demographic data. For 2 individuals we used the date of culture positivity with a 2-week correction to infer the date of diagnosis.

4.3.4 WGS analysis and transmission delineation

DNA from TB culture positive Mycobacteria Growth Indicator Tubes (Becton Dickinson) was extracted. Sequencing libraries were constructed with Nextera XT DNA Library Prep Kit (Illumina) and sequenced on the Illumina MiSeq instrument. Generated paired-end sequencing reads were trimmed, and likely contaminant reads that might be present in clinical culture were filtered using KRAKEN software [192]. The bioinformatic analysis was performed following a previous pipeline[127]. Briefly, sequencing reads were mapped and aligned to an inferred *Mycobacterium tuberculosis* complex (MTBC) most likely common ancestor genome. Next, variants were separated into INDELS (small insertions and deletions) and SNPs. Variants with at least 10 reads in both strains and a quality score of 20 were selected. Because we wanted to detect genomic transmission, we focused on SNPs that were present with at least a 90% frequency. Finally, SNPs annotated in regions difficult to map such as repetitive sequences and PPE/PE-PGRS genes were removed from the analysis, as well as those detected in a window of 10 variants near INDELS. In addition, variants known to confer drug resistance [193, 97] were removed.

This pipeline has been validated by international public health TB reference laboratories (http://tgu.ibv.csic.es/?page_id=1794) and published [97, 194]. The parameters used in the pipeline are common among the genomic TB research community [56].

4.3.5 Cluster delineation based on SNP distances and phylogeny

Transmission clusters were defined using a loose cutoff of $\leq$15 SNPs. Furthermore, all detected groups were confirmed by building a phylogeny that included all the isolates. This phylogeny was inferred using the maximum likelihood phylogenetic approach with RAxML v8.2 [182], applying the General Time Reversible model of nucleotide substitution with the gamma distribution (GTRGAMMA). Transmission clusters with more than 1 SNP between the strains and composed of at least 4 TB individuals were kept for ensuing analyses. The methods described below are agnostic to the cutoff value, but with a threshold of 15 SNPs, we were sure to incorporate recent and old transmission events. In any case, most samples were below the cutoff of 12 SNPs, and 82% were below the cutoff of 5 variants.

4.3.6 Reconstruction of genetic relatedness networks

The resulting SNP alignment for each cluster was used to infer a genetic relatedness network. Due to the monomorphic and non-recombining nature of the MTBC [136] and the possibility that the ancestral genotype was present in the samples, we used a parsimony-based algorithm for network reconstruction implemented in the PopART software [195]. We chose a median joining network (MJN) approach because it allows cases to occupy central positions in the network; genotypes at branching points in the parsimony tree are hypothesized to have been present but unsampled. In addition, a reconstructed recent ancestor of the cluster based on the phylogenetic topology was added to the network so we could (1) hypothesize the MLAG and (2) infer the directionality of a SNP (wild-type versus mutant status) given the MLAG. In the genetic network analysis, we considered that any strain matching the MLAG for its transmission cluster was a candidate to be the index case of the cluster.

4.3.7 Timed tree reconstruction

The accepted value for the substitution rate in TB is approximately 0.3–0.5 substitutions per genome per year [54, 196], though our data seem to suggest that this rate may vary both between clusters and at the individual lineage level within clusters. We first estimated timed trees for all clusters using the treedater package in R [197] with 5 different clock rate values (ranging from 0.327 to 1.103) sampled from a log-normal distribution following a meta-analysis. Although we generated predictions for a range of rates, for clarity, results in the main text will be based on a clock rate of 0.363, which closely matches the mean rate identified in our meta-analysis and in a recent publication [153] for MTBC lineage 4, which dominates our population. Parameters used to obtain the different clock rate values, as well as the meta-analysis performed, are described and shown in **Supplementary Methods** and **Supplementary Table 4.3**.

4.3.8 Transmission inference

We developed a method of simultaneous transmission inference on many clusters based on TransPhylo, a Bayesian analysis approach that uses the Markov chain Monte Carlo (MCMC) method to reconstruct transmission trees from pathogen phylogeny [154]. The main difference between our method and TransPhylo's previous capabilities is that we can perform inference with multiple transmission clusters simultaneously, choosing which parameters should be shared between clusters.

The resulting transmission tree contains information about who infected whom and when, and also whether a case is sampled or not. This information is represented by a matrix whose columns are the times of infection, times of sampling, and transmitters, and whose rows correspond to individuals in the cluster. If an individual in the reconstructed tree is not sampled, then the corresponding entry for time of sampling is empty. TransPhylo produces a posterior sample of such trees. From this collection, we can extract (1) the posterior probability that the index case of a cluster is sampled and (2) the posterior probability that each host transmitted TB in their cluster. A detailed protocol that includes all equations of the TransPhylo method can be found in **Supplementary Methods**.

In order to test and validate our method, we performed simulations of 2 outbreaks. We observed narrower widths of credible intervals for all parameters (**Supplementary Results** and **Supplementary Figures 4.6-4.7**) using the simultaneous approach. This method has been incorporated into the latest version of the TransPhylo package [154].

4.3.9 Statistical analysis

We selected the index cases and the samples with higher than 0.6 posterior probability of being transmitters as predicted by TransPhylo (23 transmitters compared to the remaining 84 clustered cases), with sensitivity analysis of the latter threshold in **Supplementary Methods** and **Supplementary Table 4.4**. Then, we computed the odds ratio (OR) and 95% confidence intervals (Fisher's exact test) to explore epidemiological variables associated with being a transmitter. Furthermore, we performed a multivariate logistic regression to confirm our univariate result. Based on peer review feedback, we statistically compared epidemiological variables associated with transmitters to those of the non-clustered cases identified in the whole dataset.

4.4 Results

4.4.1 Genetic networks suggest missing index cases

Using an initial threshold of 15 SNPs, we identified a total of 21 transmission clusters involving 117 TB cases (**Table 4.1**). This 15-SNP threshold allowed us to look at older transmission events, although most of the cases (81.2%) were within 5 SNPs of another case, consistent with very recent transmission. Most of the clusters had more than 1 case with an identical genotype (0 SNP difference); 5 clusters had no identical pairs (**Supplemetary Table 4.3**). No statistical difference was observed for available clinical, epidemiological, and demographic variables between the 21 transmission clusters that met our inclusion criteria (n = 109) and the total clustered samples in the population (n = 325) (see **Supplementary Methods** and **Supplementary Table 4.4**).

Table 4.1: Main characteristics of the study population.

Characteristic	**All Patients** **(n= 109^a)**
Age (years)	
<18	11 (10%)
19–34	20 (18%)
35–65	66 (61%)
>65	12 (11%)
Sex	
Female	33 (30%)
Male	76 (70%)
Place of birth	
Spain	80 (73%)
Other country	29 (27%)
Sputum smear	
Positive	66 (61%)

continued

Characteristic	**All Patients** **(n= 109)**[a]
Negative	41 (38%)
Disease type	
Pulmonary	100 (92%)
Extrapulmonary	9 (8%)
Alcoholism	25 (23%)
Diabetes	13 (12%)
HIV infected	10 (9%)
Social exclusion	13 (12%)
Healthcare worker	5 (5%)
Imprisonment	8 (7%)
Diagnostic delay (days)	
≥30	46 (42%)
31–60	25 (23%)
61–89	14 (13%)
≤90	32 (29%)

[a] Eight TB cases had no epidemiological data.

Genetic networks are a popular approach to try to understand transmission without the need for additional epidemiological data. Using the SNP alignment data, we applied the MJN algorithm to establish genetic relatedness between the strains. A total of 22 missing links were predicted (involving 14 out of 21 genetic networks). In 5 of the genetic networks the predicted missing genotype corresponded to the MLAG, suggesting that the index case was not sampled. In other clusters intermediate genotypes were missing. In contrast, in 7 networks (33%) we did not predict any missing links, indicating that the MLAG predicted was present among the TB cases analyzed.

In the MJN approach it is reasonable to estimate that the strain with the same genotype as the MLAG is also the most likely index case. However, in several clusters (**Figure4.1, and Supplementary Figures 4.8-4.9**), more than one strain matched the MLAG, and thus the approach, which is based solely in genotypes, cannot predict which of the matching cases is the most likely index case. One striking feature of the networks in which we can identify an MLAG among sampled TB cases is that this hypothetical index case does not always coincide with the first diagnosed case (**Figure 4.1A**). This situation occurred in 2 of the 5 networks in which there was a case with the same genotype as the MLAG (clusters CL045 and CL078). Together with the fact that in an additional 14 genetic networks the MLAG was not present, this suggests that the common assumption that the earliest diagnosed case is the index case is not necessarily correct. All the networks reconstructed by the genetic network approach can be found in **Supplementary Figures 4.8-4.9**.

Genetic networks do not necessarily reflect transmission, as they do not integrate key information. For instance, the number of substitutions observed is affected by the time elapsed since infection and by within-host diversity; multiple clones can coexist in the same individual, and they may be differentially transmitted. Thus, the assumption that the SNPs are gained from an ancestral reconstructed genotype and that diversification events represent transmission events may not be correct.

4.4.2 TransPhylo identifies index cases not detected by contact tracing

The TransPhylo approach integrates sample timing and genetic relatedness, and allows for within-host diversity, thereby avoiding the assumption that diversification represents transmission. TransPhylo produces posterior reconstructed transmission events and timing for each cluster, which can be visualized in many ways, including consensus trees (**Figure 4.1B**) and the posterior probability of infection between cases (**Figure 4.1C**). In our study, TransPhylo estimated that there were unsampled cases, with different numbers of unsampled cases in different clusters. For the main results, we selected a clock rate value of 0.363 SNPs/genome/year, which is the rate obtained by others [196, 198]. The results show that most transmission clusters had 2 or

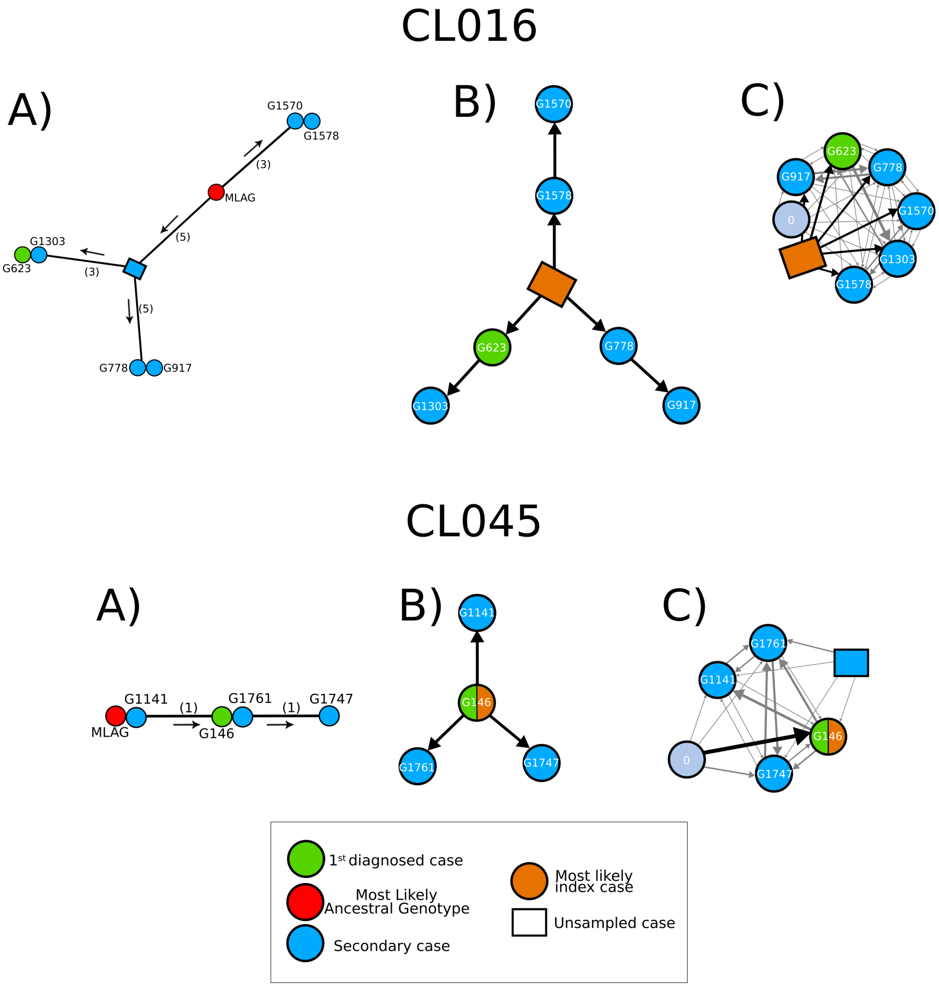


Figure 4.1: Comparison of transmission reconstruction methods. The figure shows for clusters CL045 and CL016 the inferred genetic network (A) and the consensus transmission tree inferred from TransPhylo (B and C). In addition we show the strength of the TransPhylo prediction (C). When the index case is sampled, it is depicted by a direct black arrow connecting the grey "0" circle to the respective individual. This is the case for G146 in CL045. When the index case is missing, this is represented by an orange square connected to all cases, as in CL016. Any other unsampled tuberculosis case is shown using a blue square symbol.

fewer unsampled cases (62%). Only 1 cluster (CL026) had a median number of unsampled cases greater than 5 (**Figure 4.2**). The estimated number of unsampled cases is lower if a higher substitution rate is assumed, with very few unsampled cases under a fast clock assumption (**Supplementary Figure 4.10**). This effect occurs because with a faster assumed clock rate, timed tree branches are shorter, and TransPhylo is less likely to place unsampled cases along the branches.

TransPhylo's augmented MCMC approach allows us to extract the inferred index case for each posterior tree. **Figure 4.3** shows for every cluster the probability that each diagnosed individual in the cluster was the index case, along with the individuals' diagnosis times. There are 6 clusters in which the index case was most likely unsampled. For those clusters where the index case was likely sampled, the index case is not always the first diagnosed individual (33%); the index case's diagnosis can be many months after the first diagnosis (e.g., CL005). Most of the clustered cases were not detected as contacts in the contact tracing epidemiological investigations.

There is general agreement between TransPhylo and the genetic network approach in identifying those clusters in which the index case is likely sampled. For 7 clusters (33%), both approaches predicted that the index case had been sampled. TransPhylo predicted the presence of an index case in 8 additional clusters in which the exact MLAG genotype did not occur, and consequently the genetic network approach did not predict that the index case was sampled. For the rest of the clusters (n = 6), neither TransPhylo nor the genetic network identified a likely index case. However, despite this general agreement, the methods do not always agree on which patient was the likely index case.

Genetic networks predicted the same index case as TransPhylo in only 2 (13%) of the 15 clusters with a likely sampled index case. This disagreement is likely associated with the fact that the time of sampling and rate of genetic change are not taken into account in the genetic network prediction. Also, the genetic network approach predicted more unsampled genotypes than TransPhylo, reflecting the fact that some of the missing genotypes likely existed but evolved within a host and were not transmitted (**Supplementary Figures**

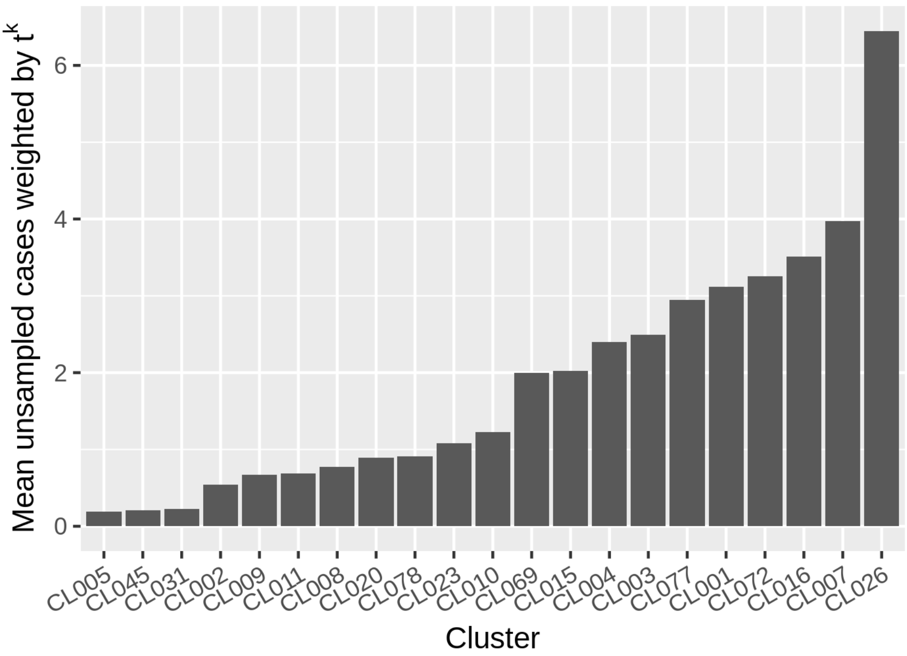


Figure 4.2: Weighted mean number of unsampled tuberculosis cases. For each posterior transmission tree, we associate a weighting factor tk, where k is the number of sampled cases for which transmission happened after diagnosis, and t = 0.1. This accounts for the fact that individuals are treated once diagnosed, and so are less likely to transmit. This figure shows the mean number of unsampled cases for one of the simulated clock rates (0.363). The results for all clock rates appear in S5 Fig.

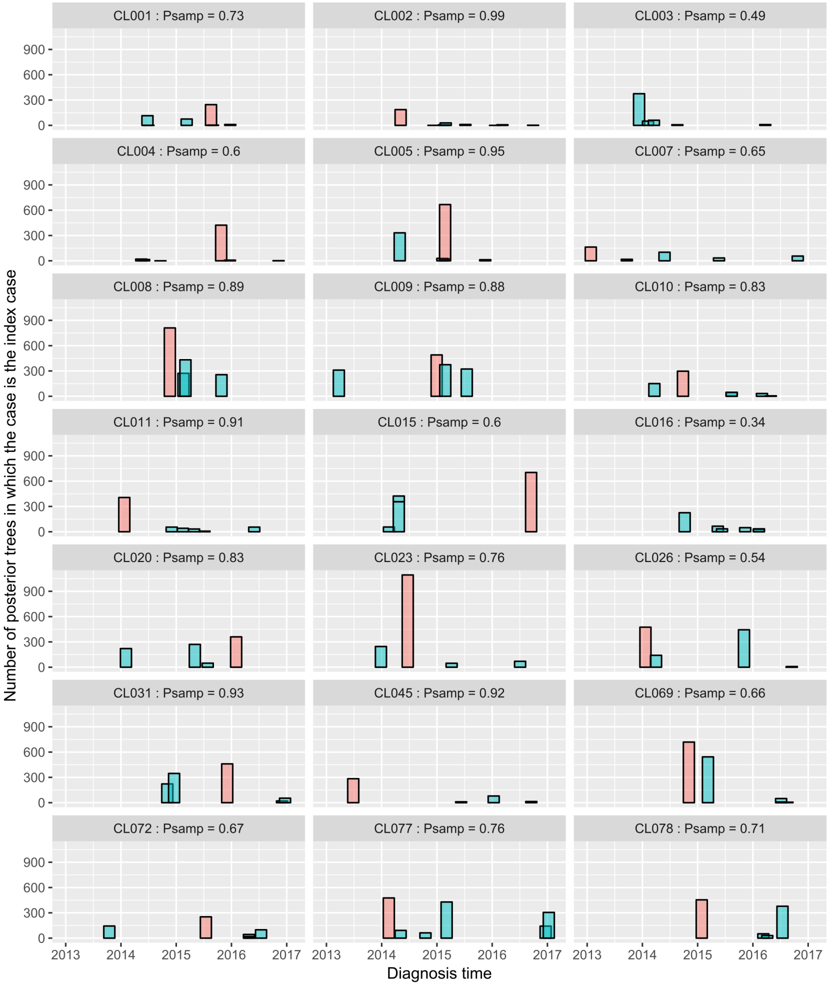


Figure 4.3: The posterior probability that each individual is the index case for a cluster versus the time of diagnosis of the individual. The individual with highest posterior probability to be the index case is shown in red for each cluster. In some clusters, the first diagnosed case was the estimated index case, in that it had the highest probability of being the index case (e.g., CL002). In contrast, in the majority of clusters the most likely index case was not the first diagnosed individual (e.g., CL010 and CL023) or was not sampled (e.g., CL016 and CL003). The Psamp values are the posterior probability that the index case was any of the sampled individuals—in some clusters (e.g., CL003) the index case was likely to have been an unsampled individual.

4.8-4.9).

4.4.3 Timing of events reveals TB cases transmitting before diagnosis or symptom onset

Because it integrates information about case timing and the molecular clock alongside genetic relatedness of isolates, TransPhylo can estimate the timing of transmission, which can be compared to diagnosis times and reported symptom times. Thus, triangulation of relevant dates and timing should allow us to use TransPhylo to evaluate how much transmission could be averted by earlier identification of individuals with TB or by isolating patients during the first stages of treatment.

First, we extracted transmission trees corresponding to one of the molecular clock rates (0.363 SNPs/genome/year) and selected all individuals for whom the probability of transmitting was greater than 0.6. We then compared inferred transmission times to diagnosis times and to the reported times of symptom onset. A total of 14 individuals had a high likelihood of being transmitters (**Figure4.4**). We reasoned that if our prediction was accurate, many transmission events should happen between the onset of symptoms and diagnosis; this is the case for 9 out of the 14 TB individuals. However, when we looked at the time of transmission in the other 5 cases, transmission occurred before symptom onset or diagnosis (G815, G258, G201, G1775, and G1449). Notably, 3 out of the 5 individuals were sputum negative at the time of diagnosis, suggesting that they were infectious before, but not at the time of detection. The time of first transmission event for all cases in every cluster is reported in **Supplementary Figures 4.11-4.17**, including combinations of different probabilities and clock rates.

To evaluate the feasibility that transmission happened before symptoms, we analyzed the contact tracing and epidemiological data available for 1 of the cases. G1449 was a credible transmitter before symptom onset (**Figure 4.4**). G1449 clustered with another case, G1011, which was the 18-year-old daughter of G1449. Both were identified almost simultaneously, but the daughter was the first to seek care. Thus, she was considered the index case, and contacts were screened. G1449 was identified during screening a few days later. We estimate that G1449 infected G1011 less than 2 years before, which is compatible with the incubation time of latent TB in persons without known risk factors. Conversely, if G1011 infected G1449 after symptom onset, then G1449 had to develop symptoms in less than 1 month since infection, which is less likely than the other scenario.

We also reasoned that the probability of transmission should be compatible with the known epidemiological characteristics of the patients. We used the time of arrival of foreign nationals to evaluate the feasibility that transmission happened when we predicted. In all individuals with a high probability of transmitting TB, transmission happened after arrival to the country. Conversely, there were 5 individuals for whom transmission was predicted to have happened before arrival, so for these individuals there is a contradiction between the prediction (if they were transmitters) and the epidemiological history. In all 5 cases, our approach did not identify them as credible transmitters (probabilities of transmission < 0.3; **Supplementary Table 4.5**).

Finally, we examined whether individuals with longer estimated times between infection and diagnosis had higher numbers of secondary TB cases. This would be expected, since delayed diagnosis gives an individual the opportunity to expose others and to become the index case of a cluster. We found that the estimated time to diagnosis was longer for those individuals predicted to have infected 2 or more secondary cases, but the results are variable, as expected given that many other factors affect probabilities of transmission and infection (see **Supplementary Figure 4.18**).

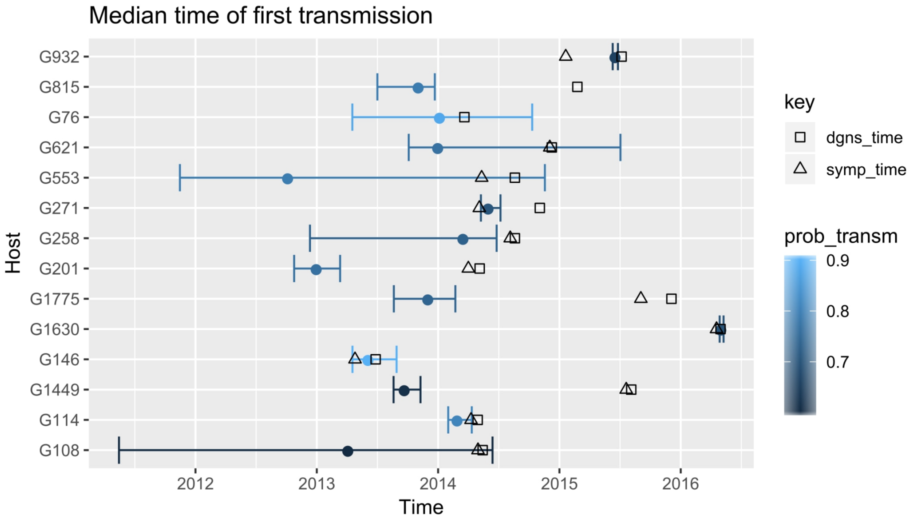


Figure 4.4: Resampled median time of first transmission. The graph represents the median time of the first highly likely transmission for individuals for whom the posterior probability of transmitting (`prob_transm`) was greater than 0.6, under a clock rate value of 0.363 SNPs/genome/year. For each case, the diagnosis time (`dgns_time`; squares) and, where known, the symptom onset time (`symp_time`; triangles) are added. Analogous graphs for different transmission probability cutoffs, and without cutoffs, are shown in Supplementary Figures.

4.4.4 Identification of transmitters allows association of risk factors to transmission

For 66% of the clusters analyzed, the index case identified by TransPhylo was either unsampled or not the first diagnosed case (14 out of 21). This suggests that index cases based on diagnostic dates can be misleading. In addition, analyses of risk factors associated with transmission using molecular epidemiology data have been traditionally performed on group measures of clustering (clustered versus unique cases, association with cluster sizes). This approach obviates the fact that not all individuals with TB are transmitters, and thus risk factors associated with transmission are difficult to disentangle from those associated with infection. Our identification of likely index cases and transmitters allows us to explore whether risk factors have a different distribution specifically among likely transmitters. We combined likely transmitter cases together with the index cases predicted by TransPhylo (n = 23) and compared them to the other clustered cases (n = 61). Our statistical analysis is limited by the low number of clusters and the low number of transmitters that were unequivocally identified. Also, clustered cases are a composite of transmitters, non-transmitters, and those cases that cannot be confidently assigned to either category. Still, relevant differences between likely transmitters and the rest of the clustered cases can be identified (**Figure** 4.5).

As a proof of concept, transmitters tended to be diagnosed later (mean diagnostic delay 85 days versus 54 days), although this difference is not statistically significant. Other variables also suggest important differences between being a transmitter and simply being part of a cluster. Transmitters were significantly enriched in diabetic patients in both univariate (Fisher's exact test; OR 0.19 [95% CI 0.02–1.10], $p < 0.003$) and multivariate (logistic regression; OR 23.77 [95% CI 2.53–339.69], $p < 0.009$) statistical analyses. It has been suggested before that diabetic patients tend to have larger TB cavities, a factor known to be associated with transmission [199]. Finally, we confirm previous reports showing that individuals who are smear negative at the time of diagnosis can be transmitters (37% in our dataset). However, we take these results with caution. We repeated the analysis comparing transmitters to non-clustered cases, and diabetes was still enriched (27% versus 10%), but not significantly (p = 0.06). While small sample sizes do not allow us to draw more conclusions, these preliminary results show the importance of differentiating between being a transmitter and being infected.

Characteristic	All Patients (n=84)	Transmitter (n=23)	Other samples (n=61)	Odd ratio (95% IC)	*p*-value*
Age (years)					
<18	10 (12%)	2 (9%)	8 (13%)	**reference**	**reference**
19-34	18 (21%)	4 (17%)	14 (23%)	0.7 (0.15-2.66)	0.76
35-65	48 (57%)	14 (61%)	34 (56%)	1.23 (0.42-3.75)	0.8
>65	8 (10%)	3 (13%)	5 (8%)	1.66 (0.23-9.51)	0.67
Sex					
Female	27 (32%)	8 (35%)	19 (31%)	1.17 (0.36-3.59)	0.79
Male	57 (68%)	15 (65%)	42 (69%)	**reference**	**reference**
Place of birth					
Spain	59 (70%)	16 (70%)	43 (70%)	**reference**	**reference**
Other country	25 (30%)	7 (30%)	18 (30%)	1.04 (0.30-3.27)	1
Sputum smear					
Positive	52 (62%)	13 (57%)	39 (64%)	0.7 (0.23-2.12)	0.61
Negative	31 (38%)	10 (43%)	21 (34%)	**reference**	**reference**
Disease type					
Pulmonary	78 (93%)	22 (96%)	56 (92%)	1.95 (0.20-97.02)	1
Extrapulmonary	6 (7%)	1 (4%)	56 (92%)	**reference**	**reference**
Alcoholism	15 (19%)	4 (17%)	11 (18%)	0.91 (0.18-3.61)	1
Diabetes	8 (9%)	5 (22%)	3 (5%)	5.14 (0.90-36.45)	**0.03**
HIV infected	7 (8%)	5 (22%)	3 (5%)	1.01 (0.09-6.87)	1
Social exclusion	10 (13%)	1 (4%)	9 (15%)	0.27 (0.01-2.22)	0.27
Health Care Worker	5 (6%)	2 (9%)	3 (5%)	1.82 (0.14-17.27)	0.61
Imprisonment	5 (6%)	2 (9%)	3 (5%)	1.79 (0.14-16.84)	0.61
Diagnostic delay	16 (20%)	7 (30%)	9 (15%)	2.25 (0.60-8.17)	0.21

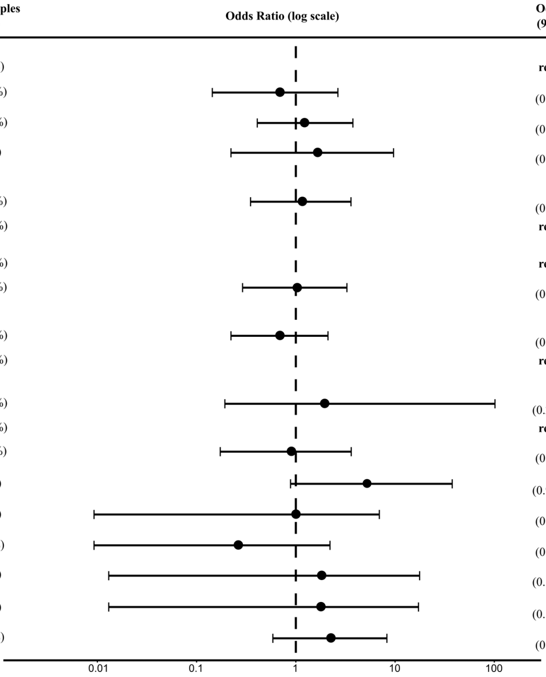


Figure 4.5: Fig 5. Epidemiological characteristics of the cases used to identify transmission risk factors. Note that the data do not include all the study samples: for 5 clusters we were not able to identify a likely transmission event, and these clusters were excluded from this analysis. Transmitters are defined as those individuals estimated to be likely transmitters and/or likely index cases detected by TransPhylo. The figure shows estimated odd ratios for each risk factor tested. *Fisher's exact test. Comparisons were made between transmitter cases and the rest of the clustered samples.

4.5 Discussion

We present a genomic-based approach to unveil individual TB transmission links between patients within transmission clusters. Importantly, our method allows us to identify, or infer the absence of, the most likely index case, as well as estimate the number of unsampled cases within a cluster. These findings may contribute to reorienting contact investigation strategies in terms of to whom and where TB testing should be done. In addition, we identify potential transmission events during the sub-clinical disease stage, suggesting the need to incorporate early disease stages in epidemiological models and TB control programs.

WGS has been shown to be superior to previous genotyping tools in identifying TB cases likely to be of recent transmission [200]. Nevertheless, there is only an agreement of 30%–50% between those identified by WGS as TB cases of recent transmission and those identified by contact tracing [149]. This scenario indicates that likely index cases are missing, and improved contact investigation strategies are required in order to detect those individuals. A recent clinical trial [42] showed that close contacts of index cases identified by active case finding have better TB cure rates than those identified by passive case finding. Thus, identification of index cases has implications at the population and at the individual care level. In this study, we showed that in up to 28% of clusters there is no evidence that the index case is included among the individuals in the cluster. For those clusters in which an index case was detected, 60% of the time the index case was not the individual first diagnosed with TB, suggesting that efforts to identify transmission are imperfect.

The reasons that index cases are not sampled in a study may be multiple and will probably vary by clinical setting. First, index case transmission could have occurred prior to the sampling time. This is very likely in our analyses, where we potentially include older transmission events, though fixed SNP cutoffs may not perfectly delineate transmission clusters [201]. Furthermore, we missed those individuals with culture negative status at the time of diagnosis, and they may have contributed to transmission. However, it is worrying that individuals with TB may have been missed by control programs and may remain actively transmitting in the population. In Valencia, around 3,000 contacts are investigated every year following the European Centre for Disease Prevention and Control guidelines. Still, a large percentage of the clustered cases were not identified as contacts, consistent with similar published studies [149, 27, 202, 147], including index cases predicted in our analysis.

With our approach we could separate likely transmitters from other clustered cases, rather than treating each cluster as a single unit, and so could associate biological, epidemiological, and demographic variables with transmission. Our dataset has 2 major shortcomings—namely the low number of transmission links with enough statistical support and the fact that only 21 clusters met the criteria for the analysis—and thus our clusters are not necessarily representative of the whole population. Still, our data suggest that certain risk and epidemiological factors are enriched among the transmitters, while others are depleted. In addition, we corroborate that individuals with negative sputum smear status can contribute to transmission (40% of index cases), as has been discussed previously [31, 203]. Larger population-based datasets including a larger number of clusters meeting the criteria will help to better define the exact role of these factors.

Our selection of TransPhylo as a tool to trace transmission was driven by the necessity of considering potential unsampled cases. There are other similar approaches that do not take unsampled cases into account [204] or that use a model more suited to environmental reservoirs [205, 206]. In addition, we could not make predictions for some transmission clusters due to the limited observed within-cluster diversity, as anticipated previously [207]. Thus, our analysis focused on those events that we could robustly estimate. It is important to note that predictions may be sensitive to molecular rate variations. We focused our discussion on analyses using a molecular rate that is appropriate for MTBC lineage 4 strains, which dominate the local setting. However, other settings will need to calibrate the model with a different rate as it is becoming apparent that the rate for different lineages may vary [153].

The fact that we estimate that approximately 35% of transmission events occurred before symptom onset could have several explanations. Patient-reported times of symptom onset are subjective, and if symptoms were mild, disease may not have been recognized for some time. However, in most cases the time difference between symptom onset and transmission spans several weeks or even months. Recently it has been speculated that sub-clinical transmission may exist and be facilitated by unrelated cough [208]. Here we show evidence for transmission during the asymptomatic phase of disease, in which the transmission probability is lower than during exacerbated disease, but non-negligible [208, 209].

There is evidence from clinical trials of sputum smear positive individuals who are otherwise healthy being potential transmitters [19]. This is in line with recent evidence showing a spectrum of different disease states (from almost healthy to diseased [15]) and the possibility that a percentage of those traditionally considered latently infected TB cases in reality are active TB cases with sub-clinical disease [19, 43]. Our transmission analysis suggests that sub-clinical disease may jeopardize current TB control strategies, in line with results from epidemiological models [43].

A limitation of our method is that we could not test it on other publicly available genomic datasets. One reason is because it is difficult to obtain cases' associated epidemiological data, especially those related to symptom onset (which is a key variable of our study). Despite this, we validated our method by (1) conducting sensitivity analyses using different TransPhylo parameters and (2) comparing the predicted transmission time for foreign-born TB cases with the time of immigration. Nevertheless, the lack of published datasets with the relevant epidemiological data highlights the need to incorporate these variables in prospective TB epidemiological studies.

In conclusion, our individual-based transmission inference method demonstrates that many likely transmitters, including index cases, are missed by contact investigations. Strikingly, a substantial proportion of these transmitters likely spread TB during sub-clinical disease. Future work aligning biomarkers and epidemiological research will help to elucidate host biomarkers of transmission during the spectrum of TB infection, to design better TB control strategies.

4.5.1 Acknowledgements

We want to thank Dick van Soolingen and Rana Jajou for assistance on benchmarking the bioinformatics pipeline.

The work described in the present chapter has been published as a Research Article in *PLoS Medicine* Journal:

Xu Y, Cancino-Muñoz I, Torres-Puente M, Villamayor LM, Borrás R, Borrás-Máñez M, et al. High-resolution mapping of tuberculosis transmission: Whole genome sequencing and phylogenetic modelling of a cohort from Valencia Region, Spain. Murray MB, editor. PLoS Med. 2019;16: e1002961. https://doi.org/10.1371/journal.pmed.1002961

4.6 Supplementary Data

4.6.1 Supplementary Methods

Ongoing population study

The present study is part of an ongoing population TB study. This global study consisted in the recollection and WGS analysis of a total of 785 positive TB clinical samples during 2014-2016. Using SNPs distances between isolates (15 SNPs), we detected that 41% (n=325) of all samples were in transmission. Although the majority of clustered cases comprised two samples, we detected transmission clusters that involved up to 12 TB cases. From this first analysis, we obtained the samples that we used in the present study. We selected to 21 genomic clusters (17.3% of all clusters identified) that corresponded 117 isolates (36% of the total transmission). These genomic groups that had at least four cases and had 2 SNPs difference between all samples involved. Furthermore, we made a comparison analysis to see whether our sampling selection was representative of the whole population (**Supplementary Table 4.4**).

Timed tree reconstruction

Although the accepted value for the TB substitution rate in the community is approximately 0·3-0·5 substitutions per genome per year[1,2], our data seem to suggest that the rate may vary both between clusters and within clusters. For example, in some clusters the SNP distances between pairs of hosts are not consistent with the case timings. For example, a host sampled earlier in time can seem to have accumulated more SNPs than a host sampled later (compared to inference of an ancestral sequence for a cluster), or vice versa. In such cases, an estimated timed phylogenetic tree using a low clock rate (as is normally assumed in TB) would place the earliest sequence in the cluster quite far back in time compared to the most recent sampled case.

We therefore need to incorporate rate uncertainty in the inference framework. However, one challenge is that we do not know if and how rates vary across clusters, and furthermore, although treedater[3] allows us to fit a relaxed clock, the consequence of increased number of parameters and lack of signal contained in small cluster data mean that the branch length estimates may not be reliable. Therefore, we adopt a simple approach whereby instead of using a single timed phylogenetic tree for each cluster, we sample clock rates from a known distribution and use treedater to estimate timed trees for all clusters by fixing the clock rate to be in the range of one of our sampled rate values with a margin of ±. So for each sampled clock rate, we obtain timed tress corresponding to all clusters and we used TransPhylo [4] and the method outlined below to infer the transmission trees. We then pool the transmission trees for each clock rate. By inspecting this combined posterior, we can compare between rates and see if any of the interesting quantities are sensitive to changes in clock rate. We perform a meta-analysis of 18 publications reporting clock rates per year from different studies of MTBC (see **Supplementary Table 4.2**). We obtained a mean rate of 0.32 (±0.022-0.44) but with very wide range of values (0.14-0.59). Thus, we chose to use a log-normal distribution with log-scale mean and standard deviation of -0.7 and 0.5, respectively, for the sampling distribution of the clock rate and δ = 0.2.

Transmission inference

We develop our method of simultaneous transmission inference on many clusters based on `TransPhylo`, a Bayesian method to reconstruct transmission trees from pathogen phylogeny. In `TransPhylo`, an MCMC method is used to draw samples from the posterior distribution of transmission tree and model parameters given a timed tree reconstructed from sequenced isolates (4.1).

$$P(T, \theta | P) \propto P(P|T, \theta) P(T|\theta) P(\theta), \qquad (4.1)$$

where T is transmission tree, P is timed tree and θ collects the model parameters. The transmission tree is represented by a matrix whose columns are the times of infection, times of sampling and the infectors, and whose rows correspond to infected individuals. If a case is not sampled, then the corresponding entry for time of sampling is empty. In the posterior trees that

TransPhylo produces, the number of rows in T can be variable across iterations, because of the addition/removal of unsampled cases; reversible-jump MCMC is used in TransPhylo to account for changes of dimensionality.

The transmission tree contains information about who infected whom and when, and also whether a case is sampled or not. The timed tree shows evolution history of pathogens sampled from hosts and is constructed from known methods of phylogenetic tree-building, such as neighbor-joining, maximum likelihood or Bayesian methods. The TransPhylo posterior thus reflects our updated belief of the transmission pattern and epidemiological parameters after observing the timed tree of the sequences.

In practice, especially in low-incidence settings, we often define transmission clusters based on our knowledge of the genomics and the epidemiology of cases, such that we are quite confident that transmissions occurred within clusters and were less likely between clusters. This makes it more amenable to analyze multiple clusters simultaneously than to work with a single large phylogeny of all sequences, because these clusters tend to be separated by long branches and TransPhylo will put many unsampled cases along these branches and so will not explore transmission within clusters efficiently.

In order to develop a framework for simultaneous transmission inference, a straightforward extension to (4.1) would be to carry out our Bayesian inference in an augmented tree and parameter space. More precisely, let $\mathbf{T}$, $\mathbf{P}$ and Θ be elements in the respective joint space of n clusters, that is, $\mathbf{T} = (T_1, \dots, T_n)$ with T_i the transmission tree for cluster i, and similarly for $\mathbf{P}$ and Θ, then

$$P(\mathbf{T}, \Theta|\mathbf{P}) \propto \prod_{i=1}^{n} P(P_i|T_i, \theta_i) P(T_i|\theta_i) P(\theta_i), \tag{4.2}$$

assuming independence between clusters. MCMC simulation of (4.2) proceeds as it would in (4.1), with each step consisting of separately updating parameters and trees for all clusters. This would be no different from independently running *TranPhylo* once for each cluster. In order to allow information to be shared between clusters, we decompose θ into sharing and non-sharing parts, $\theta = (\theta^s, \theta^{ns})$, where we can choose which parameters should be shared among clusters and which should not. The posterior distribution becomes

$$P(\mathbf{T}, \Theta|\mathbf{P}) \propto \prod_{i=1}^{n} P(P_i|T_i, \theta_i^{ns}, \theta^s) P(T_i|\theta_i^{ns}, \theta^s) P(\theta_i^{ns}) P(\theta^s). \tag{4.3}$$

Note that θ^s does not have index i because it is the same for all clusters. The update of θ^s is based on the Metropolis-Hastings ratio of likelihoods of all clusters.

With the above framework, not only can we handle the statistical inference with multiple transmission clusters simultaneously, we can also choose which parameters should be shared. The latter has both epidemiological and computational implications — if we believe that certain parameters, such as the basic reproduction number (the expected number of secondary infections from any primary infection) and/or the sampling rate will be similar within each cluster, then we can easily encode this belief into (4.3). This offers great computational savings as the number of parameters is significantly reduced — avoiding $(n-1)\mathrm{n}(\theta^s)$ parameter estimations where $\mathrm{n}(\theta^s)$ denotes the number of parameters in θ^s.

4.6.2 Supplementary Results

Simulations

We simulated two outbreaks using `TransPhylo`'s simulator function `simulateOutbreak`, with different model parameters: 1) `neg` — within-host diversity; 2) `off.r` — first parameter of negative binomial offspring distribution, or equivalently the basic reproduction number; and 3) and `pi` — sampling rate, from the posterior. Because of the stochastic nature of the simulator, we obtain different time-trees. We then apply the new joint inference routine to infer the model parameters and compare them with those obtained from running `TransPhylo` separately. For both the two independent runs and the joint routine the same number of MCMC iterations, 10^4, were used, with a burn-in of 20%, hence, the computer time of the joint routine is about the same as the total time of running the two independent runs sequentially (**Supplementary Figure 4.6**).

TransPhylo parameters

For both `neg` and `off.r`, an exponential prior $Exp(1)$ was used; while a Beta prior $Beta(5, 1)$ was used for `pi`. We see that the offspring distribution parameter, which is also the R_0, is very robust to changes in clock rate. A high sampling proportion of 0.7 was observed even with the lowest clock rate, reflecting our prior belief of sampling. In addition, the `neg` parameter is not affected by small changes in clock rate, however it is significantly lower if the clock rate is very high relative to the other rates (**Supplementary Figure 4.7**).

4.6.3 Supplementary Tables

Reference	Publication year	MTBC lineage	No. of samples	Clock rate (genome per year
Ford et al. [210]	2011	L4	33	0.34
Walker et al.[150]	2013	All	390	0.5
Bryant et al.[211]	2013	L1, L2, L3 and L4	199	0.3
Ford et al.[159]	2013	L4	36	0.36
Roetzer at al.[54]	2013	L4	86	0.44
Bos et al.[128]	2014	All	261	0.22
Merker et al.[212]	2015	L2	110	0.44
Luo et al.[213]	2015	L2	393	0.2
Eldhom et al.[214]	2015	L4	252	0.29
Kay et al.[215]	2015	L4	165	0.22
Duchêne et al.[196]	2016	All and L4	261 and 252	0.24 and 0.25
Bjorn-Mortensen et al.[216]	2016	L4	182	0.47
Liu et al.[217]	2018	All	160	0.2
Merker et al.[198]	2018	L2	220	0.41
Duchêne et al.[218]	2018	L2	110	0.41
Rutaihwa et al.[219]	2018	L2	308	0.59
Brynildsrud et al.[124]	2018	L4	269	0.21
Meehan et al.[143]	2018	L4 and L5	324	0.14

Table 4.2: Meta-analysis table for different published MTBC clock rates.

Cluster ID	Number of isolates	Number of unique strains	Alignment length (SNP)
CL001	8	1	36
CL002	12	7	12
CL003	7	2	26
CL004	6	1	16
CL005	5	2	2
CL007	5	0	19
CL008	4	0	12
CL009	4	0	9
CL010	6	3	4
CL011	6	2	5
CL015	4	1	14
CL016	6	3	16
CL020	4	0	13
CL023	4	0	15
CL026	7	1	18
CL031	5	1	5
CL045	4	1	2
CL069	5	1	14
CL072	5	1	19
CL077	6	1	31
CL078	4	1	3

Table 4.3. Characteristics and genetic information about selected clusters. The table shows characteristics of the clusters including numbers of isolates and SNPs.

Characteristic	Global clustered cases (n=325)	Selected clustered cases (n=109)
Age (years)		
<18	23 (7%)	11 (10%)
19-34	78 (24%)	20 (18%)
35-65	194 (59.7%)	66 (61%)
>65	35 (10.7%)	12 (11%)
Sex		
Female	114 (35%)	33 (30%)
Male	211 (64.9%)	76 (70%)
Place of birth		
Spanish-born	230 (70.7%)	80 (73%)
Foreign-born	95 (29.2%)	29 (27%)
Sputum smear		
Positive	197 (60.6%)	66 (61%)
Negative	126 (38.8%)	41 (38%)
Disease type		
Pulmonary	290 (89.2%)	100 (92%)
Extrapulmonary	35 (10.7%)	9 (8%)
Alcoholism	69 (21.2%)	25 (23%)
Diabetes	34 (10.4%)	13 (12%)
HIV infected	24 (7.3%)	10 (9%)
Social exclusion	36 (11%)	13 (12%)
Health care workers	9 (2.7%)	5 (5%)
imprisonment	22 (6.7%)	8 (7%)
Diagnostic delay (100 days)	61 (18.7%)	23 (21%)
30 days	125 (38.4%)	46 (42%)
31-60 days	77 (23.7%)	25 (23%)
61-89 days	35 (10.8%)	14 (13%)
90 days	76 (23.4%)	32 (29%)
Contact tracing transmission	77 (23.7%)	24 (22%)

Table 4.4: Comparison table between the clustered cases detected in the global ongoing study and those selected in this research.

Case ID	Cluster ID	Probability	Transmitter	Tr_lower	Estimated transmission year	Tr_upper	Year of arrival	Transmission since arrival (years)	Transmission	Symptoms time
G1630	CL072	0.7	YES	2016.32	2016.33	2016.35	2015	1.33	after	2016.3
G201	CL002	0.77	YES	2012.81	2012.99	2013.19	2000	12.99	after	2014.25
G146	CL045	0.9	YES	2013.29	2013.42	2013.66	2008	5.42	after	2013.32
G1761	CL045	0.12	NO	2013.39	2013.43	2013.46	2015	-1.57	before	2016.69
G1099	CL001	0.15	NO	2002.79	2002.91	2003.06	2003	-0.09	before	2014.58
G1939	CL077	0.16	NO	2005.23	2005.33	2005.6	2009	-3.67	before	2016.78
G368	CL001	0.28	NO	2011.13	2012.03	2013.15	2014	-1.97	before	2014.41
G940	CL020	0.31	NO	2008.58	2008.92	2009.93	2015	-6.08	before	2015.31

Table 4.5: Comparison between time of arrival of foreign nationals in cluster and probability of transmitting TB in the region before symptoms.

4.6.4 Supplementary Figures

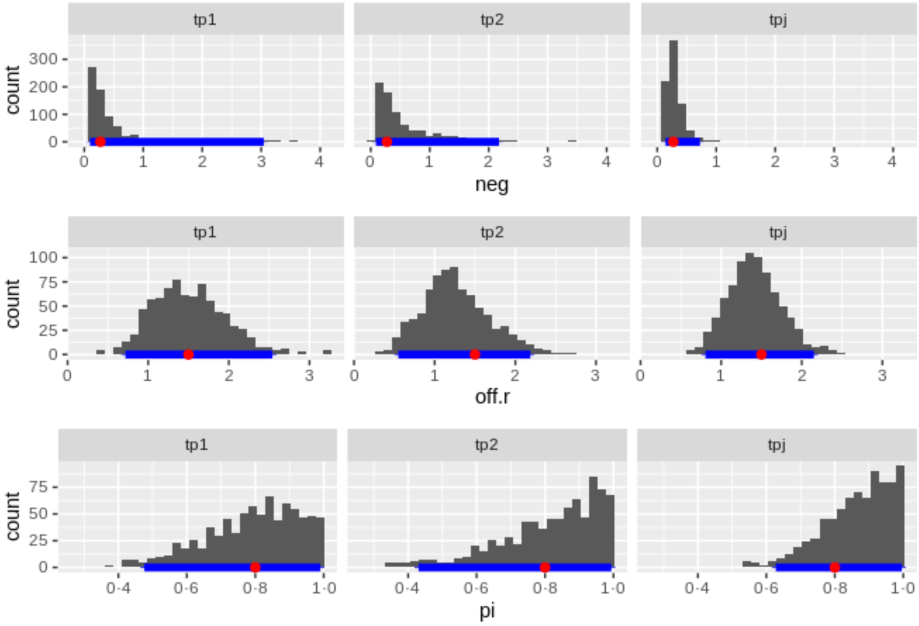


Figure 4.6. Histograms of model parameters `neg` — within-host diversity, `off.r` — first parameter of negative binomial offspring distribution, or equivalently the basic reproduction number, and `pi` — sampling rate, from the posterior. `tp1` and `tp2` — independent `TransPhylo` runs for the first and second timed tree simulated from the simulator function `simulateOutbreak`; `tpj` — `TransPhylo` run on the two timed trees with parameter sharing. Credible intervals are shown in blue and the true parameter values in red.

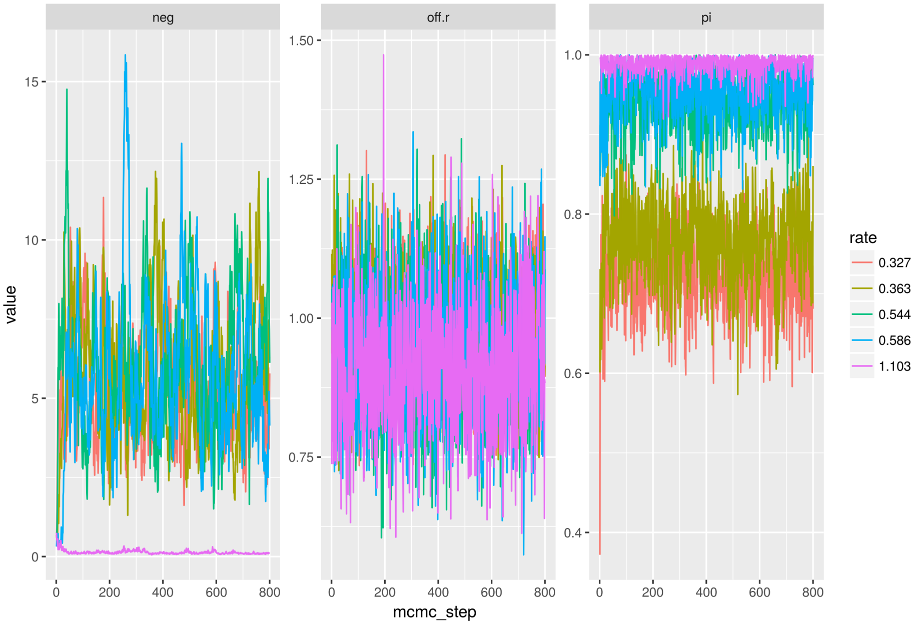


Figure 4.7. Trace plot of model parameters, colored by the simulated clock rates. The figure shows the trace plot of within-host diversity (`neg`), the offspring distribution parameter (`off.r`) and the sampling proportion (`pi`) from the MCMC run.

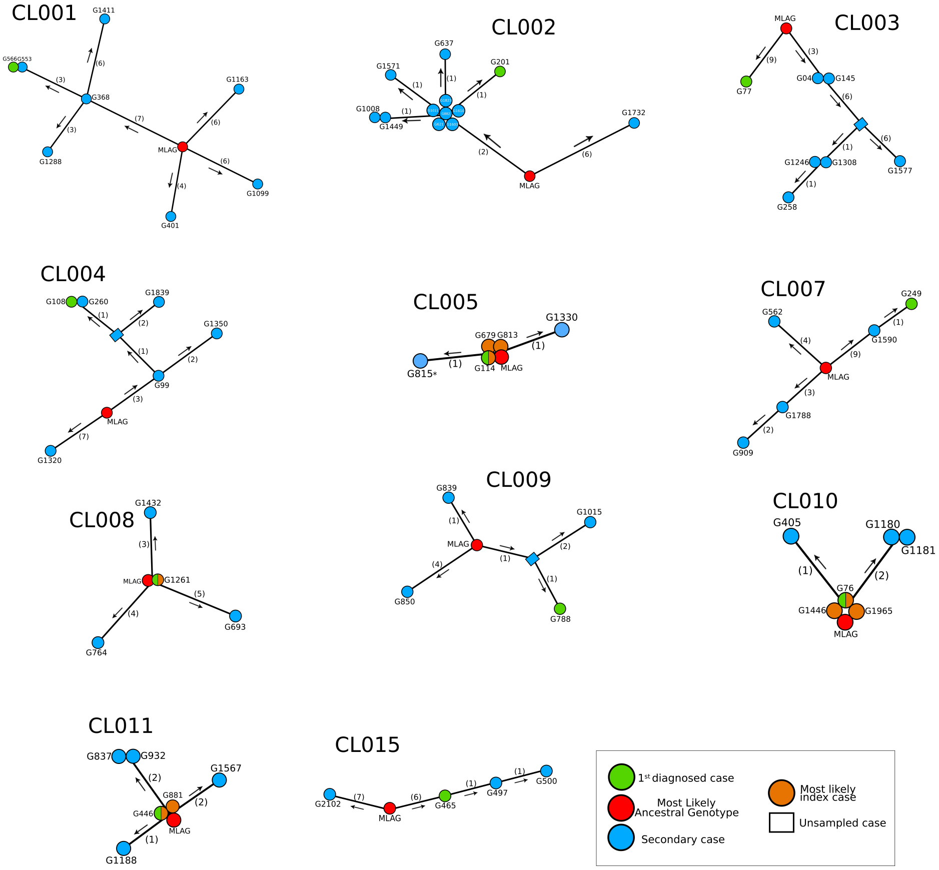


Figure 4.8. Genetic Network reconstruction of all transmission clusters used in the study(Part 1). The first diagnosed case is colour coded in green while the rest are colour coded blue. The Most Likely Index Case (brown colour) is the sample that has the same genotype as the predicted Most Likely Ancestral Genotype (red colour). The number inside brackets represents SNP difference between each isolate. The arrow denotes the high likely direction of the transmission.

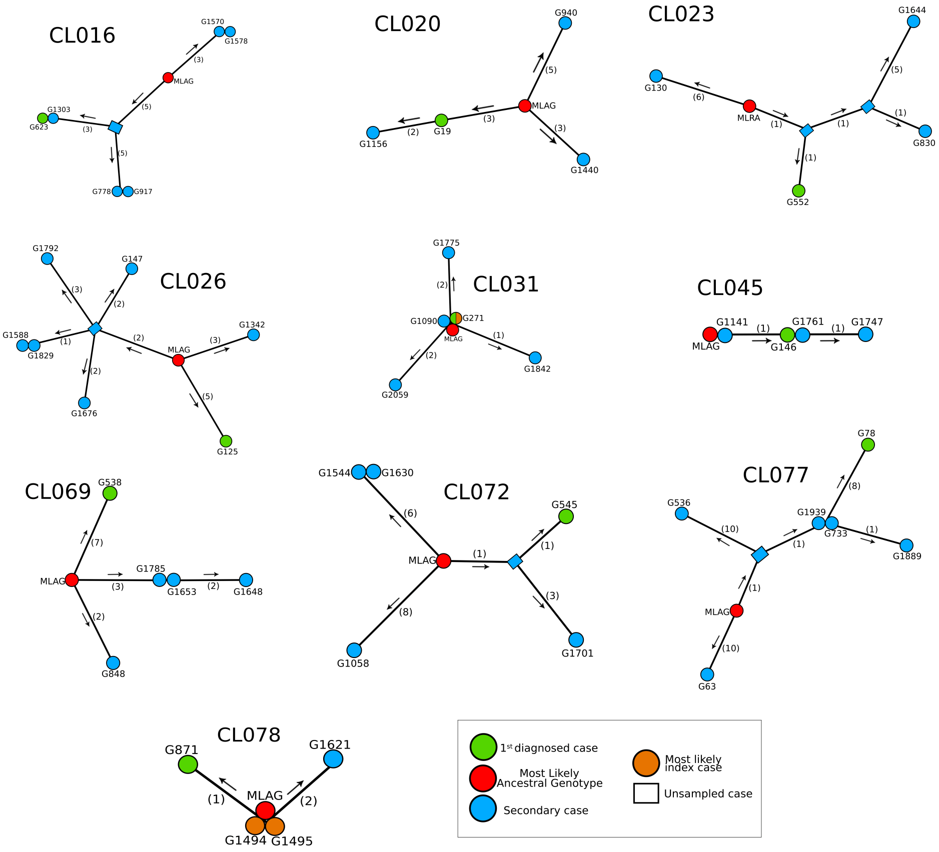


Figure 4.9. Genetic Network reconstruction of all transmission clusters used in the study (Part 2). The first diagnosed case is colour coded in green while the rest are colour coded blue. The Most Likely Index Case (brown colour) is the sample that has the same genotype as the predicted Most Likely Ancestral Genotype (red colour). The number inside brackets represents SNP difference between each isolate. The arrow denotes the high likely direction of the transmission.

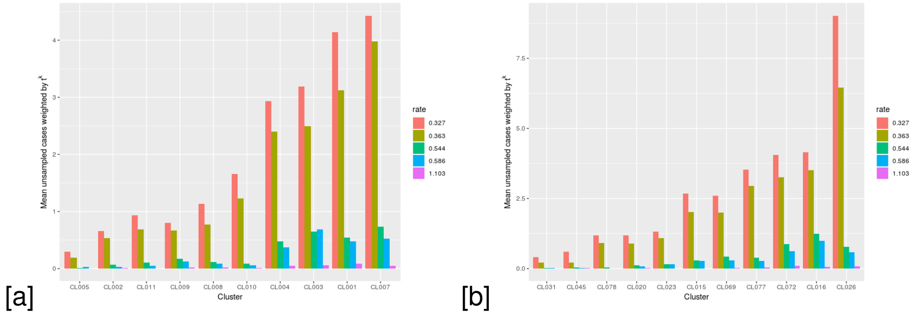


Figure 4.10. Weighted mean number of unsampled cases under different simulated clock rates.

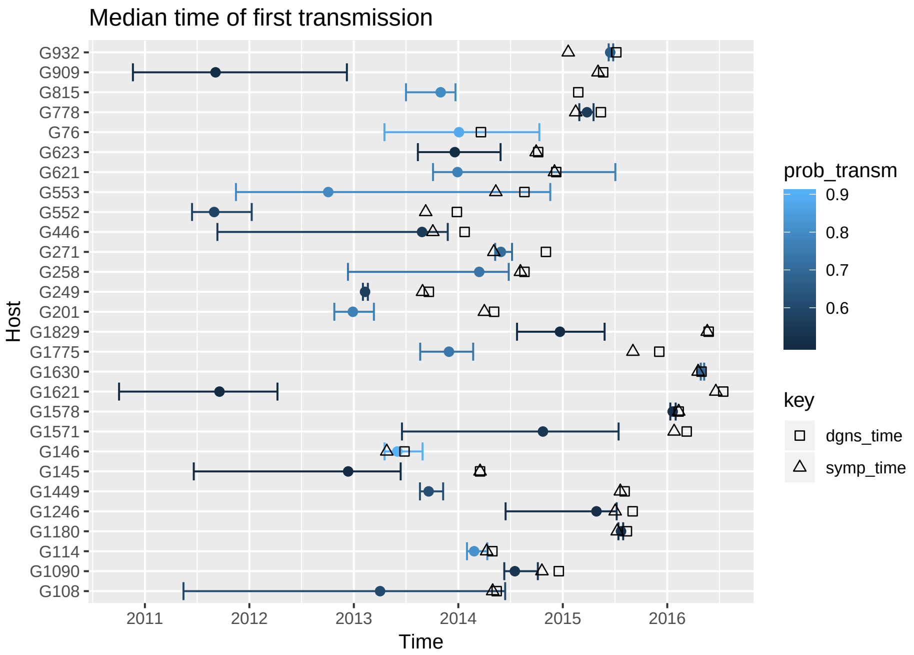


Figure 4.11. Resampled median time of first transmissions. The graph represents the median time of the first highly likely transmission for cases where the posterior probability of transmitting is greater than 0.5, under a clock rate of 0.363. For each case, the diagnosis time (square), and, where known, the symptom onset time (triangle) are added. Lighter colours indicate higher transmission probabilities. The range of the error bar indicates the 0.25 and 0.75 quantile.

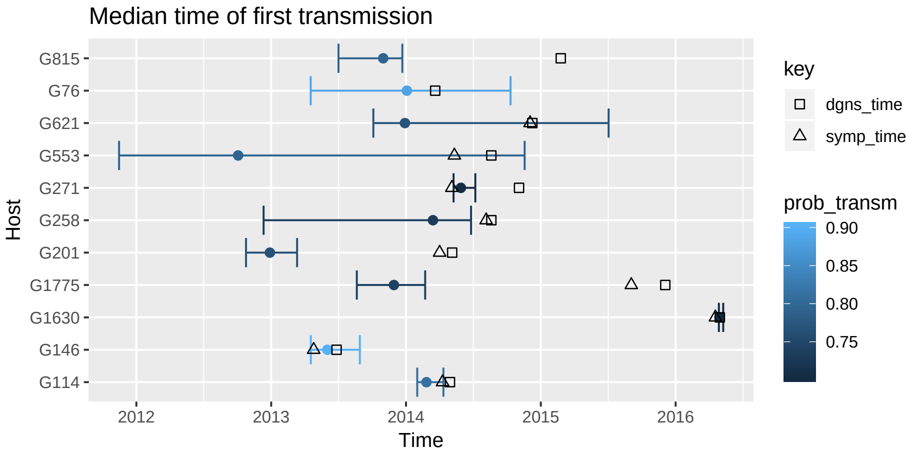


Figure 4.12. Resampled median time of first transmissions. The graph represents the median time of the first highly likely transmission for cases where the posterior probability of transmitting is greater than 0.7, under a clock rate of 0.363. For each case, the diagnosis time (square), and, where known, the symptom onset time (triangle) are added. Lighter colours indicate higher transmission probabilities. The range of the error bar indicates the 0.25 and 0.75 quantile.

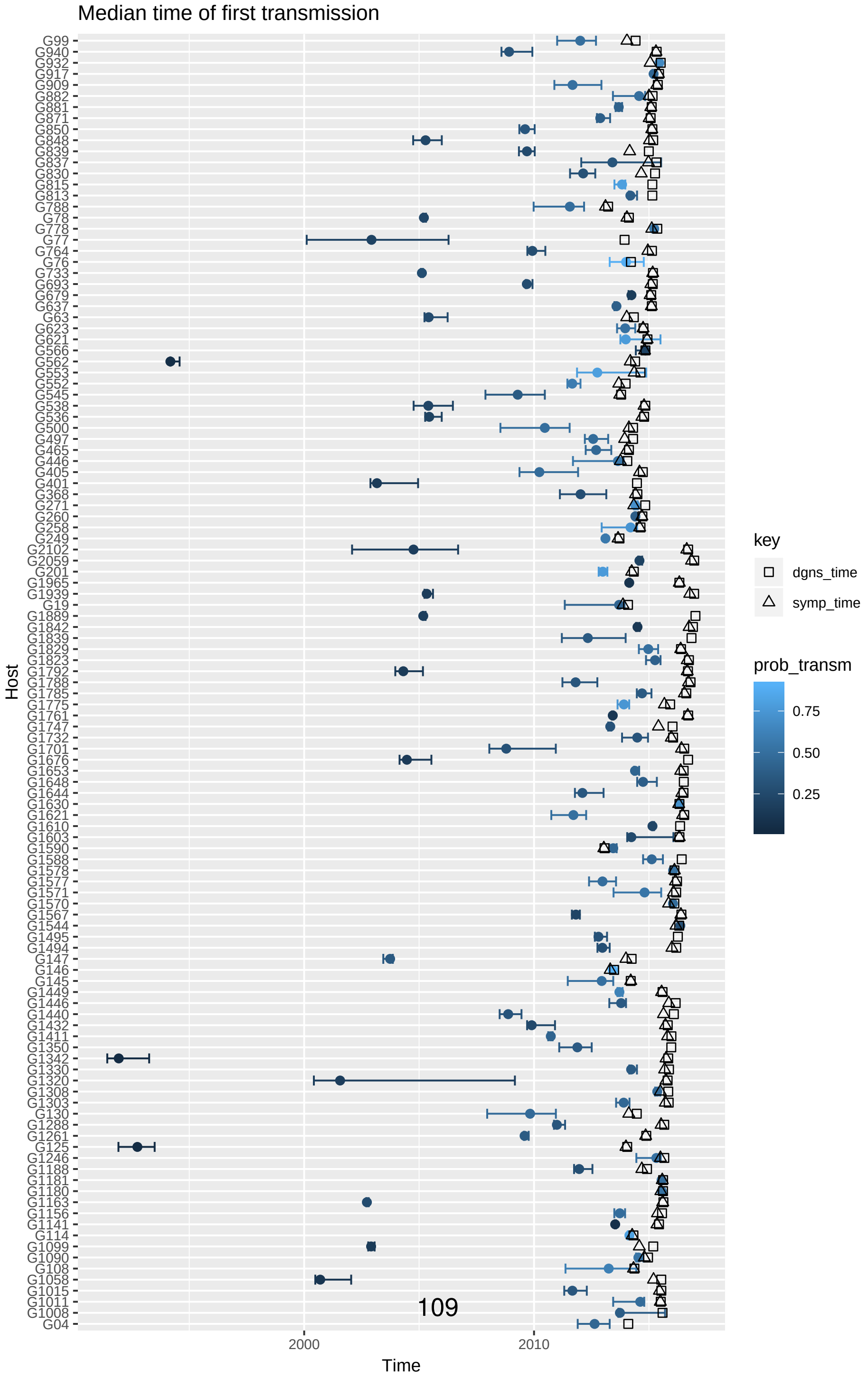


Figure 4.13. Resampled median time of first transmissions. The graph represents the median time of the first highly likely transmission for cases where the posterior probability of transmitting is greater than 0.7, under a clock rate of 0.363.

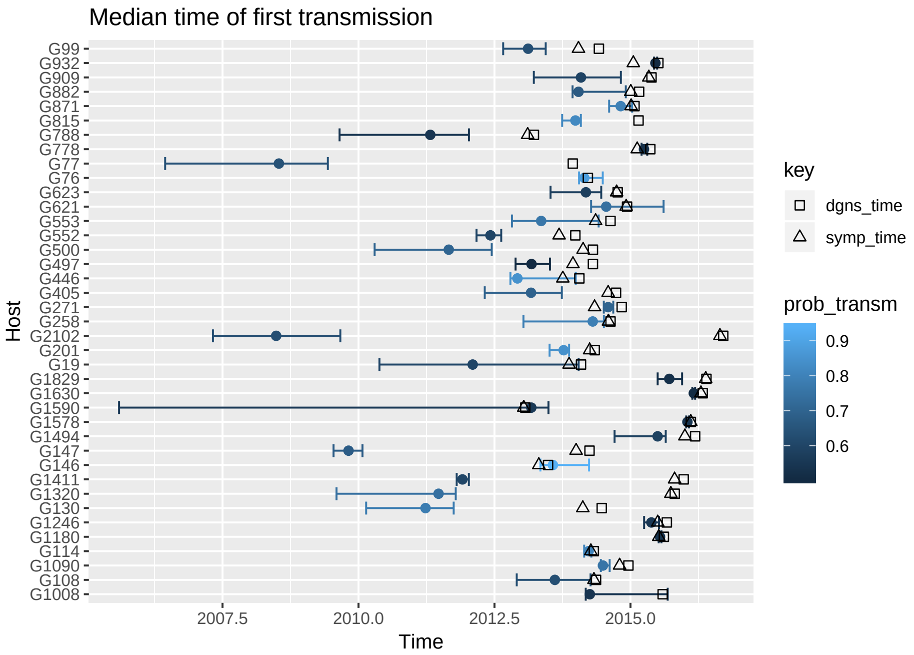


Figure 4.14. Resampled median time of first transmissions. The graph represents the median time of the first highly likely transmission for cases where the posterior probability of transmitting is greater than 0.5, under a clock rate of 0.544. For each case, the diagnosis time (square), and, where known, the symptom onset time (triangle) are added. Lighter colours indicate higher transmission probabilities. The range of the error bar indicates the 0.25 and 0.75 quantile.

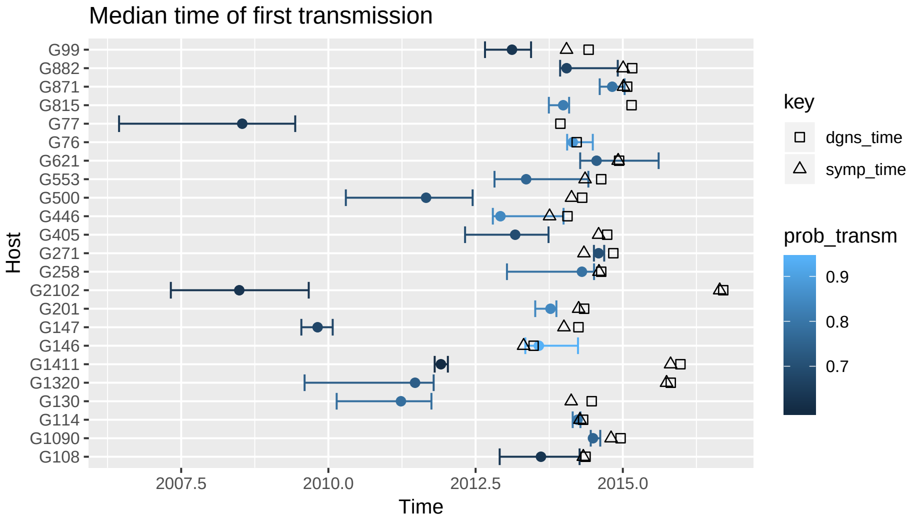


Figure 4.15. Resampled median time of first transmissions. The graph represents the median time of the first highly likely transmission for cases where the posterior probability of transmitting is greater than 0.6, under a clock rate of 0.544. For each case, the diagnosis time (square), and, where known, the symptom onset time (triangle) are added. Lighter colours indicate higher transmission probabilities. The range of the error bar indicates the 0.25 and 0.75 quantile.

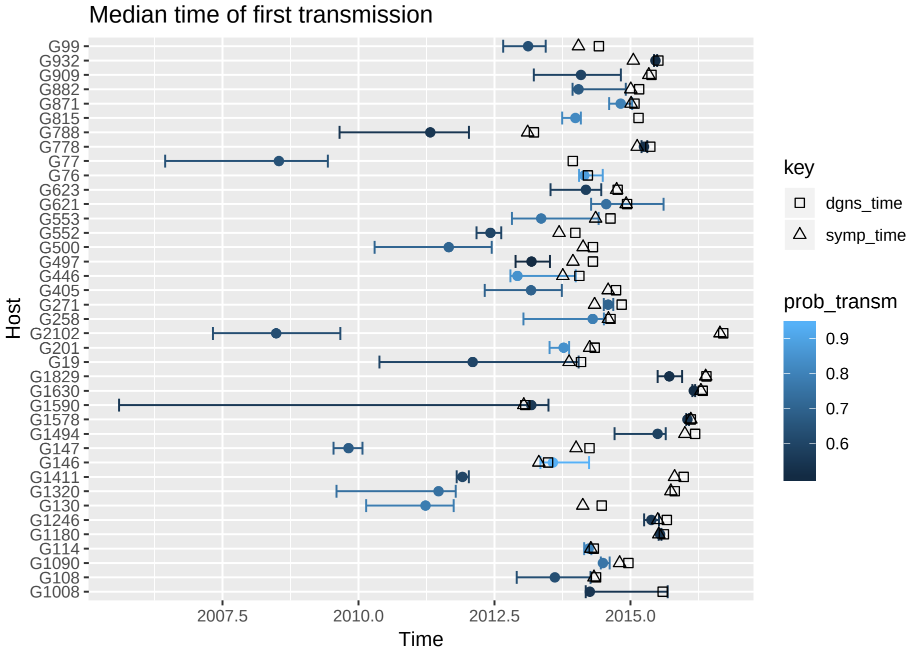


Figure 4.16. Resampled median time of first transmissions. The graph represents the median time of the first highly likely transmission for cases where the posterior probability of transmitting is greater than 0.7, under a clock rate of 0.544. For each case, the diagnosis time (square), and, where known, the symptom onset time (triangle) are added. Lighter colours indicate higher transmission probabilities. The range of the error bar indicates the 0.25 and 0.75 quantile.

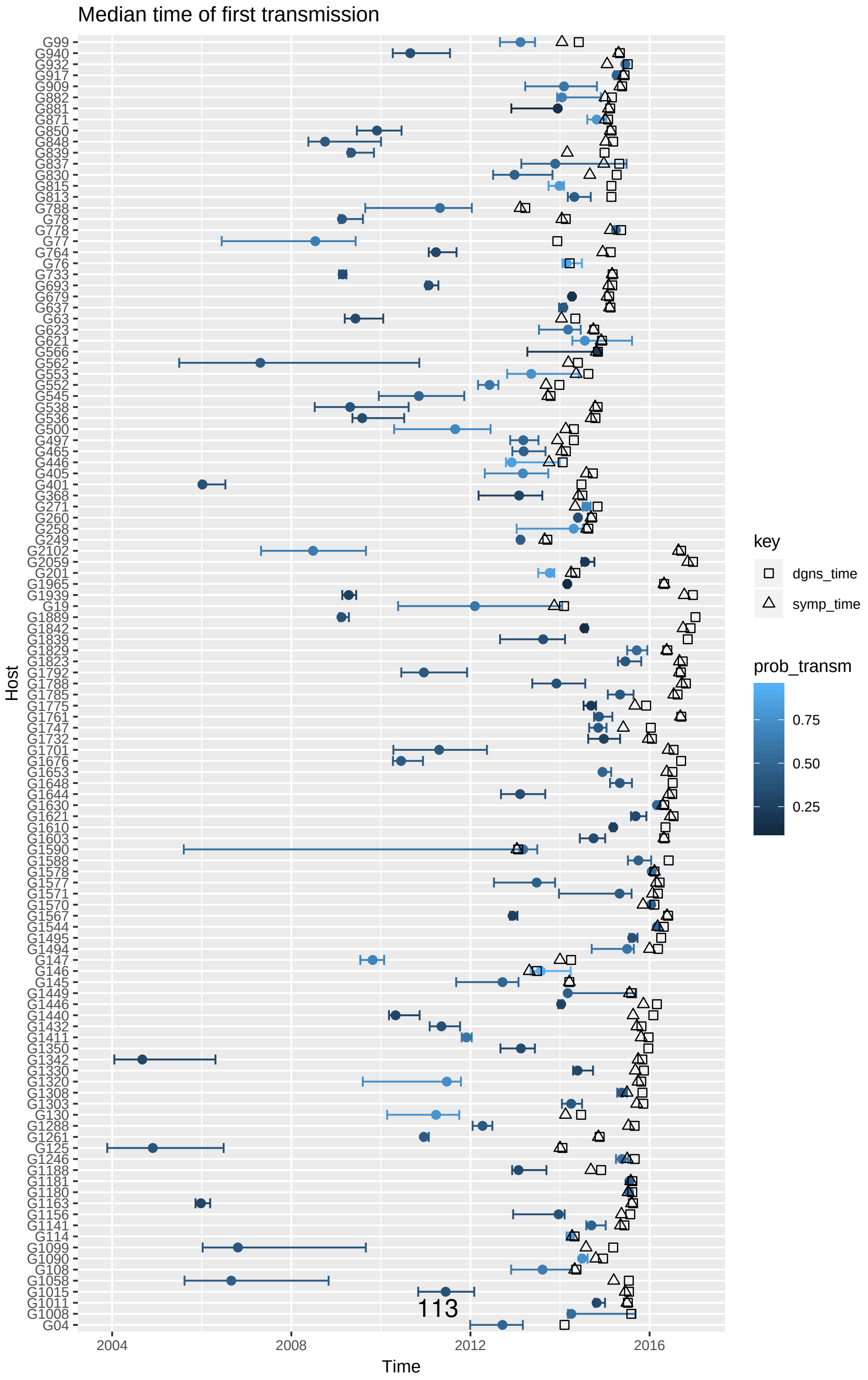


Figure 4.17. Resampled median time of first transmissions. The graph represents the median time of the first highly likely transmission for all cases, corresponding to a posterior probability of transmitting of 0.2, under a clock rate of 0.544.

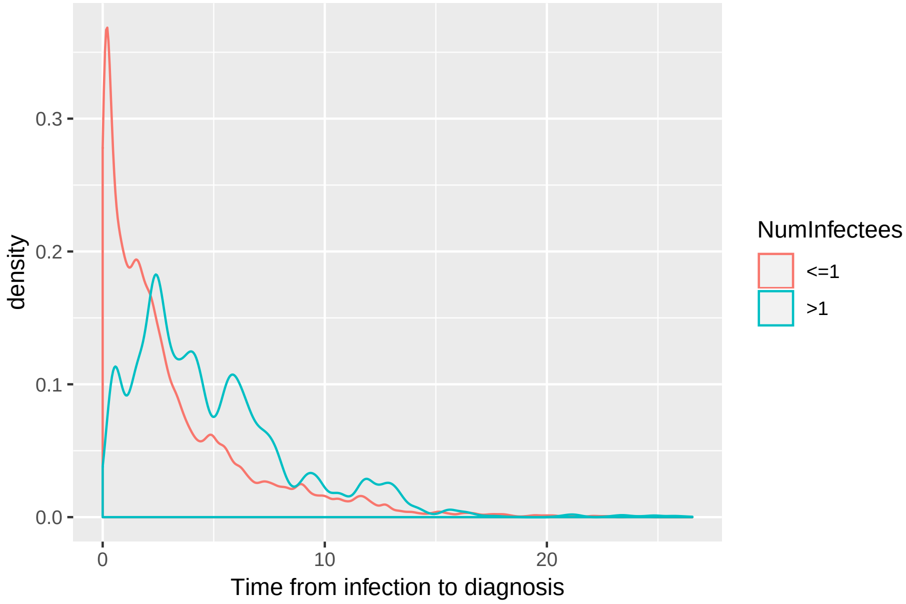


Figure 4.18. Density of times to diagnosis among those cases estimated to have caused more than one vs 0-1 secondary cases. Estimates were derived under a clock rate of 0.363 and are collected over all posterior transmission events. The means time since the infection of the transmitter to active disease of the secondary case (thus including latency periods) are 4.88 years (those infecting more than 1 secondary case) vs 3 years (those infecting 0 or 1 secondary cases).

Cryptic Resistance Mutations Associated With Misdiagnoses of Multidrug-Resistant Tuberculosis.

The work described in the present chapter has been published as a Brief Report Article in *Journal of Infectious Diseases*

Cancino-Muñoz I, Moreno-Molina M, Furió V, et al. Cryptic Resistance Mutations Associated With Misdiagnoses of Multidrug-Resistant Tuberculosis. The Journal of Infectious Diseases. 2019; 220(2):316–320. https://doi.org/10.1093/infdis/jiz104

5.1 Abstract

Understanding why some multidrug-resistant tuberculosis cases are not detected by rapid phenotypic and genotypic routine clinical tests is essential to improve diagnostics assays and advance toward personalized tuberculosis treatment. Here, we combine whole-genome sequencing with single-colony phenotyping to identify a multidrug-resistant that had infected a patient for 9 years. Our investigation revealed the failure of rapid testing and genome-based prediction tools to identify the multidrug-resistant strain. The false-negative findings were caused by uncommon rifampin and isoniazid resistance mutations. Although whole-genome sequencing data helped to personalize treatment, the patient developed extensively drug-resistant tuberculosis, highlighting the importance of coupling new diagnostic methods with appropriate treatment regimens.

Keywords: Tuberculosis, drug resistance, whole-genome sequencing, individualized treatment, cryptic mutations

5.2 Background

Personalized treatment in tuberculosis can be achieved in the next years if we are able to implement rapid, cost-effective and comprehensive drug susceptibility tests (DSTs). However, the prospects for this personalization deeply depend on our ability to identify drug resistance-associated mutations and to interpret their clinical role during management of the cases [220]. Current methods to identify and manage drug resistance are based on rapid liquid culture systems and/or molecular amplification tests [221]. Both approaches have limitations. For rifampin (RIF), some mutations, termed “disputed” mutations [84], are systematically missed by rapid automatic liquid culture methods, such as the Bactec-Mycobacteria Growth Indicator Tube (Bactec-MGIT) system, which is used in this study. These noncanonical RIF resistance mutations are involved in low-level resistance and associated with relapse [221, 222]. Likewise, the number of mutations screened by nucleic acid amplifications tests is limited [84]. Until now, no case has been reported in which both the Bactec-MGIT system and genotypic assays failed to identify multidrug-resistant strains. Here, we report a multidrug-resistant strain with cryptic mutations not detected by rapid routine clinical methods or whole-genome prediction tools. Prospective whole-genome sequencing (WGS) helped to track additional resistance mutations, but failure to provide an appropriate treatment regimen led to extensively drug-resistant tuberculosis. Despite limited therapeutic options, the patient was declared cured in 2018.

5.3 Methods

Clinical and microbiological data, together with a more-detailed description of the methods, are included in the **Supplementary Methods**.

5.3.1 Clinical case, Isolate Collection, and Routine DST Procedures

The study case was a Spanish-born patient with no common tuberculosis risk factors whose first episode of tuberculosis occurred in 2009. Findings of sputum and culture analyses became negative within 2 months after treatment initiation, and the patient was considered cured 4 months later, based on World Health Organization guidelines. However, relapse occurred in 2013 despite no risk factor for relapse during the initial episode [223]. Two years later, the patient was not responding to therapy despite compliance with treatment, close monitoring, and infection with a drug-sensitive, based on results of the hospital´s routine rapid phenotypic DST and genotypic testing. We analysed 16 serial clinical isolates recovered from the patient during 2009-2018 by the clinical microbiology unit of the Hospital Universitario General de Valencia (Valencia, Spain).

Routine phenotypic DST for the first-line antituberculosis drugs (and linezolid [LZD] in 1 isolate) was performed on all clinical samples collected during the study period, using the Bactec-MGIT 960 system (Becton Dickinson, Franklin Lakes, NJ). For second-line drugs, the Sensititre MycoTB MIC Plate (Trek Diagnostics System, Cleveland, OH) was used. Ranges of critical concentrations for all drugs are specified in the **Supplementary Methods**.

5.3.2 WGS Sequencing

Extracted DNA from diagnostic cultures was sequenced on the MiSeq platform, using standard procedures. We used Kraken [192] to identify reads belonging to the *Mycobacterium tuberculosis* complex. For mapping and variant calling, we used a previously described pipeline [127]. For details, see the **Supplementary Methods**.

5.3.3 Identification of Candidate Drug Resistance Variants

We identified candidate drug resistance variants by mapping them to known drug resistance-associated genes and confirming that they had not previously been described as phylogenetic markers (**Additional Table 2**). In addition, we screened any new variant that arose during the course of treatment in any part of the genome and reach a minimal frequency of 15% in $\leq$1 sample, to evaluate their potential role on drug resistance. Our in-house results were compared to data from 3 publically genomic resistance databases (accessed April 2018) [193, 103, 139]. The global frequencies of these candidate mutations were evaluated against an in-house database of 4762 genomes collected worldwide.

5.3.4 Isolation of Single Clones, Amplicon Sequencing, and Minimum Inhibitory Concentration (MIC) Validation for Resistance Mutations

After we identified mutations in genes or genomic regions associated with drug resistance to INH and RIF, we tested whether those variants conferred resistance, by characterizing a series of single-colony isolates. Twenty-two single clones with different genotypes, obtained from complex diagnostic cultures at different time points, were selected and isolated. Each clone was tested twice for susceptibility to INH, using the resazurin microtiter assay with 2-fold dilutions for 9 different concentrations (range, 0.06-32 μg/mL). We also confirmed that the I491F mutation conferred resistance to RIF, using the proportions method with 2-fold dilutions for 10 different concentrations (range, 0.06-64 μg/mL). Before DST, we performed ultra-deep amplicon sequencing of the regions of interest (rpoB, katG, and the ahpC promoter) to confirm the genotype of each clone, as well as to discount the presence of any unnoticed mutation with a frequency of $\leq$0.1%, the lover limit of detection (**Supplementary Methods and Supplementary Table 5.4**).

5.4 Results

5.4.1 Cryptic Variants Behind an Unnoticed Multidrug-Resistant Tuberculosis Case

A total of 16 isolates from the patient were available and sequenced during the study period (2009-2017). It is important to note that the first genome sequence was analyzed in 2015, after the patient had received standard first-line treatment for 2 years without a response (**Supplementary Figure 5.3**). Reconstruction of a phylogeny from WGS data strongly supported a scenario in which the relapse infection (which began in 2013) was caused by the strain from the first episode (which began in 2009; **Supplementary Figure 5.4**). Inspection of candidate variants only revealed likely mutations in known drug resistance genes (a complete list is shown in **Additional Table 3**). We found the *rpoB* mutation I491F, which is a noncanonical but known RIF resistance-associated variant. I491F is systematically missed by Bactec-MGIT system because of an unfortunate combination of slow mycobacterial growth and the system´s switch to an automated readout after 20 days.

WGS analysis of previous isolates revealed that I491F variant appeared to be fixed in the isolate initially cultured during the relapse episode (in September 2013) but not in the isolate from the first episode (in 2009). Thus, during the relapse episode, the mycobacteria were already resistant to RIF at the time the first positive culture result were obtained (**Figure** 5.1). Knowing this, we looked for INH resistance variants in the first isolate from the relapse episode, but we did not detect any putative mutation. However, in later isolates we identified 2 new noncanonical, mutually exclusive candidate INH resistance mutations, in *katG* gene (G249del and G273R). The G249del variant appeared as early as January 2014 but with highly variable frequency across samples, although it dominated the last mycobacteria-positive cultures (from June 2016 onward). In contrast, the G273R mutation appeared for the first time in March 2014, disappearing by December 2015. As expected for noncanonical *katG* mutations, screening of the ahpC gene and promoter region identified multiple ahpC promoter mutations whose presence fluctuated through time (**Figure** 5.1).

Switching treatment from a first-line regimen to a multidrug-resistant tuberculosis regimen is a major clinical decision. Thus, resistance variants detected by our genomic analyses needed validation. We selected 22 single clones from secondary cultures of specimens obtained at different time points during treatment and performed DST with alternative methods (**Figure** 5.2). Furthermore, we performed ultra-deep amplicon sequencing in specific RIF and INH resistance regions (**Supplementary Methods**). First, we confirmed that all 19 clones harboring the *rpoB* I491F mutation were RIF resistant (MIC, $>$ 1 μg/mL) as compared to the 3 clones from 2009 with no mutation (MIC, $<$ 1 μg/mL), which had an higher MIC than H37Rv but similar to that for other RIF-susceptible strains described elsewhere [87]. In the case of INH resistance, clones from 2009 and 2013 had a low MIC for INH ($<$ 0.25 μg/mL), consistent with the fact that no putative *katG* mutations were found in these isolates. In contrast, all clones from 2014 had either the *katG* G273R or the G249del mutation fixed and no other alternative candidate mutation at a frequency of $\geq$0.1%, as revealed by amplicon sequencing (**Figure** 5.2). All of these clones were highly resistant to INH (MIC, $>$ 32 μg/mL) based on the resazurin microtiter assay. In addition, clone haplotype analysis established a link between specific *ahpC* mutations and the 2 specific *katG* variants (**Figure 5.2**).

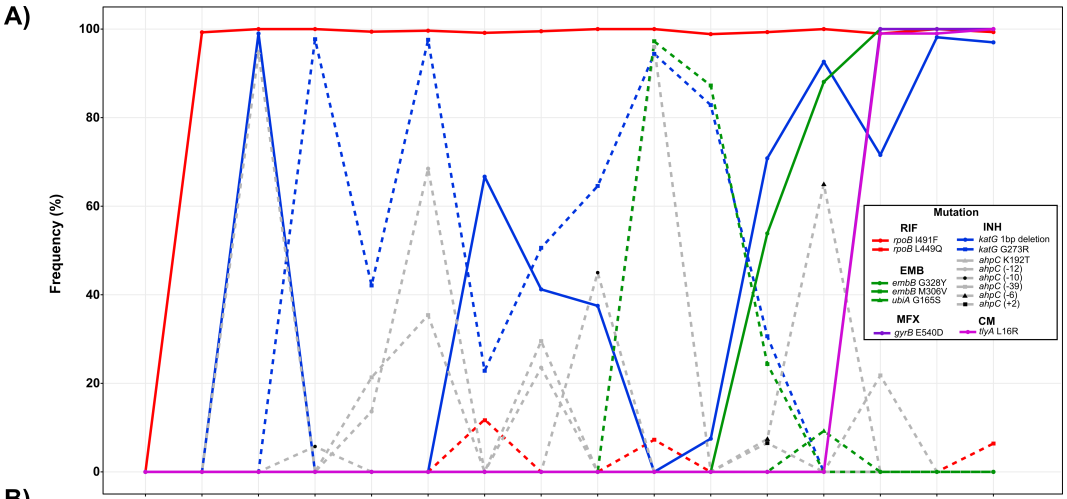
Thus, multiple lines of evidence suggested that the 2 *katG* mutations were likely involved in INH resistance: (1) genomic analyses identified the variants as mutually exclusive, suggesting selection for different resistant populations; (2) a single-clone phenotypic assay identified that these mutations were associated with high-level INH resistance; and (3) *ahpC* mutations were associated with noncanonical *katG* mutations (**Supplementary Table 5.1**). Additionally, genomic analysis detected 3 different ethambutol-associated mutations, beginning April 2015, including 2 in the *embB* genomic region (G328Y, M306V) and 1 in the *ubiA* region (G165S; **Figure 5.1**).

5.4.2 Prospective case management aided by WGS data

After validation of RIF and INH resistance, we used WGS as a primary tool for detecting resistance. Given the newly discovered multidrug-resistant tuberculosis profile of the patient and their lack of response to treatment, the clinical team decided to change the drug regimen in December 2015. Moxifloxacin (MFX) and capreomycin were added, and RIF removed. Despite the patient´s adherence to the new treatment regimen, sputum smear results were positive in June 2016, followed by another positive culture in October 2016. Rapid sequencing of this isolate revealed that it had acquired a related MFX resistance mutation (E540D, in *gyrB*) and a likely capreomycin (L16R, in *tlyA*). In parallel, a microdilution-based assay (the Sensititre MycoTB MIC Plate) confirmed resistance to MFX (MIC, $<$ 4 μg/mL) but revealed amikacin susceptibility (MIC, $\leq$ 1 μg/ml). Accordingly, drug therapy was adjusted, removing MFX and adding linezolid and amikacin. Unfortunately, bedaquiline and delamanid were not available to the hospital. With this new treatment, the last positive culture result was in February 2017, and after 18 months of culture negativity, the patient made a satisfactory recovery. All resistance variants detected are in **Supplementary Table 5.2**.

Notably, none of the publically available genomic resistance prediction databases classified any of the isolates as a multidrug-resistant or extensively drug-resistant strain (**Supplementary Table 5.3**). In agreement with this, an extensive analysis of 4762 genomes revealed that strains with noncanonical RIF resistance mutations were depleted of known *katG* resistance mutations (7.6% vs 36%; P $<$ .001, by the x2 test). A deeper analyses of *katG* in those strains revealed 7 mutations not described before, all of them leading to an amino acid change and some with phylogenetic convergence signal (**Supplementary Data**).

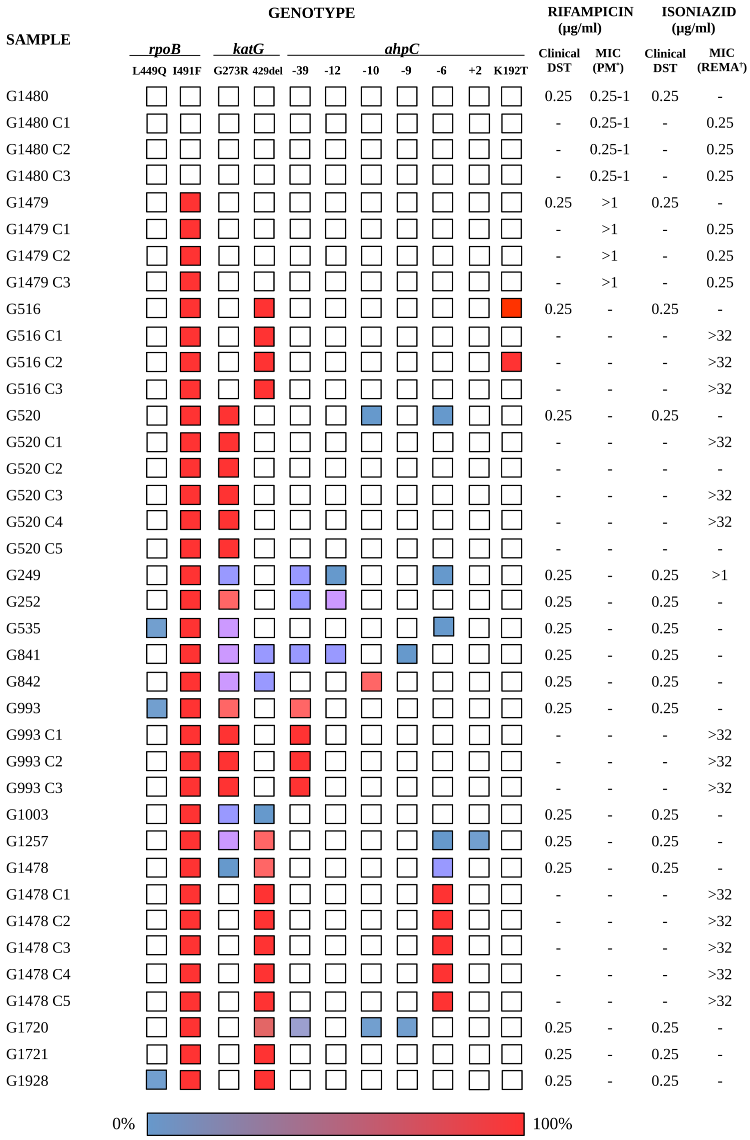
A)



B)

Sample	G1480	G1479	G516	G520	G249	G252	G535	G841	G842	G993	G1003	G1257	G1478	G1720	G1721	G1928
Date	27/04/09	20/09/13	21/01/14	17/03/14	27/06/14	11/08/14	13/10/14	09/12/14	27/01/15	15/04/15	11/06/15	09/11/15	11/12/15	02/06/16	14/10/16	09/01/17
Sputum/ Culture	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+
Hospital DST (RIF/INH)*	-/-	-/-	-/-	-/-	-/+	-/-	-/-	-/-	-/-	-/-	-/-	-/+	-/+	-/-	-/+	-/+
Genomic resistance (RIF/INH)‡	-/-	+/-	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+
Treatment	RIF,INH, EMB,PZA	RIF,INH, EMB,PZA	RIF,INH, EMB,PZA	RIF,INH, EMB,PZA	RIF,INH, EMB,PZA	RIF,INH, EMB,PZA	RIF,INH, EMB,PZA	RIF,INH, EMB,PZA	RIF,INH, EMB,PZA	RIF,INH, EMB,PZA	RIF,INH, EMB,PZA	RIF,INH, EMB,PZA	EMB,MFX, CM,PZA,INH^{+}	AMK, PZA, LZD, INH^{+}	AMK, PZA, LZD, INH^{+}	AMK, PZA, LZD, INH^{+}
Resistance Type§	DS	RR	MDR	MDR	MDR	MDR	MDR	MDR	MDR	MDR	MDR	MDR	MDR	XDR	XDR	XDR

Figure 5.1: A, All drug resistance mutations detected during the treatment. Different colors indicate drug resistance variants associated with different genes. Solid lines denote mutations at a frequency of 100% in the patient´s last mycobacteria-positive culture, whereas dashed lines indicate transitory variants over time. B, Relevant clinical findings of the case study, including antibiotic therapy, sputum and culture status, DST results in the hospital, and predicted WGS-based resistance profile. Plus signs denote that higher doses of isoniazid were added. AMK, amikacin; CM, capreomycin; DS, drug susceptible; DST, drug susceptibility testing; EMB, ethambutol; INH, isoniazid; LZD, linezolid; MDR, multidrug-resistant; MFX, moxifloxacin; PZA, pyrazinamide; RIF, rifampicin; RR rifampicin resistant; XDR, extensively-drug resistant. aPhenotypic analysis was performed using the Bactec-MGIT 960 system, and genotypic analysis was performed using the GenoType MTBDR plus system. bPredicted DST-based resistance, based on mutations detected in genomic regions associated with RIF and INH resistance. cPredicted whole-genome sequencing-based drug susceptibility profile.



SAMPLE	GENOTYPE											RIFAMPICIN (µg/ml)		ISONIAZID (µg/ml)	
	rpoB		*katG*		*ahpC*										
	L449Q	I491F	G273R	429del	-39	-12	-10	-9	-6	+2	K192T	Clinical DST	MIC (PM*)	Clinical DST	MIC (REMA†)
G1480	□	□	□	□	□	□	□	□	□	□	□	0.25	0.25-1	0.25	-
G1480 C1	□	□	□	□	□	□	□	□	□	□	□	-	0.25-1	-	0.25
G1480 C2	□	□	□	□	□	□	□	□	□	□	□	-	0.25-1	-	0.25
G1480 C3	□	□	□	□	□	□	□	□	□	□	□	-	0.25-1	-	0.25
G1479	□	🟥	□	□	□	□	□	□	□	□	□	0.25	>1	0.25	-
G1479 C1	□	🟥	□	□	□	□	□	□	□	□	□	-	>1	-	0.25
G1479 C2	□	🟥	□	□	□	□	□	□	□	□	□	-	>1	-	0.25
G1479 C3	□	🟥	□	□	□	□	□	□	□	□	□	-	>1	-	0.25
G516	□	🟥	□	🟥	□	□	□	□	□	□	🟥	0.25	-	0.25	-
G516 C1	□	🟥	□	🟥	□	□	□	□	□	□	□	-	-	-	>32
G516 C2	□	🟥	□	🟥	□	□	□	□	□	□	🟥	-	-	-	>32
G516 C3	□	🟥	□	🟥	□	□	□	□	□	□	□	-	-	-	>32
G520	□	🟥	🟥	□	□	□	🟦	□	🟦	□	□	0.25	-	0.25	-
G520 C1	□	🟥	🟥	□	□	□	□	□	□	□	□	-	-	-	>32
G520 C2	□	🟥	🟥	□	□	□	□	□	□	□	□	-	-	-	-
G520 C3	□	🟥	🟥	□	□	□	□	□	□	□	□	-	-	-	>32
G520 C4	□	🟥	🟥	□	□	□	□	□	□	□	□	-	-	-	>32
G520 C5	□	🟥	🟥	□	□	□	□	□	□	□	□	-	-	-	-
G249	□	🟥	🟪	□	🟪	🟦	□	□	🟦	□	□	0.25	-	0.25	>1
G252	□	🟥	🟥	□	🟪	🟪	□	□	□	□	□	0.25	-	0.25	-
G535	🟦	🟥	🟪	□	□	□	□	□	🟦	□	□	0.25	-	0.25	-
G841	□	🟥	🟪	🟪	🟪	🟪	□	🟦	□	□	□	0.25	-	0.25	-
G842	□	🟥	🟪	🟪	□	□	🟥	□	□	□	□	0.25	-	0.25	-
G993	🟦	🟥	🟥	□	🟥	□	□	□	□	□	□	0.25	-	0.25	-
G993 C1	□	🟥	🟥	□	🟥	□	□	□	□	□	□	-	-	-	>32
G993 C2	□	🟥	🟥	□	🟥	□	□	□	□	□	□	-	-	-	>32
G993 C3	□	🟥	🟥	□	🟥	□	□	□	□	□	□	-	-	-	>32
G1003	□	🟥	🟪	🟦	□	□	□	□	□	□	□	0.25	-	0.25	-
G1257	□	🟥	🟪	🟥	□	□	□	□	🟦	🟦	□	0.25	-	0.25	-
G1478	□	🟥	🟦	🟥	□	□	□	□	🟪	□	□	0.25	-	0.25	-
G1478 C1	□	🟥	□	🟥	□	□	□	□	🟥	□	□	-	-	-	>32
G1478 C2	□	🟥	□	🟥	□	□	□	□	🟥	□	□	-	-	-	>32
G1478 C3	□	🟥	□	🟥	□	□	□	□	🟥	□	□	-	-	-	>32
G1478 C4	□	🟥	□	🟥	□	□	□	□	🟥	□	□	-	-	-	>32
G1478 C5	□	🟥	□	🟥	□	□	□	□	🟥	□	□	-	-	-	>32
G1720	□	🟥	□	🟥	🟪	□	🟦	🟦	□	□	□	0.25	-	0.25	-
G1721	□	🟥	□	🟥	□	□	□	□	□	□	□	0.25	-	0.25	-
G1928	🟦	🟥	□	🟥	□	□	□	□	□	□	□	0.25	-	0.25	-

Figure 5.2: Percentages are given for the frequency among diagnostic cultures, as well as among individual isolated clones (identified with a "C"). DST, drug susceptibility testing; MIC, minimum inhibitory concentration; PM, proportions methods; REMA, resazurin microtiter assay. a;Phenotypic DST results for RIF and INH from individual clones. b;Phenotypic DST results for RIF and INH from clinical isolates.

5.5 Discussion

Here we described the use of WGS data to diagnose a case of multidrug-resistant tuberculosis that was missed by the commonly used Bactec-MGIT system. An uncommon RIF resistance mutation (I491F, *rpoB*) led to a systematic negative test result. Notably, this outcome affected INH DST with the Bactec-MGIT system; in contrast, our investigation clearly demonstrated the presence of high-level INH resistance at different time points. The resistance profile undetected by the Bactec-MGIT system before genome data were available explains why the patient remained culture positive and the infecting mycobacteria acquired additional resistance mutations between 2013 and 2015. In the absence of a fully reliable Bactec-MGIT result, we decided to use WGS data to aid in the clinical management of the case.

However, clinical decisions based on WGS are not straightforward. The higher resolution of next-generation sequencing approaches, combined with our increasing knowledge of drug resistance-associated mutations, provide evidence for the usefulness of designing individualized drug regimens [224]. Nevertheless, this work also reveals additional layers of complexity in the clinical decision making; for example, INH-susceptible subpopulations were still present after 2.5 years of treatment (**Supplementary Figure 5.5**). These results suggest that personalized treatment will require serial sequencing over time, preferably instead of sputum culture, to avoid culture bias and track the dynamics of susceptible and resistant subpopulations. Furthermore, rigorous standardized statistical approaches such as those developed by Miotto et al [97] should identify high likely drug resistance mutations, to avoid false-positive predictions and adverse downstream clinical consequences. In this particular case, WGS aided care management, but despite access to WGS data, treatment decisions led to the development of extensively drug-resistant tuberculosis.

The poor treatment outcome in this patient is in line with previous reports that RIF monoresistance is associated with relapse and with the acquisition of additional resistance mutations [221, 225]. Furthermore, it is also noteworthy that most of the variants described are epidemiologically rare and that none of the canonical mutations were found. Uncommon drug resistance-conferring mutations are likely more common in high-burden countries [226], and, thus personalized treatment approaches based on WGS data in those countries may be compromised. Evidence from this patient adds to the view that we need to integrate different layers of heterogeneity to understand the emergence of and predict drug resistance in a patient. Those layers include strain, lesion, pharmacodymanics, and drug penetration heterogeneity.

5.5.1 Acknowledgments

We thank Dr. Ana Gil-Brusola (Hospital Universitario y Politécnico de La Fe, Valencia, Spain) for her corrections and feedback during the manuscript preparation.

The work described in the present chapter has been published as a Brief Report Article in *Journal of Infectious Diseases*

Cancino-Muñoz I, Moreno-Molina M, Furió V, et al. Cryptic Resistance Mutations Associated With Misdiagnoses of Multidrug-Resistant Tuberculosis. The Journal of Infectious Diseases. 2019; 220(2):316–320. https://doi.org/10.1093/infdis/jiz104

5.6 Supplementary Data

5.6.1 Supplementary Methods

Case study

A 43-year old male patient with no tuberculosis (TB) clinically records was admitted to the Hospital Universitario General de Valencia, Spain in April of 2009. The patient did not present any comorbility or risk factor that supposed a bad therapeutic compliance (e.g HIV negative status). The patient was first diagnosed with pulmonary TB and treated with the standard first-line therapy for active TB (Two months with rifampicin [RIF], isoniazid [INH], ethambutol [EMB] and pyrazinamide [PZA], followed by four months of RIF and INH only), adjusting for body weight. During this period, serial sputum-positive samples was collected and culture in order to 1) confirm the presence of *Mycobacterium tuberculosis* Complex (MTBC) bacilli and 2) to performed drug susceptibility testing (DST). MTB identification was carried out using the commercial kit BACTEC MGIT TB Identification Test (Becton Dickinson and Co, Franklin Lakes, New Jersey). First-line DST was performed using rapid phenotypic BACTEC MGIT 960 system (Becton Dickinson and Co, Franklin Lakes, New Jersey) and genotypic probes (GenoType MTBDR plus, Hain Lifescience, Nehren, Germany). DST results indicated that the isolate was fully-susceptible to all first-line antibiotics. After two months of treatment, the patient became sputum smear and culture negative. Consecutive negative samples for next four months (n=4) confirmed that the patient was cured. During 2010 and 2012 the patient presented negative sputum smear and was discharged.

In September 2013, the patient was readmitted to the hospital because he presented typical TB symptoms suggesting a relapse episode. MTBC was detected in sputum smear samples and confirmed in by rapid culture-based immunoassay (BACTEC MGIT TBc Identification Test). Relapse disease has been related with the infection by an MDR TB strain. So, phenotypic and genoptypic DST was performed. Again, phenotypic DST and line-probe assays showed no evidence of drug resistance. With this scenario, the patient was treated with the same first-line antituberculous therapy maintained all effective drugs. Six months later (March 2014) the patient became sputum and culture-negative and once again, was considered cured. Unfortunately, in June 2014, a single sputum specimen was detected as an INH mono-resistant isolate by phenotypic BACTEC MGIT 960 system but no mutation was detected by genotypic Hain line-probes assay. Because the infection strains was considered INH mono-resistant and there were more four effective antituberculous agents, the drug remigen was prolongated another nine months. Notably, during this time, phenotypic DST from all the clinical culture-positive samples tested negative for resistance to first-line drugs. Due to the first episode and the continuum culture-positive sputum status during the relapse, the therapy was extended until December 2015. As a consequence of the long and ineffective TB regimen, the patient developed a great cavity in the right upper lung.

In December 2015, whole genome sequencing (WGS) was performed for a culture derived from a positive sputum sample of the patient isolated in June of 2014. WGS analysis identified mutations in the *rpoB* and *katG* genes conferring RIF and INH resistance (*rpoB* I491F and *katG* G273R, respectively), which indicated an infection caused by probable a multidrug resistant (MDR) strain. We corroborated and validated RIF and INH resistance by extensive phenotypic DST methods (Proportions method in the case of the RIF and REMA assay for INH, see below for a detailed description). With this result, the antituberculois therapy was changed, and second-line drugs were added following the recommendations by the WHO guidelines[1]. This regimen included the administration of EMB, Moxifloxacin (MFX), PZA, higher doses of INH (INH+) and the injectable agent Capreomycin (CP). We decided to use CP instead of other aminoglycoside because of the greater ease of intramuscular administration. From this point, WGS was used in retrospective manner to genomic characterirzed all the stored clinical samples from the patient, since the first TB episode in 2009. In addition, WGS was used prospective manner together with hospital phenotypic DST results to guide the antibiotic regimen in the patient.

On June 2016, after six months of the MDR treatment, a new sputum smear and culture-positive appeared. WGS analysis of this sample revealed the presence of mutations conferring EMB, MFX, and high likely CP drug resistance (*embB* G238Y, *gyrB* E540D and *tlyA* L16R, respectively), indicating that TB infecting isolate had evolved to an extensively drug-resistant (XDR) strain. Phenotypic DST performed in the hospital (Sensititre MYCOTB MIC Plate, Trek Diagnostics System, Cleveland, OH) confirmed MFX drug resistance (MIC 4 μg/ml) and susceptibility to amikacin (MIC $\leq$1µg/ml). Based on these findings and following the WHO recommendations [227], we decided to change the drug treatment in the same month adding the second-line amikacin and linezolid agents, INH+ and PZA. In addition, surgical intervention was proposed but there were several cavited foci in both lungs and the patient rejected this option. The last culture-positive sputum was detected in February 2017. A year later, amikacin was removed from the therapy due to the patient had suffered adverse effects such as hearing and sigth loss. It is notable that PZA resistance was no detected by phenotypic (BACTEC MGIT 960 system) and genotypic (WGS analysis) methods during the nine years of the disease.

Isolate collection and DNA extraction

For this study, we used sixteen serial clinical isolates from a single patient and one epidemiologically related sample, all supplied by the Hospital General de Valencia. The single-patient samples covered a total of 9 years and included a first episode of active TB and a relapse incident years later (Supplementary Table 1). Positive mycobacterial growth indicator tubes (MGIT) were subcultured in Middlebrook 7H11 agar plates supplemented with 10% OADC (Becton-Dickinson) for 5 weeks at 37℃. We scrapped bacteria with a sterile loop four times across the plate to obtain a representative sample of the population and we extracted their DNA using the CTAB protocol [228]. The rest of the bacteria was also scrapped and stored in 1ml of glycerol (20%) at -80 C°. Because of the high risk to manage and work with XDR strains, isolates G1257, G1720 and G1928 were directly extracted from the positive MGIT culture.

Routine microbiological diagnostics methods

Standard phenotypic DST for the first-line anti tuberculosis drugs (and linezolid [LZD] in one isolate) was performed in all clinical samples during whole study period using the BACTEC MGIT 960 system (Becton Dickinson and Co, Franklin Lakes, New Jersey) following the manufacturer recommended critical concentrations for each drug; RIF (1.0μg/ml), INH (0.1μg/ml), ethambutol (EMB) (5.0μg/ml), pyrazinamide (PZA) (100μg/ml), streptomycin STR (1.0μg/ml) and LZD (1.0μg/ml).In addition, genotypic line-probes assays for RIF and INH resistance was carried out with the commercial kit GenoType MTBDR plus (Hain Lifescience, Nehren, Germany). Finally, Sensititre MYCOTB MIC Plate (Trek Diagnostics System, Cleveland, OH)) was applied within a few clinical samples to determine the minimal inhibitory concentrations (MIC) for some second-line antibiotics such as moxifloxacin (MFX), ethionamide (ETH) and amikacin (AMK). The following range of concentrations were used; MFX, (0.06-8μg/ml); ETH, (0.3-40μg/ml); AMK, (0.12-16μg/ml).

Whole genome sequencing analysis

DNA from diagnostic cultures was extracted as previously described [228]. Sequencing libraries were constructed with Nextera XT DNA Library Preparation Kit (Illumina, San Diego, CA) following the manufacturer instructions. WGS was performed on the Illumina MiSeq instrument and the average sequencing depth value per base was 168-fold (range: 103-268). To account for contaminant DNA, we identified reads only belonging to the *Mycobacterium tuberculosis* Complex (MTBC) using KRAKEN software [192]. Mapping and single nucleotide polymorphism (SNP) calling was performed following a previous pipeline [127]. Briefly, MTBC reads were mapped and aligned to an inferred MTBC most likely common ancestor genome using BWA [229]. Next, we separated variants into INDELS (small insertions and deletions included) and SNPs. Single polymorphisms with at least 10 reads in both strands and a quality score of 20 were selected and classified into two categories based on their frequency in the sample. We considered fixed SNPs those with no less than 90% of frequency and low-frequency SNPs those involving variants whose frequencies range between 10 and 89%. An INDEL was considered whether the mutation was present with a minimum deep coverage of 10x. SNP annotation was performed using H37Rv annotation reference (AL123456.2). Finally, SNPs annotated in regions difficult to map such as repetitive sequences and PPE/PE-PGRS genes were removed from the analysis as well as those detected near INDELS.

Identification of relapse versus re-infection strains

In order to identify whether the isolates from the first and second disease episode were the same (relapse) or coming from two independent infections (re-infection) we applied both a phylogenetic and a SNP threshold approach. Regarding the SNP threshold approach, pairwise genetic distances were compared based on a concatenated SNP alignment obtained from every fixed SNP from every isolate. A genetic distance below 12 SNPs is indicative of clonal diversification from the first episode strain and thus is classified as a relapse case. On the contrary distances beyond 12 SNPs are indicative of exogenous infection with a different strain and thus the case is classified as likely re-infection [150, 230]. To differentiate phylogenetically between relapse and re-infection we built a phylogeny including all the isolates from the first and second episode as well as epidemiological related sample (brother). The phylogeny was inferred following maximum likelihood phylogenetic approach using RAxML v8.2 [182] applying the General Time Reversible model of nucleotide substitution with the Gamma distribution (GTRGAMMA).

Identification of known lineage and drug resistance associated mutations

Once we had determined all the polymorphic sites (including fixed and variable SNPs), we compared our SNP and INDEL data with a series of publicly available predictive databases (Phyresse [193], Mykrobe predictor [103], TB-profiler [139], latest accessed on April 2018) to determine lineage and drug resistance mutations present in our samples. Isolates were classified as susceptible, resistant or multidrug resistant (MDR) according to the mutations identified.

Identification of candidate drug resistance variants

Relevant variants during the course of treatment were selected if they appeared at least in one isolate with a minimum frequency of 15% and had at least 15% of difference between any two isolates (n=96, see Supplementary Table 2). Global frequency of the new candidate mutations was evaluated against a database of 4,762 genomes collected world-wide. The same database was used to identify strains with and without "disputed" *rpoB* mutations as well as new drug resistance determinants to INH in *katG* gene.

Isolation of single colonies and drug susceptibility testing of new drug resistance mutations

After the identification of a series of novel mutations in genes or genomic regions associated with drug resistance we decided to explicitly test if those mutations conferred resistance by phenotypically and genetically characterizing a series of single-colony isolates. Twenty-two single colonies from six clinical samples with different co-existing haplotypes were isolated from Middlebrook 7H11 agar plates, grown for 10 to 15 days in Middlebrook 7H9 at 37C and stored in 500μl aliquotes with 20% glycerol at -80C° until used.

Antibiotic stock solutions were prepared at 10 mg/ml either in sterile water (INH) or methanol (RIF, both from Sigma-Aldrich) and stored at -20ºC. Nine days before the experiments, 500μl of frozen inoculum from each of the 22 colonies were cultured in 10ml of Middlebrook 7H9 broth supplemented with 10% ADC (Becton-Dickinson) and 0.1% tween80. Exponentially-growing bacteria were adjusted to an optical density of 0.2 (which is equivalent to a 1 McFarcland turbidity standard) in tween-free Middlebrook 7H9 broth immediately prior to use.

We obtained the MIC for INH following the resazurin microtiter assay (REMA) protocol [231]. Briefly, serial two-fold dilutions ranging from 0.125 to 32μg/ml of

INH in tween-free 7H9 broth were prepared in a 96-well plate, with a volume of 100μl. We then added 100ul of a 1:100 dilution of a fresh density adjusted bacterial suspension, prepared as above and sealed the plate in a hermetic plastic bag. Because of the slow growth of the samples carried the *rpob* I491F mutations, after 9 days of incubation at 37℃, 30μl of 0.02% resazurin was added to each well and the plate was further incubated at 37ºC for 48 h. A change of color from blue to pink indicates bacterial growth. The MIC was defined as the lowest drug concentration that prevented this color change. Drug and bacteria free wells were used as control.

To determine phenotypic resistance to RIF, we used the proportion method (PM) in Middlebrook 7H11 agar according to standard procedures [81]. The reason for using this particular method is that these strains grow very slowly in RIF-containing broth and give inconsistent results with the REMA. In brief, 50μl of a 10-2 and a 10-4 dilutions of a fresh density-adjusted bacterial suspension were cultured in plates containing 0.125 to 64μg/ml of RIF. Drug-free plates were used as a positive control. After 3 weeks of incubation at 37℃, the number of colonies forming units growing on medium with antibiotic were compared with those on the positive control. The proportion of resistant bacteria was represented as percentage. The MIC was defined as the first drug concentration where there was no growth.

The H37Rv strain was used as a quality control in all experiments. In addition, the isolate from the first episode that is wild-type (G1480) for the mutations was used to measure the increase on MIC in the following isolates. Strains were classified as susceptible or resistant according to the critical concentrations recommended by WHO (guidelines 2014 [227]).

Deep amplicon sequencing: amplification, library construction and sequencing

In order to confirm the SNPs identified and the possible detection of additional very low-frequency variants (¡5%), we performed deep amplicon sequencing following and adjusting the single molecule-overlapping reads (SMOR) approach [232]. Six primer sets were designed to target specific regions of *rpoB*, *katG* and *ahpC* (promoter and CDS included) (see supplementary Table 5), producing amplicons between 285-339bp. We used two additional primers sets targeting phylogenetic lineage-diagnostic SNPs as internal controls. A total of 39 samples were analysed including 15 clinical isolates from the case, 22 from single colonies and two controls (lineage 4 H37Ra and a Lineage 2 Beijing strain). A single PCR amplification step was carried out with parameters: initial denaturation at 95ºC for 3min, 20 cycles of denaturation at 98ºC for 15 sec, annealing at 65ºC for 15s and extension at 72ºC for 30s, with a final extension at 72ºC for 2min. Each reaction contained 12.5μl of 2x KAPA HiFi HotStart ReadyMix (KAPA biosystems), 0.75μl of 10μM forward primer, 0.75 of 10μM reverse primer, 5μl of template DNA and 6μl of PCR-grade water. After amplification, reactions were pooled by sample (5μl of each amplicon and 20μl of 10μM Tris-HCl) and purified using 1X NucleoMag NGS Clean-up and Size Select (Macherey-Nagel) up to 50μl final volume.

Amplicon sequencing libraries construction and sequencing were performed as described for whole genome sequencing. Purified libraries were validated on a Bioanalyzer DNA chip (Agilent Technologies) to verify fragment size, and quantified using Qubit 3.0 Fluorometer (Thermo Fisher Scientific). The expected average coverage for this experiment was 50,000 fold per base.

Data availability

All genomic data are deposited in the European Nucleotide Archive under the Bioproject numbers PRJEB22237 and PRJEB25887.

5.6.2 Supplementary Results

Non-canonical resistance mutations can lead to under-reporting and treatment of MDR-TB cases

The fact that INH resistance went undetected (likely due to the fitness cost of the RIF mutation on replication [84] led us to question whether non-canonical (also known as “disputed”) mutations are involved in a systematic under detection of MDR-TB. We pooled together a global SNP database of 4762 strains. A total of 66 strains harbored a non-canonical RIF resistance mutation (7.6% of the strains with a RIF resistance mutation). Of these, only 24 isolates contained a known katG mutation as opposed to 92.4% in strains carrying undisputed RIF resistance mutations (36%, $P<0.001$, chi-square test). Some of the strains could be RIF mono-resistant, particularly related with relapse cases [233, 234]. However, we hypothesized that a percentage of these strains were undetected MDR cases associated to non-canonical mutations. Among strains with a "non-canonical" *rpoB* mutation we found seven mutations not described before, all of them leading to an aminoacid change. For three of them we had evidence of convergent evolution, a strong predictor of resistance particularly in drug resistance associated regions [235, 236] [19,20]. One, *katG* V1A have been recently identified as INH associated with resistance [237]. For the other two (V473L and G285V), publicly available phenotypic results are contradictory as the majority of the isolates were susceptible. However, *rpoB* I491F mutation was present in four out of the six strains indicating a possible parallelism with the phenotyping problems presented in this manuscript.

5.6.3 Supplementary Tables

Isolate	Date	Resistance profile	Genetic changes associated to resistance mutations							
			rpoB	*katG*	*oxyR-ahpC*	*ahpC*	*ubiA*	*embB*	*gyrB*	*tlyA*
G1480	27/04/09	DS	WT	WT	WT	WT	WT	WT	WT	WT
G1479	20/09/13	RR	I491F	WT	WT	WT	WT	WT	WT	WT
G516	21/01/14	MDR	I491F	P429del	WT	K192T	WT	WT	WT	WT
G520	17/03/14	MDR	I491F	G273R	G-6A C-10T	WT	WT	WT	WT	WT
G249	27/06/14	MDR	I491F	G273R	G-6A C-12T C-39T	WT	WT	WT	WT	WT
G252	11/08/14	MDR	I491F	G273R	C-12T C-39T	WT	WT	WT	WT	WT
G535	13/10/14	MDR	I491F L449Q	P429del G273R	WT	WT	WT	WT	WT	WT
G841	09/12/14	MDR	I491F	P429del G273R	G-9A C-12T C-39T	WT	WT	WT	WT	WT
G842	27/01/15	MDR	I491F	P429del G273R	C-10T	WT	WT	WT	WT	WT
G993	15/04/15	MDR	I491F L449Q	G273R	C-39T	WT	G165S	WT	WT	WT
G1003	11/06/15	MDR	I491F	P429del G273R	WT	WT	G165S	WT	WT	WT
G1257	09/11/15	MDR	I491F	P429del G273R	G-6A	WT	G165S	D328Y	WT	WT
G1478	11/12/15	MDR	I491F	P429del	G-6A	WT	WT	D328Y M306V	WT	WT
G1720	02/06/16	XDR	I491F	P429del	G-9A C-10T C-39T	WT	WT	D328Y	E540D	L16R
G1721	14/10/16	XDR	I491F	P429del	WT	WT	WT	D328Y	E540D	L16R
G1928	09/01/17	XDR	I491F	P429del	C-39T	WT	WT	D328Y	E540D	L16R

Table 5.1: **Isolate sequencing results and main drug resistance associated mutations. Abbreviations: DS, drug susceptible; MDR, multidrug resistance; MTBC, *Mycobacterium tuberculosis* Complex; XDR, extensively drug-resistant.**

Drug associated	Genomic position	Wild type allele	Mutant allele	Gene alias	Gene name	Gene type	Mutation type	Nucleotide change	Aminoacid change
Moxifloxacin	6742	A	C	gyrB	Rv0005	essential	nonsynonymous	A1620C	E540D
Rifampicin	761152	T	A	rpoB	Rv0667	essential	nonsynonymous	A1346T	L449Q
Rifampicin	761277	A	T	rpoB	Rv0667	essential	nonsynonymous	A1471T	I491F
Capreomycin	1917986	T	G	tlyA	Rv1694	nonessential	nonsynonymous	T47G	L16R
Isoniazid	2154827	C	-	katG	Rv1908c	nonessential	nonsynonymous	C1285∆	G429del
Isoniazid	2155295	C	G	katG	Rv1908c	nonessential	nonsynonymous	G817C	G273R
Isoniazid	2726112	C	T			ahpC	promoter mutation		
Isoniazid	2726139	C	T			ahpC	promoter mutation		
Isoniazid	2726141	C	A/T			ahpC	promoter mutation		
Isoniazid	2726145	G	A			ahpC	promoter mutation		
Isoniazid	2726153	G	A			ahpC	promoter mutation		
Isoniazid	2726767	A	C	ahpC	Rv2428	nonessential	nonsynonymous	A575C	K192T
Ethambutol	4247429	A	G	embB	Rv3795	essential	nonsynonymous	A916G	M306V
Ethambutol	4247495	G	T	embB	Rv3795	essential	nonsynonymous	G982T	D328Y
Ethambutol	4269341	C	T	ubiA	Rv3806c	essential	nonsynonymous	G493A	G165S

Table 5.2: Drug resistance associated variants identified in this study

Isolate	Date	Resistance profile[b]	Predicted WGS-DST Rifampicin/Isoniazid[a]		
			PhyResSe[193]	Mykrobe predictor[103]	TB profiler[139]
G1480	27/04/09	DS	-/-	-/-	-/-
G1479	20/09/13	RR	+/-	-/-	+/-
G516	21/01/14	MDR	+/-	-/-	+/-
G520	17/03/14	MDR	+/-	-/-	+/-
G249	27/06/14	MDR	+/-	-/-	+/-
G252	11/08/14	MDR	+/-	-/-	+/-
G535	13/10/14	MDR	+/-	-/-	+/-
G841	09/12/14	MDR	+/-	-/-	+/-
G842	27/01/15	MDR	+/-	-/-	+/-
G993	15/04/15	MDR	+/-	-/-	+/-
G1003	11/06/15	MDR	+/-	-/-	+/-
G1257	09/11/15	MDR	+/-	-/-	+/-
G1478	11/12/15	MDR	+/-	-/-	+/-
G1720	02/06/16	XDR	+/-	-/-	+/-
G1721	14/10/16	XDR	+/-	-/-	+/-
G1928	09/01/17	XDR	+/-	-/-	+/-

[a] Predicted whole-genome-based DST (WGS-DST) using publicly available software and after this study.

[b] Based on whole-genome data.

Table 5.3: Drug resistance profile predicted by publicly databases. Abbreviations: DS, drug susceptible; MDR, multidrug resistance; RR, rifampicin resistant; XDR, extensively drug-resistance.

Amplicon target [a]	**Primer sequences 5’ - 3’**	**Product length (bp)** [b]	**Reference**
rpoB RRDR (761020-761233)	F-CGATCACACCGCAGACGTT R-GTTTCGATCGGGCACATCC	232	[232]
rpoB 491 (761126-761362)	F-GTCGGGGTTGACCCACAAG R-CAGGTACACGATCTCGTCGC	256	This study
katG 315 (2155074-2155345)	F-CCATGAACGACGTCGAAACAG R-GCTCTTCGTCAGCTCCCACTC	272	[232]
katG 429del (2154941-2154723)	F-AGACAGTCAATCCCGATGCC R-GCGGGTGGATCCGATCTATG	257	This study
oxyR-ahpC promoter (2726015-2726251)	F-ACCACTGCTTTGCCGCCACC R-CCGATGAGAGCGGTGAGCTG	236	[238]
ahpC 192 (2726608-2726856)	F-ACCCCAACAACGAGATCCAG R-GATGTCTTTGGCGTACTCGG	218	This study

[a] Positions correspond to H37Rv genome. NCBI Reference Sequence NC000962.3.

[b] Length includes Illumina adapter sequences.

Table 5.4: Primers used for amplicon sequencing of relevant *rpoB, katG* and *ahpC* regions. Abbreviations: RRDR, rifampicin-resistance-determining-region.

5.6.4 Supplementary Figures

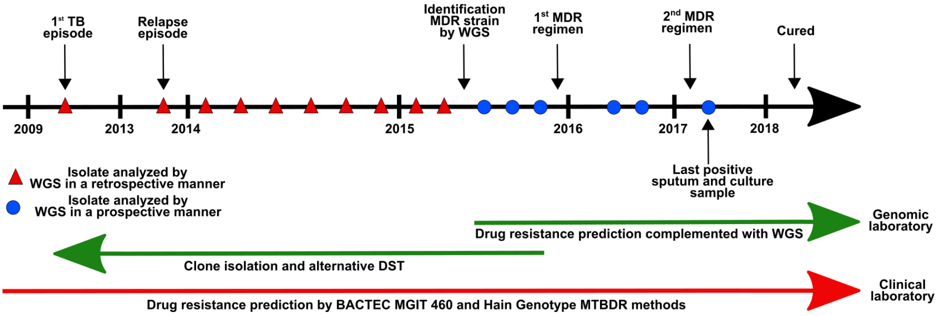
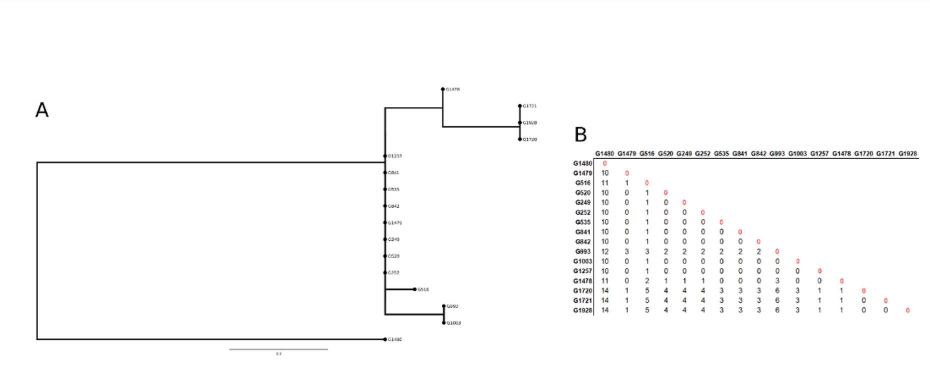

1st TB episode
Relapse episode
Identification MDR strain by WGS
1st MDR regimen
2nd MDR regimen
Cured
2009
2013
2014
2015
2016
2017
2018
Last positive sputum and culture sample
Isolate analyzed by WGS in a retrospective manner
Isolate analyzed by WGS in a prospective manner
Drug resistance prediction complemented with WGS
Genomic laboratory
Clone isolation and alternative DST
Drug resistance prediction by BACTEC MGIT 460 and Hain Genotype MTBDR methods
Clinical laboratory

Figure 5.3: Management of the case and role of the genomic and clinical laboratories. Timeline with the most relevant patient's management clinical events during 9 years of the infection. The clinical laboratory (red line) refers to the routine BACTEC-MGIT (phenotypic) and Hain Gentotype MTBDRplus (genotypic) assays performed in the hospital. The genomic laboratory (green lines) corresponds to results obtained by whole genome sequencing analysis of the isolates and available from 2015 onwards. Once the relevant mutations were identified using WGS individual clones at different time-points were isolated and tested for RIF and INH resistance using an alternative approach to BACTEC MGIT. WGS was also used in a prospective manner to predict DST to new drugs during MDR-TB treatment.



	G1480	G1479	G516	G520	G249	G252	G535	G841	G842	G993	G1003	G1257	G1478	G1720	G1721	G1928
G1480	0															
G1479	10	0														
G516	11	1	0													
G520	10	0	1	0												
G249	10	0	1	0	0											
G252	10	0	1	0	0	0										
G535	10	0	1	0	0	0	0									
G841	10	0	1	0	0	0	0	0								
G842	10	0	1	0	0	0	0	0	0							
G993	12	3	3	2	2	2	2	2	2	0						
G1003	10	0	1	0	0	0	0	0	0	0	0					
G1257	10	0	1	0	0	0	0	0	0	0	0	0				
G1478	11	0	2	1	1	1	0	0	0	3	0	0	0			
G1720	14	1	5	4	4	4	3	3	3	6	3	1	1	0		
G1721	14	1	5	4	4	4	3	3	3	6	3	1	1	0	0	
G1928	14	1	5	4	4	4	3	3	3	6	3	1	1	0	0	0

Figure 5.4: Genetic relationship between all the isolates. Sample G1480 corresponds to the first episode and presented a fully-susceptible resistance profile; G1479 corresponds to the first relapse isolate. A. Inferred maximum Likelihood tree. Isolates G1720, G1721 and G1928 presented an extensively drug-resistant profile. B. Pairwise SNP distanced between all samples.

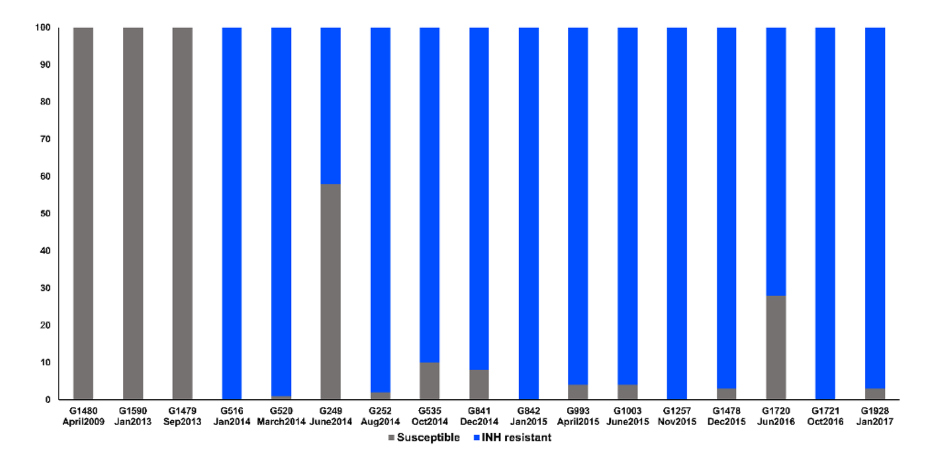


Figure 5.5: Predicted percentage of the susceptible (grey bar) versus INH resistant populations (blue bar) identified across all isolates. The percentage of the INH resistant population refers to the sum of the frequencies of the two INH mutations co-existing in the patient (*katG* G273 and G249del)

Development and application of affordable SNP typing approaches to genotype *Mycobacterium tuberculosis* complex strains in low and high burden countries.

The work described in the present chapter has been published as a Reseach Article in *Scientific Reports* Journal

Cancino-Muñoz, I. et al. Development and application of affordable SNP typing approaches to genotype *Mycobacterium tuberculosis* complex strains in low and high burden countries. Sci Rep 9, 15343 (2019).

https://doi.org/10.1038/s41598-019-51326-2

6.1 Abstract

The *Mycobacterium tuberculosis* complex (MTBC) comprises the species that causes tuberculosis (TB) which affects 10 million people every year. A robust classification of species, lineages, and sub-lineages is important to explore associations with drug resistance, epidemiological patterns or clinical outcomes. We present a rapid and easy-to-follow methodology to classify clinical TB samples into the main MTBC clades. Approaches are based on the identification of lineage and sub-lineage diagnostic SNP using a real-time PCR high resolution melting assay and classic Sanger sequencing from low-concentrated, low quality DNA. Thus, suitable for implementation in middle and low-income countries. Once we validated our molecular procedures, we characterized a total of 491 biological samples from human and cattle hosts, representing countries with different TB burden. Overall, we managed to genotype 95% of all samples despite coming from unpurified and low-concentrated DNA. Our approach also allowed us to detect zoonotic cases in eight human samples from Nigeria. To conclude, the molecular techniques we have developed, are accurate, discriminative and reproducible. Furthermore, it costs less than other classic typing methods, resulting in an affordable alternative method in TB laboratories.

6.2 Introduction

With around 10 million new cases and 1.5 million deaths, tuberculosis (TB) caused by the acid-fast bacillus *Mycobacterium tuberculosis* complex (MTBC), is the first worldwide infectious disease cause of death and remains a major global health problem in low and high burden incidence countries [239]. The MTBC comprises the species responsible for most of the human TB cases worldwide (*M. tuberculosis* and *M. africanum*) as well as those associated with animal disease (*M. bovis, M. caprae, M. canetti*) and the vaccine strain *M. bovis* BCG [240]. Human-associated strains can be further divided into seven main MTBC lineages [127, 157] using robust genomic markers as single-nucleotide polymorphisms (SNPs) that are in agreement with previous genotyping approaches [39]. Some lineages have been associated with a wide geographic distribution, such as the MTBC lineage (L) 4, which is the most predominant around the globe. On the other hand, other lineages are restricted to certain areas [40], such as L7 strains predominantly found in Ethiopia [241], the *M. africanum* L5 and L6, primarily found in West Africa [242]. MTBC L4 is considered the most frequent and genetically diverse. Recently, 10 L4 groups or sub-lineages have been proposed [40].

Lineages and sub-lineages have been associated with different functional and disease phenotypes including differences in transcription, lipids or immunological response but also with disease presentation and epidemiology [119]. However, it has been difficult to correlate specific nucleotide changes or identify regions associated with those phenotypes. Part of the problem is the complex interaction between the host, the bacteria and the environment [89]. As important as these factors are, the need for robust mycobacterial classification systems that can be used worldwide and compared across sites is sought after. Classical genotyping methods such as Spoligotyping [243] and the Mycobacterial Interspersed Repetitive Units- Variable Number of Tandem Repeats (MIRU-VNTR) [244] can fail to discriminate and classify within the MTBC phylogeny [136]. On the contrary, SNP markers are stable over time due to the absence of recombination within MTBC and are congruent with other robust markers like large deletions [136, 157]. Other new techniques like Whole-Genome sequencing (WGS) have a greater resolution but are limited by high costs, difficulties to interpret the results and limited access in low-to-middle income countries.

Recently, two new approaches became available by using real-time PCR, and a ligation-dependent PCR with Luminex flow cytometer technology for genotyping clinical MTBC strains [161]. Here, we present two methods developed for fast, accurate and less expensive MTBC genotyping using High Resolution Melting (HRM) analysis with real-time PCR reactions (real-time PCR-HRM) on multiplex and uniplex reactions with an unspecific dye, and automatized Sanger sequencing. HRM analysis assays were performed before on MTBC strains, nevertheless, all studies focused on the detection of variants related with drug resistance ([245, 246, 247, 248]), and to differentiate MTBC members in cultured and non-cultured samples [249].

In addition, we applied these new approaches to a collection of 491 clinical uncharacterized isolates from three different burdens countries: i) human derived samples from a low-burden region in Spain; ii) human derived samples from a high-burden region in Liberia, West Africa; and iii) human and cattle derived samples from abattoirs in Nigeria. The three datasets were used to show an accurate picture of the circulating lineages and sub-lineages in the different regions and to explore zoonoses between humans and cattle. Furthermore, we successfully tested our molecular assays in complex biological samples that included low DNA concentrations, unpurified heat-killed extracts, as well as contaminants that could affect the PCR performance. These methods were developed in the need to reduce the costs of typing in diagnostic laboratories, especially in high burden countries.

6.3 Methods

6.3.1 Ethics Statement

For the Valencia biological samples, the study was approved by the corresponding Ethics Committee of the Regional Health Office for Valencia (Spain), with an exemption for informed consents from the corresponding Ethics Committee on the basis that this study is part of the surveillance program of communicable diseases by the Public Health Regional Program and, as such, falls outside the mandate of the corresponding Ethics Committee for Biomedical Research. For the Nigerian samples: Based on the premise that the country has a high level of illiteracy, verbal consent was obtained before sample collection. This was approved by the UI/UCH Ethics Committee of the University of Ibadan (UI/EC/14/0198 number). For the Liberian samples: Formal ethical approval for the study was obtained from the Liberian Institute for Medical Research (EC/LIB/914/923 resolution number). Thereafter, the Liberian National Leprosy and Tuberculosis Control Programme and Ministry of Health and Social Welfare approved to collect samples from the hospitals used. For the Liberian samples, there was no direct contact with the patients and therefore exemption from informed consent was granted. All patient personal information was anonymized and no data allowing individual identification was retained. All research was performed following relevant guidelines and regulations.

6.3.2 Biological samples

We used a reference set of strain DNA samples to set-up the assays [250]. The set included 40 DNA samples corresponding to all MTBC lineages and the main L4 sub-lineages obtained in collaboration with The Swiss Tropical and Public Health Institute (Swiss TPH, [n=22]), as well as by a local ongoing TB project (n=18) (**Supplementary Table 6.5**). Swiss TPH Reference samples were well defined by WGS in previous studies [127, 40]. We used this reference set to optimize the real-time PCR-HRM assay as well as Sanger sequencing. Afterwards, we tested our molecular approaches for validation on two different TB burden settings: i) low-burden - 219 clinical isolates of the Hospital Universitario y Politécnico La Fe, Valencia, Spain obtained during the years 2011-2013; and ii) high-burden - 188 samples from Nigeria Countrywide as well as 78 clinical samples from Liberia. The Nigeria samples were enriched by isolates from cattle lesions.

6.3.3 Reference collection strains

Samples from Swiss TPH were incubated in liquid media Middlebrook 7H9 (Becton Dickinson) at 37ºC during two weeks, while as those from our laboratory were grown in commercial Middlebrook 7H10 agar (Becton Dickinson) with OADC supplement. In all cases, DNA extraction method was performed following the CTAB method [228]. Purified DNA was used to perform our molecular approaches.

6.3.4 Sample preparation

In the case of the validation datasets, we tested our molecular approaches in direct supernatants prior to an inactivation step, following the next procedure: all samples were grown in standard Lowenstein-Jensen solid media (BBL, BD) and incubated at 37℃ during 3-4 weeks; next, the inactivation was performed by a heat-kill cycle of 30min at 95℃ and centrifuged. We used the heat-inactivated supernatant to perform the molecular assays. Moreover, DNA concentration was quantified using PicoGreen® (Molecular Probes) in samples from Valencia, while Qubit fluorometer (ThermoFisher Scientific) was used on those from West African samples.

6.3.5 SNP selection for molecular assays

For real-time PCR-HRM method, we used a combination of specific-lineage and sub-lineages SNPs previously described [39, 40, 161, 120, 251] as well as new seven markers developed in the present study. Novel markers were identified by analyzing 34,167 SNPs previously identified in a dataset of 219 globally representative genomes [127]. Given that the vast majority of SNPs in MTBC are billelic and does not show evidence of convergent evolution we applied a parsimony-based approach to map and extract all the specific lineages and sub-lineages SNPs from a global MTBC phylogeny. We used the "Trace Character History" module implemented in MESQUITE (http://www.mesquiteproject.org) to obtain the polymorphisms that were common to each lineage and sub-lineage. We obtained a SNPs candidate list of 2,056 and 1,337 for lineages and sub-lineages, respectively. To choose a SNP candidate for the real-time PCR-HRM, we prioritized synonymous variants detected in essential genes. In all cases, we met this criteria, except in SNP markers for L5 and L4.1.3, in which both were non-synonymous changes. As a proof of concept, all the selected SNPs were validated against a database of 4,595 genomes recently published by our group to assure their stability as markers [194]. The same SNP dataset was used to identify genomic regions with less than 1 Kb that contain the higher number lineage-specific markers. After mapping the genomic coordinates against H37Rv reference genome (NCBI reference number NC˙000962.3), we found two candidate regions harbouring three lineage specific variants each. The first region contained the specific SNPs for L1, L2 and L3, the second region had the diagnostic variants for L3,L4 and L5.Afterwards, we designed an amplification assay including the SNPs of interest in both regions for follow-up Sanger sequencing (**Figure 6.1**).

6.3.6 Primers design and specificity

All primers for both molecular approaches were specifically designed in this project, except diagnostic SNPs for L4 and L6 that were previously published

[39, 161]. Oligosequences were designed using the Primer3Plus [252] online tool (www.primer3plus.com). To make multiplex amplifications in a single tube, we increased the melting temperature by adding AA/TT tail-bases in two primer paired sequences. After primer design, we performed a BLAST [253] search to evaluate *in silico* specificity. Furthermore, we did conventional PCR to corroborate the size and specificity of the amplicons. In all cases, we used the manufacturer conditions of the KAPA2G Fast PCR amplification kit (KAPABiosystems) adding 500 nM of each forward and reverse primer and 10 ng of DNA in a final volume of 25μl. PCR products were visualized using standard agarose gel electrophoresis (1.4%) for 1hr at 110V. Data containing the diagnostic SNP positions, as well as the primer sequences used for all molecular approaches are in **Tables 6.1-6.2** (real-time PCR-HRM approaches), and **Supplementary Table 6.6** (Sanger sequencing approach).

6.3.7 Real-time PCR-HRM for lineages and sub-lineages

We optimized the use of the real-time PCR-HRM technology for the rapid detection of specific MTBC lineages (from L1 to L6, and *M. bovis* clade), as well as L4 sub-lineages. For the most common lineages, we designed two multiplex reactions: one containing the specific oligonucleotides for L2, L3 and L4 and the second including the primers for specific L1 and L6. In contrast, uniplex reactions were developed for L5 and *M. bovis* identification. In all cases, each PCR reaction had a final volume of 10 μl. This reaction consisted in 5μl of the KAPA HRM FAST PCR Master Mix (KAPABiosystems), which includes the unspecific EvaGreen® dye and dNTPs; 2.5 mM of MgCl2 (25mM) and 10 ng of template DNA. Every forward and reverse primers had different concentrations depending on the multiplex reaction. In the first multiplex reaction, we added 200 nM of each primer for L2 and L3, while for L4, we used 600 nM of each oligonucleotide. In the second multiplex reaction, the concentration of each pair of primers was 200 nM. In the case of the Uniplex assay (L5 and *M.bovis*), a total of 400 nM of each primer were added to the mix. Finally, distilled water was added to complete a 10 μl final volume. For all lineages the PCR amplification step consisted in an initial denaturation step at 95℃ for 5 min; 40 cycles of denaturation (95℃, 10 s), annealing (57℃, 20 s) and final extension (72℃, 20 s).

Real-Time PCR-HRM for L4 specific sub-lineages was performed following uniplex reaction conditions described above, with the exception of the annealing step, which was 55℃ due to the oligonucleotides melting temperature. All reactions were performed in three technical replicates per sample to test the reproducibility of the assay.

The HRM assay was performed at the end of each reaction and consisted of one cycle of increasing temperature from 45℃ to 97℃ at ramps rate of 2.2 ℃/s. Fluorescence signal changes were collected at the end for posterior analysis. In every run we used several controls. A free-DNA well serve as a non-template control. In addition, reference DNAs for all lineages were used to assign the melting curve of the sample to one for the lineage/sub-lineage and as controls.

6.3.8 Analysis of the melting curves in HRM assay

All the real-time PCRs (multiplex and uniplex) reactions and HRM curves analyses were performed with a Roche LightCycler480 instrument (Roche Applied Science, Germany) and Gene Scanning software, respectively. This analysis consists of four steps: 1) identify samples that did not amplify as negatives to exclude them from the analysis; 2) normalize the melting curve data indicating the values of initial and final signal fluorescence for all samples; 3) adjust the temperature of the normalized melting curves at the point where the DNA is denatured to distinguish the changes in the shape of samples; and 4) finally, the melting curves are represented by a difference plot. The program allows selecting a baseline curve to show the differences based on this. In all the cases negative controls for the lineage or sub-lineages evaluated were used for this baseline curve.

6.3.9 Sanger sequencing

We performed Sanger sequencing on the reference set to identify major lineages. First, the amplification step was carried out with conventional PCR using the same reaction conditions as described above (primers specificity section). Then, PCR products were labeled and purified according to the BigDye Terminator v3.1 Cycle Sequencing Kit (AppliedBiosystems) and PCR ExcelaPure 96-Well UF Purification Kit (EdgeBioSystems) protocols, respectively. The two target regions were sequenced with the ABI 3037xl DNA analyzer (AppliedBiosystems). Finally, the resulting sequences were analyzed using Pregap5 and Gap5 programs [254], both included in the Staden package.

6.3.10 Performance of the molecular techniques

We evaluate the performance of the techniques by calculating the accuracy, sensitivity, specificity, positive predictive value and negative predictive value of each. WGS lineage definition was used as a gold standard genotyping method. We tested a total of 76 whole-genome sequenced heat-inactivated clinical samples, 43 from Liberia dataset and 33 from Valencia dataset.

6.3.11 Biosafety procedures

DNA extractions from cultures were done in a BSL-3 facility as per WHO recommendations. Samples from Nigeria and Liberia were heat-inactivated at the Nigeria laboratory. Once arrived, s second inactivation step to assure killing of the bacteria was done at FISABIO‘s (Spain) BSL-3 facility. In the case of the Valencia samples, culture procedures as well as heat-inactivated step were performed at the Hospital Universitario Politécnico la Fe BSL-3.

6.3.12 Data availability

The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request. The Sanger sequences generated were uploaded to the NCBI database under submission numbers 2262663 and 2262671.

6.4 Results

6.4.1 Lineage and sub-lineage identification by real-time PCR-HRM and Sanger sequencing

To identify diagnostic SNPs for typing all MTBC lineages and L4 sub-lineages, we analyzed a global reference collection of 219 MTBC strain genomes. A total of 34,167 SNPs were found affecting a variable number of strains as previously published. We mapped the variants to the corresponding MTBC phylogeny and extracted all that were common to the strains belonging to a lineage or sub-lineage. A total of 2,056 lineage SNPs and 1,337 sub-lineage SNPs were identified as candidate markers. Specific common L7 variants were discarded. The diagnostic potential of all variant candidate were corroborated in-silico in a larger collection of 4,495 MTBC genomes [194] to assure their stability (see **Methods** section and **Figure 6.1**).

Using the annotated specific lineage variant list, we identified two short genomic regions (less than 500bp) containing SNP markers for the six main human lineages (L1-L6). The first region involved the specific phylogenetic markers for L1, L2 and L5 at H37Rv (NCBI, NC_000962.3) reference genomic positions 4357773GA, 4357804TG, 4357657GA, while the second region included SNPs for L3, L4 and L6 at positions 1281984GA, 1281771CT, 1281685CG, respectively (**Supplementary Table 6.6**). After amplifying both regions and testing the specificity of the primers (**Supplementary Figures 1-2**), we used 18 DNA strains representative of the MTBC lineages diversity as controls (reference dataset, **Supplementary Table 6.5**) to performed Sanger sequencing. A visual analysis of the sequences generated showed that no nucleotide differences in the amplified regions were detected against the wild type sequences, except for those mutations corresponding to each specific lineage (**Supplementary Figure 3**). The diagnostic amplicon positions in the first region for L1, L2 and L5 markers were 243GA, 274GA and 127TG, respectively. In the case of the second region, amplicon positions for detecting L3, L4 and L6 were 310GA, 274GA and 127TG, respectively.

Different candidate SNP were selected based on different features to design a set of primers for each lineage and sub-lineage for real-time PCR-HRM assay (**Tables 6.1-6.2**). Seven diagnostic SNPs were tested for the real-time PCR-HMR detection of MTBC lineages on the reference dataset (n=40) (**Supplementary Table 6.5**). To optimize the reaction, we developed two multiplex real-time PCR reactions to detect the most common MTBC lineages (L1-L4 and L6). While specific lineages (L5 and *M. bovis*) were detected by a uniplex PCR reaction each, analysis of melting curves on HRM assays denoted the differences between positive controls according to their melting temperature and signal fluorescence (**Figure 6.2**).

Given the importance of L4 as the most successful MTBC lineage, we developed a real-time PCR-HRM assay to type the most common sub-lineages defined by WGS. We used six diagnostic SNPs previously described [40], while the rest were designed after screening of a global SNP database 3 In all cases, primers were designed to assure the specificity of the real-time PCR-HRM assay (**Figure 6.2**). Once more, we used 24 DNAs representing the most common L4 sub-lineages (reference collection). HRM analysis showed a clear discrimination between positive samples (those with the specific SNP) against those with wild-type genotype (**Figure 6.3**).

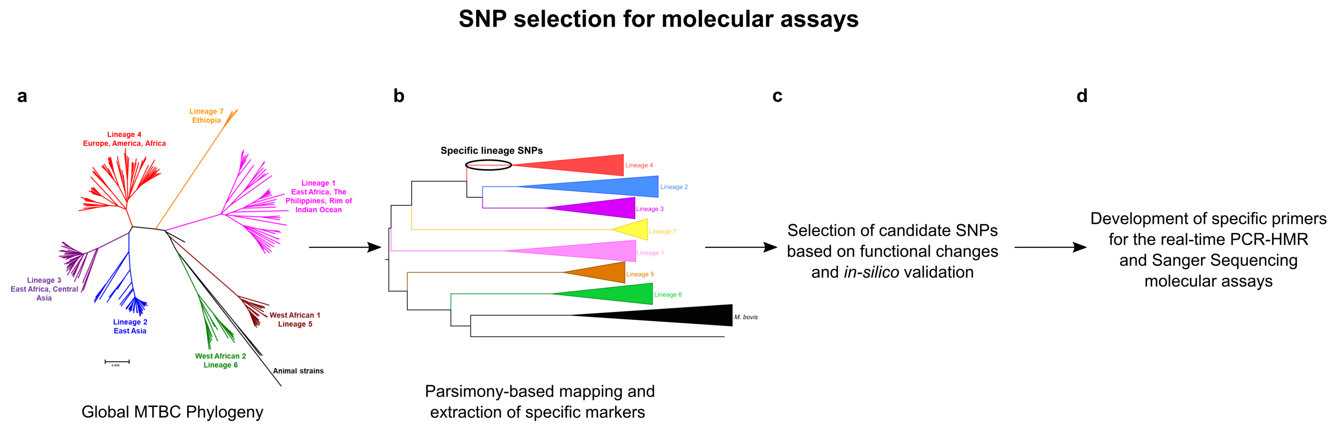

SNP selection for molecular assays
a
b
c
d
Lineage 7
Ethiopia
Lineage 4
Europe, America, Africa
Lineage 1
East Africa, The Philippines, Rim of Indian Ocean
Lineage 3
East Africa, Central Asia
Lineage 2
East Asia
West African 1
Lineage 5
West African 2
Lineage 6
Animal strains
Global MTBC Phylogeny
Specific lineage SNPs
Lineage 4
Lineage 2
Lineage 3
Lineage 7
Lineage 1
Lineage 5
Lineage 6
M. bovis
Parsimony-based mapping and extraction of specific markers
Selection of candidate SNPs based on functional changes and in-silico validation
Development of specific primers for the real-time PCR-HMR and Sanger Sequencing molecular assays

Figure 6.1: Workflow to identify and select diagnostic lineage/sub-lineage SNP for the development of the molecular assays. (a). We used a public available variant list containing 34,167 SNPs from 219 MTBC global strains. (b). First, we performed a parsimony-based approach to map and obtain 2,056 specific SNPs markers for all lineages, and 1,337 for all L4 sub-lineages. In the figure, the specific SNPs for lineage 4 in the phylogeny are marked with a circle. (c). Then, candidate SNPs were selected according to their functional change (we used synonymous SNPs in essential genes, whenever possible), and tested *in-silico* against a global dataset of 4,495 MTCB strains. (d). Finally, we developed seven specific primers to implement in the real-time PCR-HRM molecular assay, and two genomic regions for Sanger sequencing approach.

MTBC lineage[a]	**SNP position**[b]	**Nucleotide change**[c]	**Gene**	**Primers sequences 5'-3'**	**Amplicon size**	**Reference**
1	115499	T/G	Rv0101	F-ATAATATTGCGTCGGTGTTGG R-TTATATATTAATGGGCAGGCC	81bp	This study
2	3304966[d]	G/A	Rv2952	F-TGTTACCCGCACTTTCGGCGTTT R-AGGTCGGCGTATGGGAGGTA	80bp	This study
3	4266647[d]	A/G	Rv3804c	F- GCGACATACCCGTGACGGC R- CGTTGAGATGAGGATGAGGG	92bp	This study
4	2154724	A/C	Rv1908c	F-CCGAGATTGCCAGCCTTAAG R-GAAACTAGCTGTGAGACAGTC	64bp	[39]
5	456731	C/T	Rv0380c	F-GCATCGTGTCCGAAGTTCTC R-ATCATCGCCGACATCGATAC	68bp	This study
6	378404	G/A	Rv0309	F-CCGACAGCGAGAACCTGC R-CCATCACGACCGAATGCTT	54bp	[136]
M.bovis	2831482	T/G	Rv2515c	F-GTGTTGCTGTCGATGACGC R-ACTGGTACCGCAATACCGTC	91bp	This study

[a] Nomenclature proposed by Comas et al [120]

[b] Genomic position on the H37Rv reference genome (NCBI, NC_000962.3).

[c] Allelic change in the reference genome.

[d] Mutation previously described by Fenner at al [251]

Table 6.1: Specific lineage primers used in the study.

MTBC lineage[a]	SNP position[b]	Nucleotide change[c]	Gene	Primers sequences 5’-3’	Amplicon size	Reference
L4.1.1	3798451[d]	C/G	Rv3383c	F-ATCGACTCAATGGCCCGATG R-TGACTCTGGATGCGGTTTT	112bp	This study
L4.1.2	4323348[d]	C/T	Rv3848 /3849	F-AAATCCGTTCGTCGTGTGGA R-CTGACGTTGTGAGGGGTCAA	82bp	This study
L4.1.3	4409231	T/G	Rv3921c	F-GACCGCCTCCTGCTTTTTG R- ACGTCTTCGGCATGATCGAA	53bp	This study
L4.2	2942377	C/T	Rv2614c	F-GAGTAGTCCTCCAGTTCGCG R-TCAGCTTCCCCGACGAAATC	85bp	This study
L4.3	1480024[d]	G/T	Rv1318c	F-CAGGCCAGGATCCACATCAG R-TGCTGCTCAATCTCACTCGG	100bp	This study
L4.4	4307886	G/A	Rv3834c	F-AAGGTGGTGCAGTTCGAC R-ACTGCGAGGCGTGGATTC	69bp	This study
L4.5	2789341[d]	A/C	Rv2483c	F-GGAGGCCTCACCATCCTTG R-ACGAAGGCGGCTACAAAGAA	81bp	This study
L4.6.1	435708	G/A	Rv0357c	F-CAAAGATCCCGCTGGGTCAT R-GATATGAGATCGACGGCCGG	58bp	This study
L4.6.2	3191099[d]	C/A	Rv2881c	F-CATCATGCAGAACACCCATC R-CCCATTGTTCTGCTCTTTCG	72bp	This study
L4.10	1692141[d]	C/A	Rv1501	F-GCTCGGTGTTCTTCGACTCA R-TGGCCGTTTCAGATAGCACA	107bp	This study

[a] Nomenclature proposed by Stucki et al [40].

[b] Genomic position on the H37Rv reference genome (NCBI, NC_000962.3).

[c] Allelic change in the reference genome.

[d] Mutation previously described by Stucki et al [40].

Table 6.2: Specific L4 sub-lineages primers used in the study

6.4.2 Performance of typing methodologies

In the previous section we used a reference dataset to set-up the real-time PCR-HRM and Sanger sequencing assays. To evaluate and test their robustness, we calculated some performance parameters from a collection of 76 heat-inactivated clinical samples that had been whole genome sequenced in our laboratory (43 from Liberia dataset and 33 from Valencia), ranging from 0.054-8.08 ng/μl DNA concentrations .In both cases, WGS was use as a gold standard genotyping method.

The sensitivity if the real-time PCR-HRM assay was 97.37% (74/76, [95% CI: 98.82-99.68%]), while the specificity was 100% (368/369, [95% CI: 99.01-100%]).Overall, real-time PCR-HRM assay accuracy was 99.33% (95% CI: 98.04-99.86%).

The sensitivity of the Sanger sequencing was 71.05% (54/76, [95% CI: 59.51-80.89%]), while the specificity was 100% (305/305, [95% CI: 98.80-100.00%]), giving a negative predictive value of 92.44% (95% CI: 89.52-94.56%). Overall, the Sanger Sequencing assay accuracy was 94.23% (95% CI: 91.39-96.35%) (**Figure 6.3**). Thus, although Sanger sequencing allowed us to accurately identify the lineage, a positive result strongly depends on the successful amplification of at least one genomic region.

Screening method	No. of samples analyzed (n=76)	Sensitivity (95% CI)	Specificity (95% CI)	PPV	NPV	Accuracy
Real-time PCR-HRM	74	97.37% (98.82-99.68)	100.00% (99.01.100)	100%	99.46% (97.91-99.86)	99.33% (98.04-99.86)
Sanger Sequencing	54	71.05% (59.51-80.89)	100.00% (98.80-100)	100%	92.44% (89.52-94.56)	94.23% (91.39-96.35)

Table 6.3: Performance values of the molecular techniques used in the study. These values were extracted from 76 whole-genome sequenced heat-inactivated clinical samples. Abbreviations: PPV, Predicted Positive Value; NPV, Negative Predictive Value.

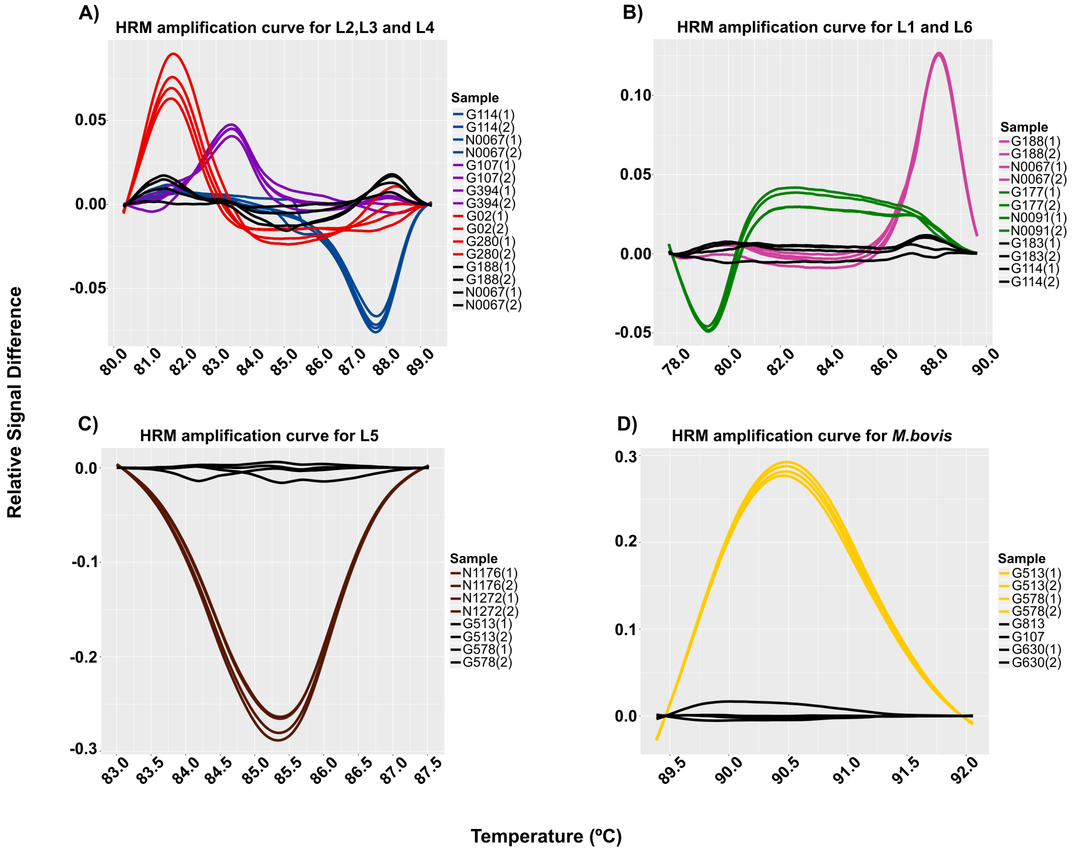


Figure 6.2: Amplified melting curves for the detection of the MTBC lineages. The graphs show the difference of the melting curves obtained by HRM analysis. Each line indicates a different sample. (A,B) Curves obtained by Multiplex PCR; (A), blue lines represent lineage 2; purple lines represent lineage 3; and red lines indicate lineage 4; (B), pink lines represent lineage 1; green lines represent lineage 6. (C,D) Curves obtained by Uniplex PCR; (C), brown lines indicate lineage 5; (D), yellow lines represent *M.bovis* lineage. In all cases black lines indicate a wild type genotype.

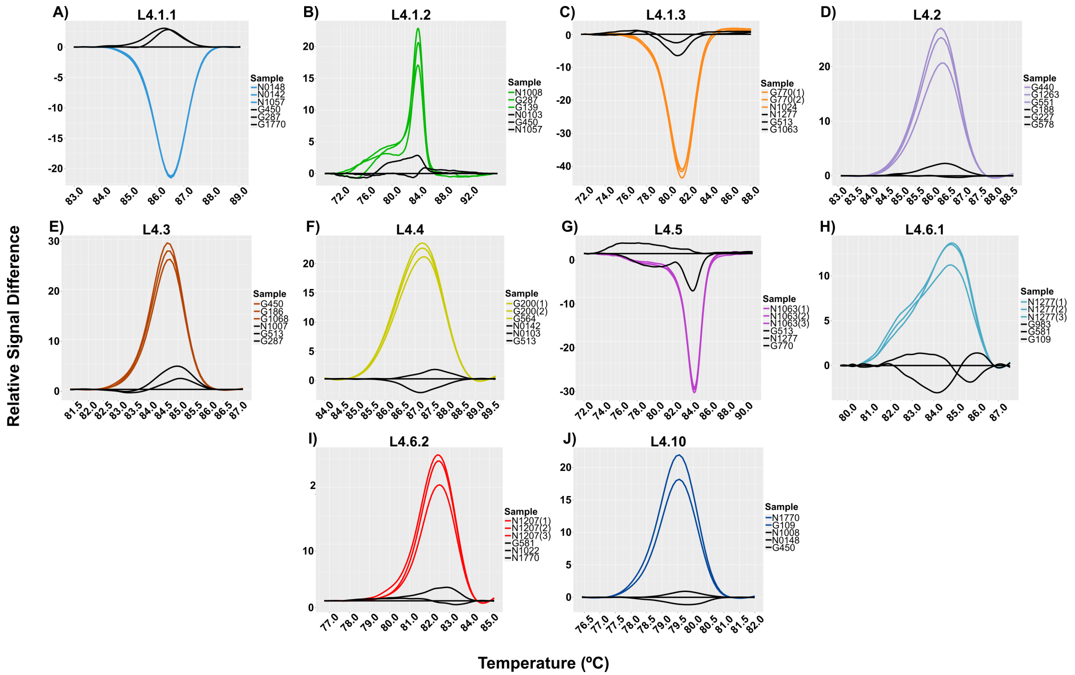


Figure 6.3: Amplified melting curves for the detection of the MTBC L4 sub-lineages. The graphic shows the difference of the melting curves obtained by HRM analysis. Each letter represent a difference plot of every sub-lineage tested. (A) sub-lineage L4.1.1. (B) sub-lineage L4.1.2. (C) sub-lineage L4.1.3. (D) sub-lineage L4.2. (E) sub-lineage L4.3. (F) sub-lineage L4.4. (G) sub-lineage L4.5. (H) sub-lineage L4.6.1. (I) sub-lineage L4.6.2 and (J) sub-lineage L4.10. In all cases black lines indicate a wild type genotype.

6.4.3 Molecular characterization in a low-burden region

Once the methods were optimized and validated, we sought to apply the real-time PCR-HRM technique to heat-inactivated bacteria from a clinical laboratory in order to test the robustness of the typing scheme to DNA amount and purity. We used 219 DNA samples from Valencia (Spain) ranging from 0.02 to 18.3ng/μl total DNA amount. These samples were heat-inactivated extracts and had not been molecularly characterized before. SNP typing identified 191 isolates (87.2%) as L4, whereas the L3 and L2 were identified in nine (4.1%) and eight (3.6%) samples each. L1 and *M. bovis* lineages were present in four and two cases, respectively. Finally, L5 was not identified. The most frequent sub-lineages identified were L4.1.2 (Haarlem family) with 66 (30%) cases, followed by L4.10 with 51 (23.2%) cases, and L4.3 (LAM family) with 49 (22.3%) isolates (**Figure 6.4**).Furthermore, with the patient origin data obtained from 140 cases, we constructed a map including the sub-lineages frequencies to see whether they are globally spread or just found in restricted areas (**Figure 6.5**). We observed that the sub-lineages L4.1.2 and L4.10 are more frequent in East Europe, while L4.3 is more common in Latin America and East Africa. In contrast, L4.6.2 (Cameroon family) and L4.4 are more specific for some regions of Africa as previously reported [40].

6.4.4 Molecular characterization in a high-burden region

We genotyped 266 tuberculosis samples from five different Nigeria districts as well as from Liberia. For this dataset, we also used heat-inactivated extracts. DNA concentrations ranged from 0.05 to 33.2ng/μl. We were not able to characterize 17 samples due to amplification problems.

The sampling scheme from Nigeria (n=171) was thought to increase the chances to identify zoonoses events between humans and cattle. Thus 58% of the samples (n=99) were obtained from cattle lesions while 42% (n=72) derived from human patients. We screened these samples with our specific *M. bovis* real-time PCR-HRM assay to distinguish between M. bovis and the rest of the

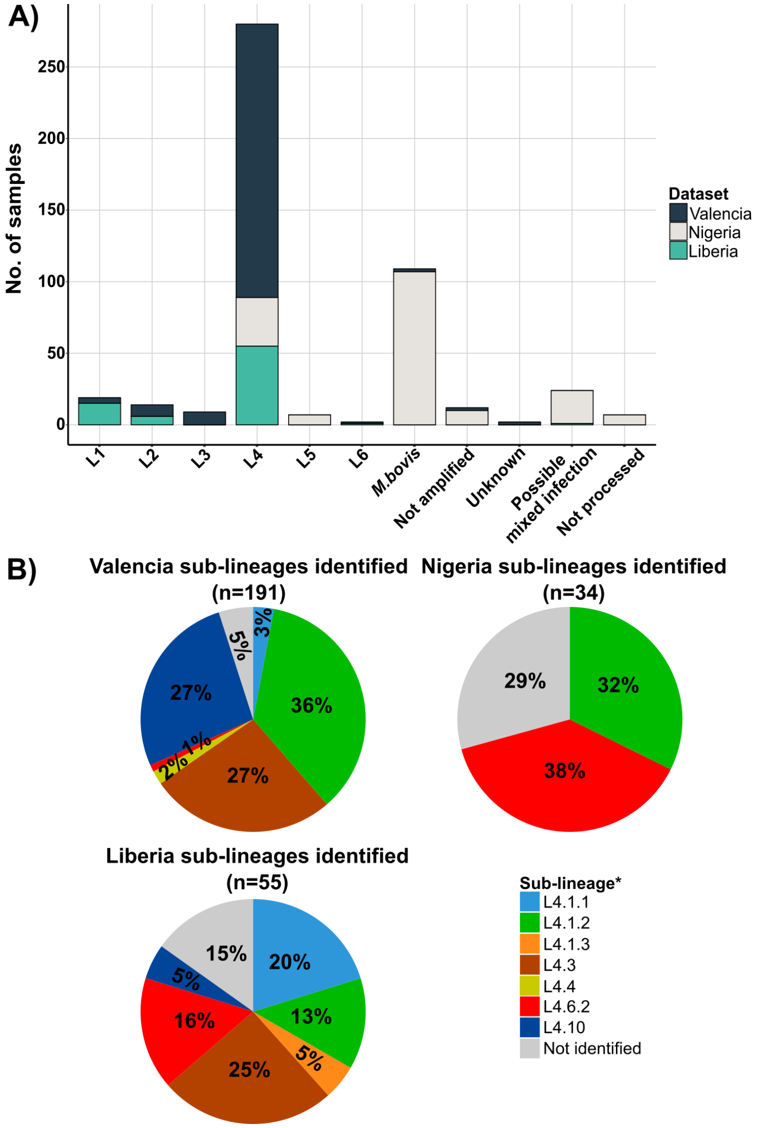


Figure 6.4: MTBC genotypes identified by real-time PCR-HRM assay in the three study regions. a, Proportion of the main lineages detected. b, Principal sub-lineages belong to L4. Numbers inside of the pie charts represents percentage. * Nomenclature proposed by Stucki et al [40]

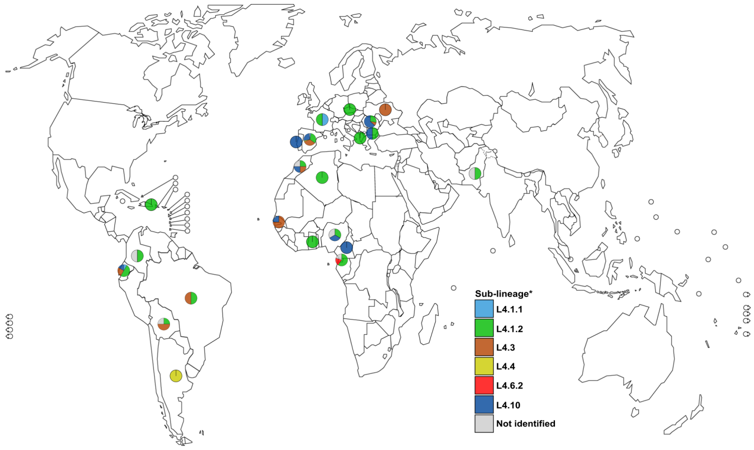


Figure 6.5: Global distribution of all lineage 4 samples in Valencia region by patient origin country. * Nomenclature proposed by Stucki et al [40]

MTBC strains (**Figure 6.2D**). We detected the presence of an *M. bovis* infection in 99 (93%) cattle lesions. No human tuberculosis was identified in any cattle-derived sample (**Figure 6.4**). In contrast, we identified bovine tuberculosis infecting eight human cases. We identified that six cases were from Makurdi, North-central-Nigeria. For the rest of the human samples (n=64), the most common genotype detected was L4 with 34 cases (20%), with the specialist L4.6.2 as the most frequent sub-lineage (13 cases). Surprisingly, we detected the presence of unclassified sub-lineages in 10 samples. L5 was classified in seven isolates. Additionally, ambiguous HRM profiles were identified in 13.5% of the human samples (n=23). This ambiguous results could be due to the presence of mixed infections by different MTBC strains, potential contamination errors and/or, less likely, to possible PCR artefacts. In addition, with the geographic data, we created a distribution map of the lineages (sub-lineages included), in order to have a snapshot of the MTBC diversity affecting the abattoirs in the zone (**Figure 6.6**).

In the case of the Liberian samples (n=78), the most frequent lineage detected was L4 corresponding to 70.5% (n=55) of all MTBC strains, followed by Indio-Oceanic L3 with 19% (n=15). L2 and L6 were the less frequent lineages with six and one cases, respectively. Within L4, we detected that the L4.3 was the most common with 25.5% of the samples (n=14), followed by L4.4.1 (X genotype) with 20% of the infecting strains (n=11). The specific L4.6.2 was found only in nine cases (16.3%) of all the clinical samples (**Figure 6.4**). Furthermore, we were not able to classify eight heat-inactivated L4 MTBC samples, suggesting that local sub-lineages are circulating in the country.

	Host	
MTBC lineage	**Cattle**	**Human**
Lineage 4	0 (0%)	34 (20%)
Lineage 5	0 (0%)	7 (4%)
M.bovis	99 (58%)	8 (5%)
Possible mixed infection	0 (0%)	23 (13%)
Total	99 (58%)	72 (42%)

Table 6.4: Global distribution of all lineages (sub-lineages included) identified in all Nigeria regions. The map shows the MTBC diversity circulating in Nigeria study regions. The diameter of each circle represents the number of samples obtained from each zone.

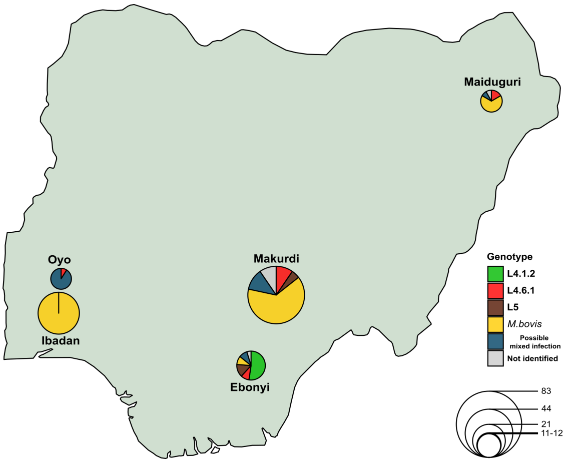


Figure 6.6: Global distribution of all lineages (sub-lineages included) identified in all Nigeria regions. The map shows the MTBC diversity circulating in Nigeria study regions. The diameter of each circle represents the number of samples obtained from each zone.

6.5 Discussion

In this study, we developed two genotyping methods using faster and less expensive technologies such as real-time PCR-HRM and Sanger automatized Sequencing with informative SNPs for the main six lineages of MTBC, including the *M. bovis* clade and the most common L4 sub-lineages. First, we set up Multiplex and Uniplex real-time PCR-HRM reactions for the main MTBC lineages and specific L4 sub-lineages using an unspecific dye instead of specific probes, which helps to reduce the cost of the assay. We optimized the PCR assay to obtain a reliable result using lower reagents volumes than the manufacturer's recommendation (10 vs 20μl of total reaction volume). By comparing this method with classical genotypic approaches such as spoligotyping which costs 26US$ per isolate [164], we demonstrated a relatively lower cost of our assay, being of 2.4US$ in the case of a uniplex PCR reaction and lower than 0.9US$ using multiplex condition. Additionally, our molecular techniques have a similar cost when we compared them with other SNP-based genotyping methods such a Luminex MOL-PCR assay, which cost 0.8 and 0.15EUR for uniplex and multiplex reactions, respectively [161]. Nevertheless, Luminex platform is considered a not conventional laboratory equipment, especially in low- and middle-income countries. Moreover, we show that our real-time PCR-HRM assay works with heat-inactivated, low-concentrated DNA samples, which are commonly generated in TB diagnostic laboratories.

It is well known that WGS are decreasing every year, mostly based on the high-throughput capacity for sequence several samples per run (up to 24 isolates with the Illumina MiSeq instrument), or because third generation sequencers technologies (such as the Oxford Nanopore MinION device) are more accessible (between 100-150EUR per isolate). Besides this, many low- and middle- income countries do not have the necessary equipment. In addition, in many high burden countries, the number of TB cases is larger so it is not feasible to perform WGS all of them [56], thus providing installations for

PCR-based approaches as a quick and affordable screening method.

Using a validation dataset of 76 clinical strains well-defined by WGS to test the performance of the techniques, we were able to genotype up to 97.3% and 71% of samples by real-time PCR-HRM and Sanger sequencing, respectively, both of the with specificity values of 100%. The fact that we could not perform Sanger sequencing assay on some samples, could be due to the low amount of DNA that we were able to recover from them, being insufficient for the minimum required concentration for the amplification on conventional PCR assays. Despite this, all the samples that were amplified by both methods (n=54), show a 100% of concordance with lineage defined by WGS, even if the techniques differed in the SNP position target. This result suggests that both molecular assays work with heat-inactivated, and low-concentrated DNA samples, which are commonly generated in TB diagnostic laboratories. We applied our real-time PCR-HRM approach to analyze three uncharacterized collections from different TB burdens countries to test its efficiency. First, we were able to classify 98.6% (215 out of 219) and 93.6% (249 out of 266) of the clinical samples from the low-burden (Valencia, Spain) and high-burden settings (Nigeria and Liberia), respectively. These results are in agreement with studies in which authors mention that up to 6% of the strains were not classified [255, 162]. Additionally, we wanted to test the sensitivity of the approach, and we found a positive result while using heat-inactivated clinical samples concentrations below the limit of detection of fluorometric quantitation. These results indicate reliable results, even by using non purified DNA as a template, and as a consequence, a rapid detection method.

Regarding the genotypes identified in Valencia, we found out that majority of the MTBC cases belonged to the global L4 (87%) being subdivided into the generalists L4.1.2 (36%), L4.10 (27.3%), and L4.3 (26.7%). The frequencies detected were in concordance to those reported before worldwide [40], in the same region [256], and the same country [257]. Moreover, we identified the rest of the MTBC lineages, except the specific L5, reflecting the high MTBC diversity circulating in the region. Regarding the global distribution of the country of origin of foreign-patients, the analysis shows a congruent result with data reported before. For example, patients infected by L4.10 strains were mostly from East Europe (16 out of 23) also, TB cases from Latin America countries were infected by L4.3 MTBC strains (9 out of 15).

Similar results were obtained with the Liberian samples. We detected that the dominant lineage of the MTBC is L4 (70.5%), followed by L1 (19.2%). As for sub-lineages, we identified the presence of the generalists (L4.3, 25.5%), and specialists (L4.6.2, 16.3%) clades almost with the same frequencies, corresponding with those reported in West African border countries [40, 258]. The presence of a high percentage of L4 unclassified samples (14.5%) suggested that endemic sub-lineages (highly likely to be specialist clades) are circulating in the region. This snapshot of the MTBC diversity in the Liberian population denotes the influence of migration inside the country, probably increased after country foundation at the beginning of the 19th century. To our knowledge, this is the first time that Liberian TB samples are genotyped.

The use of a specific lineage marker that identifies the *M. bovis* clade could be helpful to promptly distinguish this genotype from the *M. tuberculosis* strains in clinical samples. As a proof of concept, we performed our real-time PCR-HRM technique on 171 uncharacterized samples mostly from cattle. We found out that the majority of cases (60%, n=107) belonged to *M. bovis* species. Furthermore, we identified six human-samples harbouring bovine tuberculosis from the same region.

One limitation of these assays is that we only interrogated one specific SNP, and as a consequence, additional biological information such as epidemiological markers (e.g. transmission clusters detection) will be missed. Nevertheless, these techniques are flexible and could be adapted to identify other specific polymorphisms like, for example, the detection of antibiotic resistant MTBC samples [248, 259, 167], or the rapid identification of local transmission clusters [255]. We optimized the protocol using the Roche LightCycler480 system which could be an uncommon laboratory device. Nevertheless, we obtained reliable results using the less expensive

LightCycler96 instrument. In fact, West African samples were genotyped using this system.

In summary, the molecular approaches developed here show an accurate, discriminative and reproducible methodology to genotype MTBC strains. Due to a need for common and affordable reagents, these techniques could be useful in TB diagnostic laboratories from low- to middle-income countries.

6.5.1 Acknowledgements

This work was supported by European Research Council (grant, 638553-TB-ACCELERATE) and the Ministerio de Economía y Competitividad grants (SAF2013-43521-R and SAF2016-77346-R) both to Iñaki Comas.

The work described in the present chapter has been published as a Reseach Article in *Scientific Reports* Journal

Cancino-Muñoz, I. et al. Development and application of affordable SNP typing approaches to genotype *Mycobacterium tuberculosis* complex strains in low and high burden countries. Sci Rep 9, 15343 (2019).

https://doi.org/10.1038/s41598-019-51326-2

6.6 Supplementary Data

6.6.1 Supplementary Tables

Sample ID	Source	MTBC lineage	LSP lineage	Spoligotyping family
N0067	Swiss TPH	L1	Indo-Oceanic	EAI
N0153	Swiss TPH	L1	Indo-Oceanic	EAI
N1068	Swiss TPH	L1	Indo-Oceanic	EAI
N0053	Swiss TPH	L2	East-Asian	Beijing
N0150	Swiss TPH	L2	East-Asian	Beijing
N1007	Swiss TPH	L3	East-African-Indian	CAS
N1022	Swiss TPH	L3	East-African-Indian	CAS
N1057	Swiss TPH	L4.1.1	Euro-American	X
N0148	Swiss TPH	L4.1.1	Euro-American	X
N0142	Swiss TPH	L4.1.1	Euro-American	X
G02	This study	L4.1.2	Euro-American	Haarlem
G287	This study	L4.1.2	Euro-American	Haarlem
G1010	This study	L4.1.2	Euro-American	Haarlem
N1204	Swiss TPH	L4.1.3	Euro-American	Ghana
G770	This study	L4.1.3	Euro-American	Ghana
N1263	Swiss TPH	L4.2	Euro-American	
G440	This study	L4.2	Euro-American	
G551	This study	L4.2	Euro-American	
G450	This study	L4.3	Euro-American	LAM
G186	This study	L4.3	Euro-American	LAM
G1068	This study	L4.3	Euro-American	LAM
G200	This study	L4.4	Euro-American	
G564	This study	L4.4	Euro-American	
N0163	Swiss TPH	L4.5	Euro-American	
N1277	Swiss TPH	L4.6.1	Euro-American	Uganda
N1207	Swiss TPH	L4.6.2	Euro-American	Cameroon
G630	This study	L4.6.2	Euro-American	Cameroon
G818	This study	L4.6.2	Euro-American	Cameroon
N1770	Swiss TPH	L4.10	Euro-American	
G109	This study	L4.10	Euro-American	
G280	This study	L4.10	Euro-American	
N1176	Swiss TPH	L5	West-African-1	AFRI2
N1063	Swiss TPH	L5	West-African-1	AFRI2
N1272	Swiss TPH	L5	West-African-1	AFRI2
N0091	Swiss TPH	L6	West-African-2	AFRI1
N1202	Swiss TPH	L6	West-African-2	AFRI1
N1177	Swiss TPH	L6	West-African-2	AFRI1
G578	This study	*M.bovis*	*M.bovis*	*M.bovis*
G513	This study	*M.bovis*	*M.bovis*	*M.bovis*
G1020	This study	*M.bovis*	*M.bovis*	*M.bovis*

Table 6.5: **Reference samples used in this study.** All isolates were previously characterized by whole-genome sequencing. Abbreviations: LSP, Long Sequence Polymorphisms; TPH, Tropical Health Institute.

MTBC lineage[a]	SNP position[b]	Nucleotide change[c]	Gene	Primer sequences 5’-3’	Amplicon size	Reference
L1	4357773	G/A	Rv3878/ Rv3879c	F-ACCCTCAACAACCACAACGT R-CGACACTACCGATCAGCGTT	386bp	This study
L2	4357804	T/G				
L5	4357657	G/A				
L3	1281984	G/A	Rv1155/ intergenic region	F-GATGGTCATACGCCGTTGCT R-CTCTTGCGGGGACTTCGATT	402bp	This study
L4	1281771	C/T				
L6	1281685	C/G				

[a] Nomenclature proposed by Comas et al [120].

[b] Genomic position on the H37Rv reference genome (NCBI, NC_000962.3).

[c] Allelic change in the reference genome.

Table 6.6: Specific primers used in Sanger sequencing molecular approach.

6.6.2 Supplementary Figures

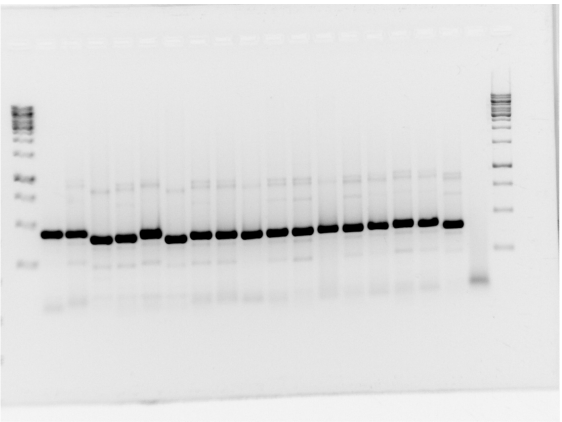


Figure 6.7: PCR products for the Sanger sequencing molecular assay (part 1). The figure shows the amplified products for Region 1 (specific markers for L1, L2 and L5). Three samples for each lineage were tested, except for L5, which only had 2 samples (wells 14-15). The amplicon size is 386bp. Well 19 was DNA-free and was considered as a negative control. The molecular weight-size ranged from 250-10,000bp.

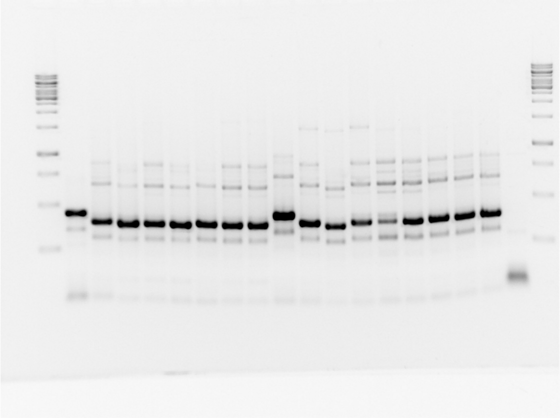


Figure 6.8: PCR products for the Sanger sequencing molecular assay (part 2). The figure shows the amplified products for Region 1 (specific markers for L3, L4 and L6). Three samples for each lineage were tested, except for L5, which only had 2 samples (wells 14-15). The amplicon size is 402bp. Well 19 was DNA-free and was considered as a negative control. The molecular weight-size ranged from 250-10,000bp.

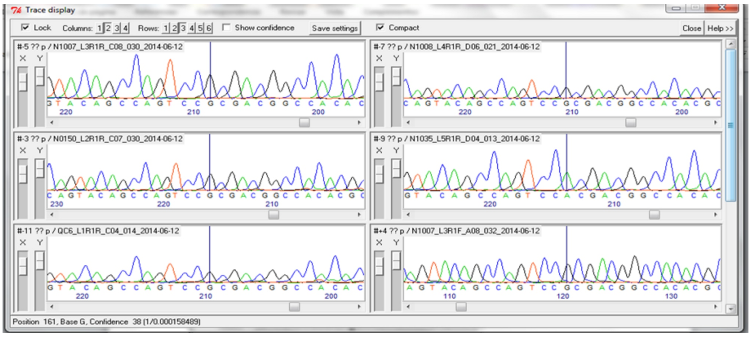

Trace display
☑ Lock
Columns: 1 2 3 4
Rows: 1 2 3 4 5 6
□ Show confidence
Save settings
☑ Compact
Close
Help >>
#5 ?? p / N1007_L3R1R_C08_030_2014-06-12
#7 ?? p / N1008_L4R1R_D06_021_2014-06-12
#3 ?? p / N0150_L2R1R_C07_030_2014-06-12
#9 ?? p / N1035_L5R1R_D04_013_2014-06-12
#11 ?? p / QC6_L1R1R_C04_014_2014-06-12
#+4 ?? p / N1007_L3R1F_A08_032_2014-06-12
Position 161, Base G, Confidence 38 (1/0.000158489)

Figure 6.9: Sanger sequencing amplified Region 1 to identify lineage 5. The chromatograms shows the specific region that contains the lineage 5 diagnostic SNP for different samples. The specific polymorphism is marked in blue. An adenine (represented in green) is detected instead of a guanine (represented in black) at the position 169. In this case, the sample N0135 harbors the diagnostic marker for lineage 5, while the rest presented the wild-type allele.

General discussion

Tuberculosis (TB) is considered one of the major causes of death worldwide. Thus, identification of latent individuals at risk of developing active disease, rapid diagnostics, as well as an efficient and early transmission detection, are essential to decrease its incidence. Novel technologies such as whole genome sequencing (WGS) are improving the identification of infectious diseases and the development of diagnostic tools with a theoretically ultimate resolution. In this thesis, we used WGS in order to obtain a genomic snapshot of all MTBC cases in a local region during a three year period. We used the isolates genomic information for several purposes: 1) to describe the bacteria population structure and transmission patterns; 2) to predict drug resistance phenotypes of isolates circulating in the region and 3) to highlight the importance of applying WGS to personalize treatment in TB cases of difficult management. As an in-depth discussion of each topic is presented in each chapter, in this general discussion we will focus on hot topics regarding the use of WGS at the epidemiological and diagnostic levels.

7.1 Comparison of transmission rates across settings based on WGS

In the third chapter, we present the genomic characterization of MTBC isolates in the Valencia Region over a three year period, 2014-2016. We applied WGS in 785 clinical isolates, described the structure of the bacilli population, carried out an epidemiological study and also we identified and delimited TB transmission. As expected, L4 is the most prevalent lineage since it is the most common and widespread lineage [40]. This is also in accordance with our chapter 6 results in which we genotype populations from Valencia and compared to other settings. In Valencia Region local-born TB individuals are the major contributors to the overall disease incidence, contrary to the situation in other low-burden regions, where the majority of TB cases are contributed by immigrants [149, 185]. At the same time it is known that in low-burden countries transmission is associated with local-born individuals. Thus, given the major contribution of local-born cases, it is not surprising that rates of genomic transmission are also higher in Valencia Region as compared to other TB settings (**Additional Tables 10.1-10.2**). To illustrate this point and put transmission in Valencia Region into context we used comparable datasets from different TB burden settings, such as The United Kingdom [27] and Malawi [151] publicly available. The United Kingdom reports a TB incidence of 8 per 100,000 population similar to Valencia Region but mainly contributed by immigrants (72%). On the other hand, Malawi is an endemic TB country which reports an incidence of 181 per 100,000 people, and a high HIV-positive TB coinfection rate (88 per 100,000 people). Using only their respective local-born individuals and up to a 12 SNP threshold to delineate transmission, we note that in Valencia Region as well as in Malawi, almost half of their respective local-born cases are involved in genomic transmission [47.4% in Valencia Region (216/456), 49.3% in Malawi (108/219)], independently of the TB burden. In contrast, only 32% (24/74) of UK-born TB patients are in transmission. Notably in **Figure 7.1** (see below), we observe several Spanish-born patients were between 15 and 50 SNPs of difference among them. In contrast not such cases exist in UK-born cases. As 15-50 SNPs reflect transmission events that occurred decades ago, our analysis suggests that in Valencia Region, in contrast to UK, older contagion events still contribute to TB cases today (**Figure 7.1**). Notably, Malawi shows a similar pattern than Valencia Region but more exacerbated. The data shows that to reach a situation as in UK efforts to halt transmission is key.

We hypothesize that the higher values of our local-born cases are the consequence of a continued transmission over past decades that was not halted and that now is reflected in the local TB incidence. On the contrary, successful past efforts on transmission control can be observed in the UK where only very recent transmission is contributing to local-born TB incidence and strains sharing ancestors 15-50 years before are absent. Data suggests that the UK situation is similar in The Netherlands, Canada, Germany and US where most cases are due to reactivation episodes in immigrants. The results suggest that control of TB in the Valencia Region and probably Spain is lagging behind with respect other low-burden countries, this is not only reflected in the different patterns of transmission but also in that the incidence in Spanish-born population (6.7 per 100,000 people) is still much higher than in UK-born population (3.5 per 100,000 people). Given that the actual TB control in Valencia Region is meeting the targets of WHO to reduce TB, it is also plausible that the efforts controlling transmission are efficient but more time is needed to reach the results observed in other low-burden countries such as the UK.

7.2 WGS to identify recent transmission in different settings

Another interesting and valuable lesson that we can extract from this analysis is about how to define genomic TB transmission. The term recent transmission is used to define those contagion events that occurred in a short period of time, typically revealed by contact investigations. A timeframe to define recent transmission is not unique and depends on the public health agency definitions but it is usually assumed that less than 2-5 years after contagion are cases of recent transmission. When using WGS this translates in a mean number of 0-5 SNP between two cases as the bacteria is assumed to accumulate 0.3-0-5 SNP/year. Higher thresholds has been used, for example 12 SNPs are well accepted a transmission cluster [152] but the disagreement with

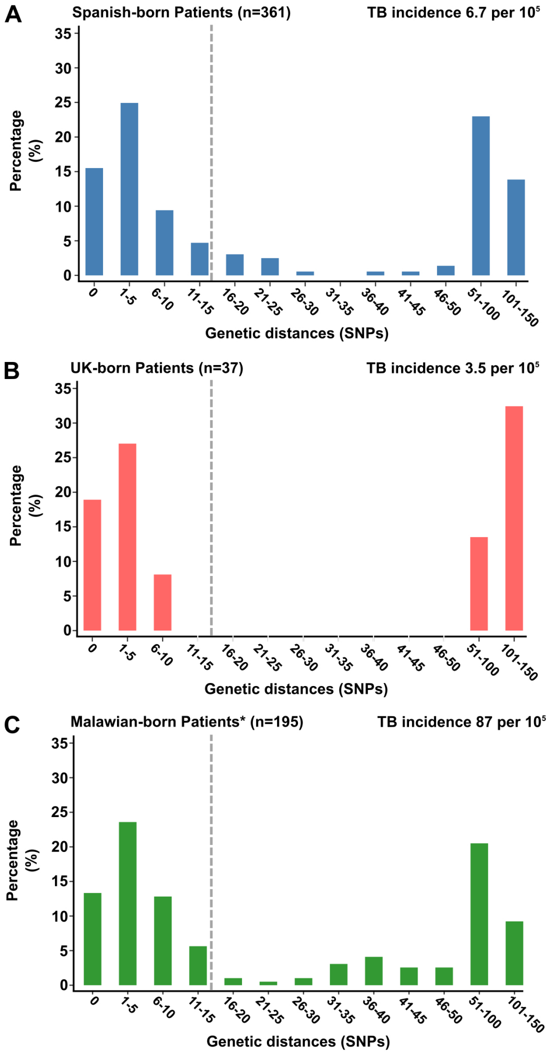


Figure 7.1: **Genetic distances between local-born patients in different TB-burden settings.** A) Pairwise distances between Spanish-born people (n= 361). B) Pairwise distances between UK-born people (n= 37). This dataset was used from Walker et al [27]. C) Pairwise distances between Malawi-born people during the 2008-2010 (n= 195). This dataset was used from Guerra-Assunção et al [151]. The grey dashed line separates the genomically related samples (clustered cases) from those that are not (unique cases). For plotting and representing purposes, we only used data up to 150 SNP threshold, which cleary means not transmission.

epidemiological investigations increases at those thresholds (this dissertation, [149]) and thus many transmission events by WGS cannot easily be validated. Thus, designing effective control measures is not only relevant to measure recent transmission but also the contribution of older transmission events. While in the UK (**Figure 7.1**) those older transmission events do not exist, in our study, but also in Malawi setting, genetic distances analysis shows a continuum of genetic distances not fitting a strict threshold like 12 SNPs and reflecting a continuum of transmission events. Thus in Valencia Region use of recent transmission SNP thresholds is useful to reveal transmissions that have happened very recently but miss the complete picture on how transmission is contributing to the yearly TB burden. As shown in Malawi, this situation may likely be common across epidemiological settings.

As expected, epidemiologically linked cases identified by public health are detected as very recent transmission events by WGS, defined by a 5 SNP threshold roughly equivalent to 5 years. As discussed above, this is related to the fact that epidemiological investigations look for recent contacts and not older transmissions. However even in that timeframe of 5 years, local investigations missed 60% of transmission cases (**Figure 3.2, chapter 3**). This suggests that while local investigations are very good at tracing some close contacts (family members, workplace) they missed many cases that probably occur outside the boundaries of the questionnaires. Some of these cases may be contacts not revealed by the index case because of the limitations of the epidemiological questionnaire. In that regard, investigations incorporating social contact networks have helped to identify additional transmission cases in some settings [175]. The application of these strategies could help to improve the transmission detection in Valencia Region.

But our data also suggests that transmision happens during casual contacts in the community or before symptoms (as shown in **chapter 4**). Identification of those individuals is more difficult. In those cases implementation of some sort of active case finding may help to identify transmitters. For example, by implementing targeted community-based screening in risk groups or large-scale screening in transmission hotspots as informed by WGS.

7.3 Transmission during subclinical disease

Controlling human-to-human transmission is key for achieving the targets of the End TB Strategy stated by the WHO [13]. Concerning this, it is mandatory to understand the complex dynamics of TB transmission and associated risk factors. Using WGS combined with epidemiological data, we investigated the dynamics of transmission within a fraction of the clusters previously identified in Valencia Region using TransPhylo, a bayesian-based phylogenetic modelling approach (**chapter 4**). These estimations included the probability of the index case being sampled, which of all clustered cases was the index case, and when a transmission event happened. Interestingly, we discovered that in some individuals, transmission occurs before symptoms onset, likely during subclinical disease. This striking result supports the idea of the existence of different infection status before developing active TB beyond the dichotomy latent/active TB [15]. We show for the first time that MTBC strains are transmissible during some of these newly recognized infection states. Nevertheless, more WGS-based studies are needed to evaluate the amount of subclinical transmission in different clinical settings.

These findings provide novel avenues to understand how TB is transmitted and the relationship of the pathogen and the host during infection that ultimately leads to transmission. We suggest that omic-based approaches that can link transmission timing with host biosignatures (for example using host transcriptomics [260]) have the potential to increase the identification TB cases at risk of transmitting the disease during the different stages of TB infection.

7.4 Active case finding to halt transmission

As we demonstrated through this dissertation, a great percentage of transmission (most of it related with local-born patients) is missed by local surveillance systems. Moreover, we showed that some TB cases likely transmit before developing symptoms. Therefore, novel diagnostics tools as well as new epidemiological interventions are needed in order to stop this autochthonous TB transmission. In addition, these approximations will reduce new infections and, hence, the TB incidence will decrease. Probably, the most important strategy with short-term effects is to change from passive to an active case finding. This strategy aims to find TB cases before the patient seeks for health care. In this way, community-based interventions include screening people in crowded areas searching for TB [261]. Many reports have been demonstrated that active case finding, increases the detection of TB in individuals with no typical symptoms, and smear negative status, in both, low-burden [31] and high-burden TB settings [262]. Despite the improvements of this active case finding intervention, studies regarding public health viability and cost-effectiveness as well as rapid screening tools are required. To cover this latter issue, recent development of molecular tests that are capable of detecting ultra-low level of *M.tuberculosis* DNA are helping to identify more cases than sputum microscopy [95, 263]. Furthermore, eight whole-blood transcriptional signatures have been shown to differentiate between latent, subclinical and active TB [22]. Given the low prevalence of TB disease in the general population, we need to find cost-effective strategies to do active case finding. In that regard the transmission picture revealed by WGS can help to design and guide epidemiological interventions.

7.5 Drug resistance prediction

Another important and essential intervention to control TB at individual-care and population levels is to stop drug resistance emergence and its transmission. Despite Valencia Region is not considered a high MDR setting, accurate and fast drug resistance prediction tools are essential for patient management. Recent studies have proved that DR prevalence can be accurately predicted by genomic surveillance [177], and especially to first-line drugs [137]. In **chapters 3 and 5** we used the bacterial genome sequence as a diagnostic for drug resistance. First, we demonstrated that WGS is a reliable prediction tool for detecting drug resistance, at least in our study region (**Table 3.5, chapter 3**). High agreement values of accuracy and specificity between WGS and phenotypic DST were estimated for first-line drugs. The few discrepant results that we identified could be due to either, phenotypic DST-related issues or by the presence of unknown mutations. In this regard, WGS-DST depends on the availability of catalogues with high-confidence causative mutations. In a recent meta-analysis, ethambutol, streptomycin and pyrazinamide have reported low sensitivities and specificities values, due sometimes to the lack of data regarding resistance genes and mutations for these drugs or to standardization problems on phenotypic-based assays [140]. Although there are web-based tools that contain several resistance mutations catalogues [193, 139, 97] there are still unknown DR variants for first- and particularly second-line and new anti-TB-drugs, however, new mutations are periodically reported [264, 186]. Despite these limitations, there is enough confidence to predict susceptibility to first-line drugs as demonstrated by a study involving 10,000 isolates from different parts of the world [137] which has led to some countries to phase-out phenotypic-DST and replace it by WGS. Nevertheless, phenotypic-DST is then focused in those cases for which there is a prediction of resistance or cases where inspection of the WGS is inconclusive.

7.6 Personalized treatment for TB based on WGS

In this thesis, we used WGS to discover and describe novel mutations related to isoniazid resistance (**chapter 5**). These variants were undetected by automated DST approaches and, thus, isolates harbouring these mutations were considered susceptible. It is well known that some of these variants conferred ‘low-level’ phenotypic resistance, resulting in an outcome difficult to interpret [84]. Moreover, we used WGS DR prediction in a prospective manner to guide TB treatment. Although the patient became XDR-TB, a close monitoring of the patient using the genomic, clinical and microbiological information led to physicians to successfully treat the patient. To our knowledge, this is one of the first cases that use WGS to personalize the treatment of a patient using genomic information. In the context of this patient this was especially relevant as DST tests for some first and second line drugs were inconclusive. These results highlight the capacity of WGS for fast and accurately predict drug resistance, at least in our study region. Additionally, it is also remarkable that WGS can solve complex or phenotypically undetermined cases, in this sense personalized treatments are imperative to save patients life. There is no doubt that personalizing treatment based on pathogen genomic data will help in the management of cases. This is really relevant for high-burden MDR-TB countries where more complicated management situations are found and where many times treatment is empirical as DST for second-line drugs is not as standardized as for first-line drugs. However, we need to expand the catalogues of mutations for second-line drugs. More importantly, the new recommended regimens by the WHO are based on all-oral drugs including drugs like bedaquiline, linezolid and delamanid [76]. Our knowledge about the genetic bases of resistance to those drugs is far from complete. Combined with the fact that there are no DST commercial tests favailable, the patient monitoring and follow-up to detect acquisition of resistance will be extremely challenging. In fact there is an increasing number of reports informing about the emergence of bedaquiline resistance both in personalized therapies [265] and under programmatic conditions [266, 267].

7.7 Strain classification

As was stated throughout this thesis, WGS is a valuable tool for identifying SNPs that could be further used for different applications in population studies, such as detecting and predicting transmission and DR. In addition, WGS can be used as a typing tool to classify MTBC strains into lineages. Although this technology has become cheaper over the last years, specific infrastructure as well as qualified personnel are needed, these conditions are not always available in low and middle-income regions. Nevertheless, the knowledge obtained from WGS could be translated to develop simpler laboratory methods in order to obtain specific and accurate information in a cheaper and faster manner. As a proof of concept, in **chapter 6**, two molecular tools were developed for classifying MTBC isolates into the main lineages, using routine PCR-based reagents and laboratory instruments. One of the aims was to have an alternative and affordable typing method that can be useful in resource-limited settings. In addition, we demonstrated that our approaches work with low DNA concentration, which is a common problem in routine clinical samples (eg, DNA obtained from heat-inactivated and sputum samples). Thus, a clinical collection of around 500 isolates from different TB incidence settings was characterized. As an example, our results were in agreement with those recently published [186]. Thanks to the low MTBC diversity and its clonal genomic features, these approaches are not limited only to the detection of phylogenetic SNPs, in fact, they can be adapted to other specific applications. For instance, SNPs panels related to DR are commonly used in routine clinical diagnostics. Markers with epidemiological interest (to identify a local TB outbreak or transcontinental spread) can be unmasked with these molecular techniques [268, 146, 166]. However, a previous notion is always required to proceed to WGS of representative cases and identify specific SNP positions.

7.8 Further considerations

From the analysis and results reported here, we can state that implementation of WGS in public health surveillance systems will improve drug resistance prediction and case management at the individual level. In addition, it will help to characterize TB transmission patterns accurately at the population level. However, its implementation requires the integration of other related disciplines such as microbiology, epidemiology and bioinformatics. Thus, a multidisciplinary team is needed involving personnel with knowledge regarding genomics and skills in analyzing big data. Currently, a few high-income countries are using WGS as a routine diagnostic tool, but it is estimated that more countries will adopt it in the next few years. In our local setting, many clinicians and epidemiologists are conscious about the use and applications of WGS in TB field. In addition, we have worked with them on many occasions in order to exploit the most relevant clinical and epidemiological results from our genomic analysis. In Valencia Region, its integration to the local surveillance system does not seem very far.

Finally, the results presented have been only possible by the invaluable contribution of many actors of the Valencia Region health system. A total of 18 clinical microbiology units from hospitals in the region have generously contributed to the study in what represents, one of the best examples of multicenter studies in infectious diseases in the country. Public health officials and particularly those devoted to the control of TB in the region have provided the invaluable demographic and epidemiological information needed to reach the conclusions exposed. Thanks to this coordinated effort, we described and proved how this technology complements the current approaches used in Valencia Region to handle the disease. The implementation of WGS as a supplementary tool for the diagnosis and epidemiology of TB in Valencia Region is already helping local TB control and, we think can serve as a template for using WGS into other infectious diseases in the local public health surveillance system.

Conclusions

- Tuberculosis transmission burden in Valencia Region is high based on WGS compared to other low-burden countries like The United Kingdom. Risk factors of genomic transmission in Valencia Region are being born in Spain (associated), young age (associated) and older age (non-associated).
- Although current TB epidemiological investigations (contact tracing) maintain a low TB incidence, they underestimate the real TB transmission burden.
- Our analysis demonstrated that WGS could be an alternative and trustworthy tool for detecting both recent and older TB transmission events. Furthermore, it could become a transformative methodology for the public health surveillance system.
- Sensitivity and specificity values demonstrated that WGS is an accurate and reliable tool for predicting drug resistance.
- WGS-based phylogenetic modelling combined with epidemiological data allows to infer high likely transmitters and index cases within a transmission cluster. However, its application is limited by the null diversity observed many times between clustered cases.
- In some individuals, TB transmission can occur during subclinical disease and can jeopardize the progress of passive case finding and contact

tracing. Together with emerging data from different fields, controlling subclinical disease through active case finding approaches will become relevant in the near future.

- Real-time WGS can be used to detect drug resistance variants during TB treatment. More importantly, it helps to identify uncommon mutations that arise through time.
- The PCR-based SNP-typing molecular approaches give an alternative rapid tool for classification of MTBC strains and it is a method that could be translated to multiple TB applications.

Bibliography

[1] Gagneux S. Ecology and evolution of mycobacterium tuberculosis: **16**(4):202–213. ISSN 1740-1526, 1740-1534. doi:10.1038/nrmicro.2018.8.

[2] Brites D, Loiseau C, Menardo F, Borrell S, Boniotti MB, Warren R, Dippenaar A, Parsons SDC, Beisel C, Behr MA *et al.* A new phylogenetic framework for the animal-adapted mycobacterium tuberculosis complex: **9**:2820. ISSN 1664-302X. doi:10.3389/fmicb.2018.02820.

[3] Fedrizzi T, Meehan CJ, Grottola A, Giacobazzi E, Fregni Serpini G, Tagliazucchi S, Fabio A, Bettua C, Bertorelli R, De Sanctis V *et al.* Genomic characterization of nontuberculous mycobacteria: **7**(1):45258. ISSN 2045-2322. doi:10.1038/srep45258.

[4] Porvaznik I, Solovič I and Mokrý J. Non-tuberculous mycobacteria: Classification, diagnostics, and therapy. In Pokorski M, ed., *Respiratory Treatment and Prevention*, pages 19–25. Springer International Publishing. ISBN 978-3-319-44488-8: doi:10.1007/5584_2016_45.

[5] Yeung MW, Khoo E, Brode SK, Jamieson FB, Kamiya H, Kwong JC, Macdonald L, Marras TK, Morimoto K and Sander B. Health-related quality of life, comorbidities and mortality in pulmonary nontuberculous mycobacterial infections: A systematic review: Health outcomes in NTM: **21**(6):1015–1025. ISSN 13237799. doi:10.1111/resp.12767.

[6] Prevots DR, Loddenkemper R, Sotgiu G and Migliori G. Nontuberculous mycobacterial pulmonary disease: an increasing burden with substantial costs: **49**(4):1700374. ISSN 0903-1936, 1399-3003. doi:10.1183/13993003.00374-2017.

[7] Cook GM, Berney M, Gebhard S, Heinemann M, Cox RA, Danilchanka O and Niederweis M. Physiology of mycobacteria. In *Advances in Microbial Physiology*, volume 55, pages 81–319. Elsevier. ISBN 978-0-12-374790-7: doi:10.1016/S0065-2911(09)05502-7.

[8] Manual of clinical microbiology, 10th edition.

[9] Jarlier V and Nikaido H. Mycobacterial cell wall: Structure and role in natural resistance to antibiotics: **123**(1):11–18. doi:10.1111/j.1574-6968.1994.tb07194.x.

[10] WORLD HEALTH ORGANIZATION. *GLOBAL TUBERCULOSIS REPORT 2019.* WORLD HEALTH ORGANIZATION. ISBN 978-92-4-156571-4. OCLC: 1132243244.

[11] WHO Regional Office for Europe/European Centre for Disease Prevention and Control. *Tuberculosis surveillance and monitoring in Europe 2019 - 2017 data*. ISBN 978-92-9498-297-1.

[12] MINISTERIO DE SANIDAD, CONSUMO Y BIENESTAR SOCIAL. PLAN PARA LA PREVENCIÓN y CONTROL DE LA TUBERCULOSIS EN ESPAÑA.

[13] World Health Organization. *IMPLEMENTING THE END TB STRATEGY: THE ESSENTIALS*. WORLD HEALTH ORGANIZATION. ISBN 978-92-4-150993-0.

[14] Pai M, Behr MA, Dowdy D, Dheda K, Divangahi M, Boehme CC, Ginsberg A, Swaminathan S, Spigelman M, Getahun H *et al.* Tuberculosis: **2**(1):16076. ISSN 2056-676X. doi:10.1038/nrdp.2016.76.

[15] Lin PL and Flynn JL. The end of the binary era: Revisiting the spectrum of tuberculosis: **201**(9):2541–2548. ISSN 0022-1767, 1550-6606. doi:10.4049/jimmunol.1800993.

[16] Zumla A, Raviglione M, Hafner R and Fordham von Reyn C. Tuberculosis: **368**(8):745–755. ISSN 0028-4793, 1533-4406. doi:10.1056/NEJMra1200894.

[17] Lin PL, Maiello P, Gideon HP, Coleman MT, Cadena AM, Rodgers MA, Gregg R, O'Malley M, Tomko J, Fillmore D *et al.* PET CT identifies reactivation risk in cynomolgus macaques with latent m. tuberculosis: **12**(7):e1005739. ISSN 1553-7374. doi:10.1371/journal.ppat.1005739.

[18] Barry CE, Boshoff HI, Dartois V, Dick T, Ehrt S, Flynn J, Schnappinger D, Wilkinson RJ and Young D. The spectrum of latent tuberculosis: rethinking the biology and intervention strategies: **7**(12):845–855. ISSN 1740-1526, 1740-1534. doi:10.1038/nrmicro2236.

[19] Drain PK, Bajema KL, Dowdy D, Dheda K, Naidoo K, Schumacher SG, Ma S, Meermeier E, Lewinsohn DM and Sherman DR. Incipient and subclinical tuberculosis: a clinical review of early stages and progression of infection: **31**(4):e00021–18. ISSN 0893-8512, 1098-6618. doi:10.1128/CMR.00021-18.

[20] Singhania A, Verma R, Graham CM, Lee J, Tran T, Richardson M, Lecine P, Leissner P, Berry MPR, Wilkinson RJ *et al.* A modular transcriptional signature identifies phenotypic heterogeneity of human tuberculosis infection: **9**(1):2308. ISSN 2041-1723. doi:10.1038/s41467-018-04579-w.

[21] Esmail H, Lai RP, Lesosky M, Wilkinson KA, Graham CM, Coussens AK, Oni T, Warwick JM, Said-Hartley Q, Koegelenberg CF *et al.* Characterization of progressive HIV-associated tuberculosis using 2-deoxy-2-[18 f] fluoro-d-glucose positron emission and computed tomography: **22**(10):1090.

[22] Gupta RK, Turner CT, Venturini C, Esmail H, Rangaka MX, Copas A, Lipman M, Abubakar I and Noursadeghi M. Concise whole blood transcriptional signatures for incipient tuberculosis: a systematic review and patient-level pooled meta-analysis:

page S2213260019302826. ISSN 22132600. doi:10.1016/S2213-2600(19)30282-6.

[23] Stein CM, Zalwango S, Malone LL, Thiel B, Mupere E, Nsereko M, Okware B, Kisingo H, Lancioni CL, Bark CM *et al.* Resistance and susceptibility to mycobacterium tuberculosis infection and disease in tuberculosis households in kampala, uganda: **187**(7):1477–1489. ISSN 0002-9262, 1476-6256. doi:10.1093/aje/kwx380.

[24] Williams CM, Abdulwhhab M, Birring SS, De Kock E, Garton NJ, Townsend E, Pareek M, Al-Taie A, Pan J, Ganatra R *et al.* Exhaled mycobacterium tuberculosis output and detection of subclinical disease by face-mask sampling: prospective observational studies: page S1473309919307078. ISSN 14733099. doi:10.1016/S1473-3099(19)30707-8.

[25] Balcells ME, Thomas SL, Godfrey-Faussett P and Grant AD. Isoniazid preventive therapy and risk for resistant tuberculosis: **12**(5):744–751. ISSN 1080-6040, 1080-6059. doi:10.3201/eid1205.050681.

[26] Yates TA, Khan PY, Knight GM, Taylor JG, McHugh TD, Lipman M, White RG, Cohen T, Cobelens FG, Wood R *et al.* The transmission of mycobacterium tuberculosis in high burden settings: **16**(2):227–238. ISSN 14733099. doi:10.1016/S1473-3099(15)00499-5.

[27] Walker TM, Lalor MK, Broda A, Ortega LS, Morgan M, Parker L, Churchill S, Bennett K, Golubchik T, Giess AP *et al.* Assessment of mycobacterium tuberculosis transmission in oxfordshire, UK, 2007-12, with whole pathogen genome sequences: an observational study: **2**(4):285–292. ISSN 2213-2619. doi:10.1016/S2213-2600(14)70027-X.

[28] Lee RS, Radomski N, Proulx JF, Levade I, Shapiro BJ, McIntosh F, Soualhine H, Menzies D and Behr MA. Population genomics of *Mycobacterium tuberculosis* in the inuit: **112**(44):13609–13614. ISSN 0027-8424, 1091-6490. doi:10.1073/pnas.1507071112.

[29] Jones-López EC, Acuña-Villaorduña C, Ssebidandi M, Gaeddert M, Kubiak RW, Ayakaka I, White LF, Joloba M, Okwera A and Fennelly KP. Cough aerosols ofMycobacterium tuberculosisin the prediction of incident tuberculosis disease in household contacts: **63**(1):10–20. doi:10.1093/cid/ciw199.

[30] Acuña-Villaorduña C, White LF, Fennelly KP and Jones-López EC. Tuberculosis transmission: sputum vs aerosols: **16**(7):770–771. ISSN 14733099. doi:10.1016/S1473-3099(16)30075-5.

[31] Behr M, Warren S, Salamon H, Hopewell P, de Leon AP, Daley C and Small P. Transmission of mycobacterium tuberculosis from patients smear-negative for acid-fast bacilli: **353**(9151):444–449. ISSN 01406736. doi:10.1016/S0140-6736(98)03406-0.

[32] Lin HH, Ezzati M and Murray M. Tobacco smoke, indoor air pollution and tuberculosis: A systematic review and meta-analysis: **4**(1):e20. ISSN 1549-1676. doi:10.1371/journal.pmed.0040020.

[33] Lönnroth K, Williams BG, Stadlin S, Jaramillo E and Dye C. Alcohol use as a risk factor for tuberculosis – a systematic review: **8**(1):289. ISSN 1471-2458. doi: 10.1186/1471-2458-8-289.

[34] Menon S, Rossi R, Nshimyumukiza L, Wusiman A, Zdraveska N and Eldin MS. Convergence of a diabetes mellitus, protein energy malnutrition, and TB epidemic: the neglected elderly population: **16**(1):361. ISSN 1471-2334. doi: 10.1186/s12879-016-1718-5.

[35] Corbett EL, Steketee RW, ter Kuile FO, Latif AS, Kamali A and Hayes RJ. HIV-1/AIDS and the control of other infectious diseases in africa: **359**(9324):2177–2187. ISSN 0140-6736. doi:10.1016/S0140-6736(02)09095-5.

[36] Jeon CY and Murray MB. Diabetes mellitus increases the risk of active tuberculosis: A systematic review of 13 observational studies: **5**(7):e152. ISSN 1549-1676. doi:10.1371/journal.pmed.0050152.

[37] Cheng S, Chen W, Yang Y, Chu P, Liu X, Zhao M, Tan W, Xu L, Wu Q, Guan H *et al.* Effect of diagnostic and treatment delay on the risk of tuberculosis transmission in shenzhen, china: An observational cohort study, 1993–2010: **8**(6):e67516. ISSN 1932-6203. doi:10.1371/journal.pone.0067516.

[38] Storla DG, Yimer S and Bjune GA. A systematic review of delay in the diagnosis and treatment of tuberculosis: **8**(1):15. ISSN 1471-2458. doi:10.1186/1471-2458-8-15.

[39] Gagneux S, DeRiemer K, Van T, Kato-Maeda M, de Jong BC, Narayanan S, Nicol M, Niemann S, Kremer K, Gutierrez MC *et al.* Variable host-pathogen compatibility in mycobacterium tuberculosis: **103**(8):2869–2873. ISSN 0027-8424, 1091-6490. doi:10.1073/pnas.0511240103.

[40] Stucki D, Brites D, Jeljeli L, Coscolla M, Liu Q, Trauner A, Fenner L, Rutaihwa L, Borrell S, Luo T *et al.* Mycobacterium tuberculosis lineage comprises globally distributed and geographically restricted sublineages.: **48**:1535–1543.

[41] Mathema B, Andrews JR, Cohen T, Borgdorff MW, Behr M, Glynn JR, Rustomjee R, Silk BJ and Wood R. Drivers of tuberculosis transmission: **216**:S644–S653. ISSN 0022-1899, 1537-6613. doi:10.1093/infdis/jix354.

[42] Fox GJ, Nhung NV, Sy DN, Hoa NL, Anh LT, Anh NT, Hoa NB, Dung NH, Buu TN, Loi NT *et al.* Household-contact investigation for detection of tuberculosis in vietnam: **378**(3):221–229. ISSN 0028-4793, 1533-4406. doi:10.1056/NEJMoa1700209.

[43] Dowdy DW, Basu S and Andrews JR. Is passive diagnosis enough?: The impact of subclinical disease on diagnostic strategies for tuberculosis: **187**(5):543–551. ISSN 1073-449X, 1535-4970. doi:10.1164/rccm.201207-1217OC.

[44] Begun M, Newall AT, Marks GB and Wood JG. Contact tracing of tuberculosis: A systematic review of transmission modelling studies: **8**(9):e72470. ISSN 1932-6203. doi:10.1371/journal.pone.0072470.

[45] Reichler MR, Reves R, Bur S, Thompson V, Mangura BT, Ford J, Valway SE and Onorato IM. Evaluation of investigations conducted to detect and prevent transmission of tuberculosis: **287**(8):991–995. ISSN 0098-7484. doi:10.1001/jama.287.8.991. Publisher: American Medical Association.

[46] Fox GJ, Barry SE, Britton WJ and Marks GB. Contact investigation for tuberculosis: a systematic review and meta-analysis: **41**(1):140–156. ISSN 0903-1936. doi:10.1183/09031936.00070812.

[47] Control and prevention of tuberculosis in the united kingdom: code of practice 2000. joint tuberculosis committee of the british thoracic society: **55**(11):887–901. ISSN 0040-6376. doi:10.1136/thorax.55.11.887.

[48] Hwang TJ, Ottmani S and Uplekar M. A rapid assessment of prevailing policies on tuberculosis contact investigation: **15**(12):1620–1623. ISSN 1815-7920. doi:10.5588/ijtld.11.0222.

[49] Hanrahan CF, Nonyane BAS, Mmolawa L, West NS, Siwelana T, Lebina L, Martinson N and Dowdy DW. Contact tracing versus facility-based screening for active TB case finding in rural south africa: A pragmatic cluster-randomized trial (kharitode TB): **16**(4):e1002796. ISSN 1549-1676. doi:10.1371/journal.pmed.1002796.

[50] Kasaie P, Mathema B, Kelton WD, Azman AS, Pennington J and Dowdy DW. A novel tool improves existing estimates of recent tuberculosis transmission in settings of sparse data collection: **10**(12):e0144137. ISSN 1932-6203. doi:10.1371/journal.pone.0144137.

[51] Wobudeya E, Lukoye D, Lubega IR, Mugabe F, Sekadde M and Musoke P. Epidemiology of tuberculosis in children in kampala district, uganda, 2009–2010; a retrospective cross-sectional study: **15**(1):967. ISSN 1471-2458. doi:10.1186/s12889-015-2312-2.

[52] ValenS, Scaini JLR, Abileira FS, Gonves CV, von Groll A and Silva PEA. Prevalence of tuberculosis in prisons: risk factors and molecular epidemiology: **19**(10):1182–1187. ISSN 1815-7920. doi:10.5588/ijtld.15.0126.

[53] García De Viedma D and Pérez-Lago L. The evolution of genotyping strategies to detect, analyze, and control transmission of tuberculosis: **6**(5). ISSN 2165-0497. doi:10.1128/microbiolspec.MTBP-0002-2016.

[54] Roetzer A, Diel R, Kohl TA, Rückert C, Nübel U, Blom J, Wirth T, Jaenicke S, Schuback S, Rüsch-Gerdes S *et al.* Whole genome sequencing versus traditional genotyping for investigation of a mycobacterium tuberculosis outbreak: A longitudinal molecular epidemiological study: **10**(2):e1001387. ISSN 1549-1676. doi:10.1371/journal.pmed.1001387.

[55] Wyllie DH, Davidson JA, Grace Smith E, Rathod P, Crook DW, Peto TE, Robinson E, Walker T and Campbell C. A quantitative evaluation of MIRU-VNTR typing against whole-genome sequencing for identifying mycobacterium tuberculosis transmission: A prospective observational cohort study: **34**:122–130. ISSN 23523964. doi:10.1016/j.ebiom.2018.07.019.

[56] Meehan CJ, Goig GA, Kohl TA, Verboven L, Dippenaar A, Ezewudo M, Farhat MR, Guthrie JL, Laukens K, Miotto P *et al.* Whole genome sequencing of mycobacterium tuberculosis: current standards and open issues: **17**(9):533–545. ISSN 1740-1526, 1740-1534. doi:10.1038/s41579-019-0214-5.

[57] Nahid P, Dorman SE, Alipanah N, Barry PM, Brozek JL, Cattamanchi A, Chaisson LH, Chaisson RE, Daley CL, Grzemska M *et al.* Executive summary: Official american thoracic society/centers for disease control and prevention/infectious diseases society of america clinical practice guidelines: Treatment of drug-susceptible tuberculosis: **63**(7):853–867. ISSN 1058-4838, 1537-6591. doi:10.1093/cid/ciw566.

[58] Tiberi S, du Plessis N, Walzl G, Vjecha MJ, Rao M, Ntoumi F, Mfinanga S, Kapata N, Mwaba P, McHugh TD *et al.* Tuberculosis: progress and advances in development of new drugs, treatment regimens, and host-directed therapies: **18**(7):e183–e198. ISSN 14733099. doi:10.1016/S1473-3099(18)30110-5.

[59] Consortium TTT. Rifapentine and isoniazid once a week versus rifampicin and isoniazid twice a week for treatment of drug-susceptible pulmonary tuberculosis in HIV-negative patients: a randomised clinical trial: **360**(9332):528–534. ISSN 01406736. doi:10.1016/S0140-6736(02)09742-8.

[60] Gillespie SH, Crook AM, McHugh TD, Mendel CM, Meredith SK, Murray SR, Pappas F, Phillips PP and Nunn AJ. Four-month moxifloxacin-based regimens for drug-sensitive tuberculosis: **371**(17):1577–1587. ISSN 0028-4793, 1533-4406. doi:10.1056/NEJMoa1407426.

[61] Weiner M, Burman W, Vernon A, Benator D, Peloquin CA, Khan A, Weis S, King B, Shah N, Hodge T *et al.* Low isoniazid concentrations and outcome of tuberculosis treatment with once-weekly isoniazid and rifapentine: **167**(10):1341–1347. ISSN 1073-449X, 1535-4970. doi:10.1164/rccm.200208-951OC.

[62] Goldstein BP. Resistance to rifampicin: a review: **67**(9):625–630. ISSN 0021-8820, 1881-1469. doi:10.1038/ja.2014.107.

[63] Vall-Spinosa A, Lester W, Moulding T, Davidson PT and McClatchy JK. Rifampin in the treatment of drug-resistant mycobacterium tuberculosis infections: **283**(12):616–621. ISSN 0028-4793. doi:10.1056/NEJM197009172831202. Publisher: Massachusetts Medical Society.

[64] Mitchison DA and Nunn AJ. Influence of initial drug resistance on the response to short-course chemotherapy of pulmonary tuberculosis: **133**(3):423–430. ISSN 0003-0805. doi:10.1164/arrd.1986.133.3.423.

[65] Telenti A, Imboden P, Marchesi F, Matter L, Schopfer K, Bodmer T, Lowrie D, Colston M and Cole S. Detection of rifampicin-resistance mutations in mycobacterium tuberculosis: **341**(8846):647–651. ISSN 01406736. doi:10.1016/0140-6736(93)90417-F.

[66] Van Deun A, Aung KJM, Hossain A, de Rijk P, Gumusboga M, Rigouts L and de Jong BC. Disputed rpoB mutations can frequently cause important rifampicin resistance among new tuberculosis patients: **19**(2):185–190. ISSN 1815-7920. doi:10.5588/ijtld.14.0651.

[67] Stagg HR, Lipman MC, McHugh TD and Jenkins HE. Isoniazid-resistant tuberculosis: a cause for concern?: **21**(2):129–139. ISSN 1815-7920. doi:10.5588/ijtld.16.0716.

[68] Zhang Y, Heym B, Allen B, Young D and Cole S. The catalase-peroxidase gene and isoniazid resistance of mycobacterium tuberculosis: **358**:591–593. doi:10.1038/358591a0.

[69] Seifert M, Catanzaro D, Catanzaro A and Rodwell TC. Genetic mutations associated with isoniazid resistance in mycobacterium tuberculosis: A systematic review: **10**(3):e0119628. ISSN 1932-6203. doi:10.1371/journal.pone.0119628.

[70] Sherman DR, Mdluli K, Hickey MJ, Arain TM, Morris SL, Barry CE 3rd and Stover CK. Compensatory ahpC gene expression in isoniazid-resistant mycobacterium tuberculosis: **272**(5268):1641–1643. ISSN 0036-8075. doi:10.1126/science.272.5268.1641.

[71] Telenti A, Philipp WJ, Sreevatsan S, Bernasconi C, Stockbauer KE, Wieles B, Musser JM and Jacobs WR. The emb operon, a gene cluster of mycobacterium tuberculosis involved in resistance to ethambutol: **3**(5):567–570. ISSN 1078-8956, 1546-170X. doi:10.1038/nm0597-567.

[72] Safi H, Lingaraju S, Amin A, Kim S, Jones M, Holmes M, McNeil M, Peterson SN, Chatterjee D, Fleischmann R *et al.* Evolution of high-level ethambutol-resistant tuberculosis through interacting mutations in decaprenylphosphoryl--d-arabinose biosynthetic and utilization pathway genes: **45**(10):1190–1197. ISSN 1061-4036, 1546-1718. doi:10.1038/ng.2743.

[73] Zhang Y and Yew WW. Mechanisms of drug resistance in mycobacterium tuberculosis: update 2015: **19**(11):1276–1289. ISSN 1815-7920. doi:10.5588/ijtld.15.0389.

[74] Lange C, Abubakar I, Alffenaar JWC, Bothamley G, Caminero JA, Carvalho ACC, Chang KC, Codecasa L, Correia A, Crudu V *et al.* Management of patients with multidrug-resistant/extensively drug-resistant tuberculosis in europe: a TBNET consensus statement: **44**(1):23–63. ISSN 0903-1936, 1399-3003. doi:10.1183/09031936.00188313.

[75] Miotto P, Cabibbe AM, Feuerriegel S, Casali N, Drobniewski F, Rodionova Y, Bakonyte D, Stakenas P, Pimkina E, Augustynowicz-Kopeć E *et al.* Mycobacterium tuberculosis pyrazinamide resistance determinants: a multicenter study: **5**(5):e01819–14. ISSN 2150-7511. doi:10.1128/mBio.01819-14.

[76] World Health Organization. *Rapid Communication: Key changes to the treatment of durg-resistant tuberculosis*. WORLD HEALTH ORGANIZATION.

[77] Fitzpatrick C and Floyd K. A systematic review of the cost and cost effectiveness of treatment for multidrug-resistant tuberculosis:: **30**(1):63–80. ISSN 1170-7690. doi:10.2165/11595340-000000000-00000.

[78] Seddon JA, Thee S, Jacobs K, Ebrahim A, Hesseling AC and Schaaf HS. Hearing loss in children treated for multidrug-resistant tuberculosis: **66**(4):320–329. ISSN 01634453. doi:10.1016/j.jinf.2012.09.002.

[79] Pontali E, Sotgiu G, Tiberi S, D'Ambrosio L, Centis R and Migliori GB. Cardiac safety of bedaquiline: a systematic and critical analysis of the evidence: **50**(5):1701462. ISSN 0903-1936, 1399-3003. doi:10.1183/13993003.01462-2017.

[80] Crofton J and Mitchison DA. Streptomycin resistance in pulmonary tuberculosis: **2**(4588):1009. doi:10.1136/bmj.2.4588.1009.

[81] Canetti G, Fox W, Khomenko Aa, Mahler HT, Menon NK, Mitchison DA, Rist N and Šmelev NA. Advances in techniques of testing mycobacterial drug sensitivity, and the use of sensitivity tests in tuberculosis control programmes: **41**(1):21.

[82] Schön T, Miotto P, Köser C, Viveiros M, Böttger E and Cambau E. Mycobacterium tuberculosis drug-resistance testing: challenges, recent developments and perspectives: **23**(3):154–160. ISSN 1198743X. doi:10.1016/j.cmi.2016.10.022.

[83] Mitchison DA. Drug resistance in tuberculosis: **25**(2):376–379. ISSN 0903-1936, 1399-3003. doi:10.1183/09031936.05.00075704.

[84] Miotto P, Cabibbe AM, Borroni E, Degano M and Cirillo DM. Role of disputed mutations in the *rpoB* gene in interpretation of automated liquid MGIT culture results for rifampin susceptibility testing of *Mycobacterium tuberculosis*: **56**(5):e01599–17, /jcm/56/5/e01599–17.atom. ISSN 0095-1137, 1098-660X. doi: 10.1128/JCM.01599-17.

[85] Sanchez-Padilla E, Merker M, Beckert P, Jochims F, Dlamini T, Kahn P, Bonnet M and Niemann S. Detection of drug-resistant tuberculosis by xpert MTB/RIF in swaziland: **372**(12):1181–1182. ISSN 0028-4793, 1533-4406. doi:10.1056/NEJMc1413930.

[86] Makhado NA, Matabane E, Faccin M, Pinçon C, Jouet A, Boutachkourt F, Goeminne L, Gaudin C, Maphalala G, Beckert P *et al.* Outbreak of multidrug-resistant tuberculosis in south africa undetected by WHO-endorsed commercial tests: an observational study: **18**(12):1350–1359. ISSN 14733099. doi:10.1016/S1473-3099(18)30496-1.

[87] Schon T, Jureen P, Giske CG, Chryssanthou E, Sturegard E, Werngren J, Kahlmeter G, Hoffner SE and Angeby KA. Evaluation of wild-type MIC distributions as a tool for determination of clinical breakpoints for mycobacterium tuberculosis: **64**(4):786–793. ISSN 0305-7453, 1460-2091. doi:10.1093/jac/dkp262.

[88] Ängeby K, Juréen P, Kahlmeter G, Hoffner S and Schön T. Challenging a dogma: antimicrobial susceptibility testing breakpoints for mycobacterium tuberculosis: **90**(9):693–698. ISSN 00429686. doi:10.2471/BLT.11.096644.

[89] Comas I and Gagneux S. The past and future of tuberculosis research: **5**(10):e1000600. ISSN 1553-7374. doi:10.1371/journal.ppat.1000600.

[90] Eddabra R and Ait Benhassou H. Rapid molecular assays for detection of tuberculosis: **10**(1):4. ISSN 2200-6133. doi:10.1186/s41479-018-0049-2.

[91] World Health Organization. *The use of molecular line probe assays for the detection of resistance to isoniazid and rifampicin*. WORLD HEALTH ORGANIZATION. ISBN 978-92-4-151126-1.

[92] Ling DI, Zwerling AA and Pai M. GenoType MTBDR assays for the diagnosis of multidrug-resistant tuberculosis: a meta-analysis: **32**(5):1165–1174. ISSN 0903-1936, 1399-3003. doi:10.1183/09031936.00061808.

[93] Denkinger CM, Schumacher SG, Boehme CC, Dendukuri N, Pai M and Steingart KR. Xpert MTB/RIF assay for the diagnosis of extrapulmonary tuberculosis: a systematic review and meta-analysis: **44**(2):435–446. ISSN 0903-1936, 1399-3003. doi:10.1183/09031936.00007814.

[94] World Health Organization. *Rapid Communication: Molecular assays as initial tests for the diagnosis of tuberculosis and rifampicin resistance*. WORLD HEALTH ORGANIZATION.

[95] Dorman SE, Schumacher SG, Alland D, Nabeta P, Armstrong DT, King B, Hall SL, Chakravorty S, Cirillo DM, Tukvadze N *et al.* Xpert MTB/RIF ultra for detection of mycobacterium tuberculosis and rifampicin resistance: a prospective multicentre diagnostic accuracy study: **18**(1):76–84. ISSN 14733099. doi:10.1016/S1473-3099(17)30691-6.

[96] Ahmad S, Mokaddas E, Al-Mutairi N, Eldeen HS and Mohammadi S. Discordance across phenotypic and molecular methods for drug susceptibility testing of drug-resistant mycobacterium tuberculosis isolates in a low TB incidence country: **11**(4):e0153563. ISSN 1932-6203. doi:10.1371/journal.pone.0153563.

[97] Miotto P, Tessema B, Tagliani E, Chindelevitch L, Starks A, Emerson C, Hanna D, Kim PS, Liwski R, Zignol M *et al.* A standardised method for interpreting the association between mutations and phenotypic drug resistance in *Mycobacterium tuberculosis*: **50**(6):1701354. ISSN 0903-1936, 1399-3003. doi:10.1183/13993003.01354-2017.

[98] Köser CU, Ellington MJ, Cartwright EJP, Gillespie SH, Brown NM, Farrington M, Holden MTG, Dougan G, Bentley SD, Parkhill J *et al.* Routine use of microbial whole genome sequencing in diagnostic and public health microbiology: **8**(8):e1002824. ISSN 1553-7374. doi:10.1371/journal.ppat.1002824.

[99] Jackson BR, Tarr C, Strain E, Jackson KA, Conrad A, Carleton H, Katz LS, Stroika S, Gould LH, Mody RK *et al.* Implementation of nationwide real-time whole-genome sequencing to enhance listeriosis outbreak detection and investigation: **63**(3):380–386. ISSN 1058-4838, 1537-6591. doi:10.1093/cid/ciw242.

[100] Sánchez-Busó L, Comas I, Jorques G and González-Candelas F. Recombination drives genome evolution in outbreak-related legionella pneumophila isolates: **46**(11):1205–1211. ISSN 1061-4036, 1546-1718. doi:10.1038/ng.3114.

[101] Harris SR, Cole MJ, Spiteri G, Sánchez-Busó L, Golparian D, Jacobsson S, Goater R, Abudahab K, Yeats CA, Bercot B *et al.* Public health surveillance of multidrug-resistant clones of neisseria gonorrhoeae in europe: a genomic survey: **18**(7):758–768. ISSN 14733099. doi:10.1016/S1473-3099(18)30225-1.

[102] Stoesser N, Batty EM, Eyre DW, Morgan M, Wyllie DH, Del Ojo Elias C, Johnson JR, Walker AS, Peto TEA and Crook DW. Predicting antimicrobial susceptibilities for escherichia coli and klebsiella pneumoniae isolates using whole genomic

sequence data: **68**(10):2234–2244. ISSN 1460-2091, 0305-7453. doi:10.1093/jac/dkt180.

[103] Bradley P, Gordon NC, Walker TM, Dunn L, Heys S, Huang B, Earle S, Pankhurst LJ, Anson L, de Cesare M *et al.* Rapid antibiotic-resistance predictions from genome sequence data for staphylococcus aureus and mycobacterium tuberculosis: **6**(1):10063. ISSN 2041-1723. doi:10.1038/ncomms10063.

[104] Kos VN, Déraspe M, McLaughlin RE, Whiteaker JD, Roy PH, Alm RA, Corbeil J and Gardner H. The resistome of pseudomonas aeruginosa in relationship to phenotypic susceptibility: **59**(1):427–436. ISSN 0066-4804, 1098-6596. doi:10.1128/AAC.03954-14.

[105] Zanini F, Brodin J, Thebo L, Lanz C, Bratt G, Albert J and Neher RA. Population genomics of intrapatient HIV-1 evolution: **4**:e11282. ISSN 2050-084X. doi:10.7554/eLife.11282.

[106] Carroll MW, Matthews DA, Hiscox JA, Elmore MJ, Pollakis G, Rambaut A, Hewson R, García-Dorival I, Bore JA, Koundouno R *et al.* Temporal and spatial analysis of the 2014–2015 ebola virus outbreak in west africa: **524**(7563):97–101. ISSN 0028-0836, 1476-4687. doi:10.1038/nature14594.

[107] Gwinn M, MacCannell DR and Khabbaz RF. Integrating advanced molecular technologies into public health: **55**(3):703–714. ISSN 0095-1137, 1098-660X. doi:10.1128/JCM.01967-16.

[108] Muñoz JF, Gade L, Chow NA, Loparev VN, Juieng P, Berkow EL, Farrer RA, Litvintseva AP and Cuomo CA. Genomic insights into multidrug-resistance, mating and virulence in candida auris and related emerging species: **9**(1):5346. ISSN 2041-1723. doi:10.1038/s41467-018-07779-6.

[109] Talundzic E, Ravishankar S, Kelley J, Patel D, Plucinski M, Schmedes S, Ljolje D, Clemons B, Madison-Antenucci S, Arguin PM *et al.* Next-generation sequencing and bioinformatics protocol for malaria drug resistance marker surveillance: **62**(4). ISSN 1098-6596. doi:10.1128/AAC.02474-17.

[110] Qvarnstrom Y, Wei-Pridgeon Y, Van Roey E, Park S, Srinivasamoorthy G, Nascimento FS, Moss DM, Talundzic E and Arrowood MJ. Purification of cyclospora cayetanensis oocysts obtained from human stool specimens for whole genome sequencing: **10**(1):45. ISSN 1757-4749. doi:10.1186/s13099-018-0272-7.

[111] Gwinn M, MacCannell D and Armstrong GL. Next-generation sequencing of infectious pathogens: **321**(9):893. ISSN 0098-7484. doi:10.1001/jama.2018.21669.

[112] Satta G, Lipman M, Smith G, Arnold C, Kon O and McHugh T. Mycobacterium tuberculosis and whole-genome sequencing: how close are we to unleashing its full potential?: **24**(6):604–609. ISSN 1198743X. doi:10.1016/j.cmi.2017.10.030.

[113] Armstrong GL, MacCannell DR, Taylor J, Carleton HA, Neuhaus EB, Bradbury RS, Posey JE and Gwinn M. Pathogen genomics in public health: **381**(26):2569–2580. ISSN 0028-4793, 1533-4406. doi:10.1056/NEJMsr1813907.

[114] Quick J, Loman NJ, Duraffour S, Simpson JT, Severi E, Cowley L, Bore JA, Koundouno R, Dudas G, Mikhail A *et al.* Real-time, portable genome sequencing for ebola surveillance: **530**(7589):228–232. ISSN 0028-0836. doi:10.1038/nature16996.

[115] Faria NR, Sabino EC, Nunes MRT, Alcantara LCJ, Loman NJ and Pybus OG. Mobile real-time surveillance of zika virus in brazil: **8**(1):97. ISSN 1756-994X. doi:10.1186/s13073-016-0356-2.

[116] Besser J, Carleton H, Gerner-Smidt P, Lindsey R and Trees E. Next-generation sequencing technologies and their application to the study and control of bacterial infections: **24**(4):335–341. ISSN 1198743X. doi:10.1016/j.cmi.2017.10.013.

[117] Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, Gordon SV, Eiglmeier K, Gas S, Iii CEB *et al.* Deciphering the biology of mycobacterium tuberculosis from the complete genome sequence: **396**:27.

[118] Bastos HN, Osório NS, Gagneux S, Comas I and Saraiva M. The troika host–pathogen–extrinsic factors in tuberculosis: Modulating inflammation and clinical outcomes: **8**:1948. ISSN 1664-3224. doi:10.3389/fimmu.2017.01948.

[119] Coscolla M and Gagneux S. Consequences of genomic diversity in mycobacterium tuberculosis: **26**(6):431–444. ISSN 10445323. doi:10.1016/j.smim.2014.09.012.

[120] Comas I, Chakravartti J, Small PM, Galagan J, Niemann S, Kremer K, Ernst JD and Gagneux S. Human t cell epitopes of mycobacterium tuberculosis are evolutionarily hyperconserved: **42**(6):498–503. ISSN 1061-4036, 1546-1718. doi:10.1038/ng.590.

[121] Riojas MA, McGough KJ, Rider-Riojas CJ, Rastogi N and Hazbón MH. Phylogenomic analysis of the species of the mycobacterium tuberculosis complex demonstrates that mycobacterium africanum, mycobacterium bovis, mycobacterium caprae, mycobacterium microti and mycobacterium pinnipedii are later heterotypic synonyms of mycobacterium tuberculosis: **68**(1):324–332. ISSN 1466-5026, 1466-5034. doi:10.1099/ijsem.0.002507.

[122] Semuto Ngabonziza JC, Loiseau C, Marceau M, Jouet A, Menardo F, Tzfadia O, Niyigena EB, Mulders W, Fissette K, Diels M *et al.* A sister lineage of the *Mycobacterium tuberculosis* complex discovered in the african great lakes region. doi:10.1101/2020.01.20.912998.

[123] Gagneux S and Small PM. Global phylogeography of mycobacterium tuberculosis and implications for tuberculosis product development: **7**(5):328–337. ISSN 14733099. doi:10.1016/S1473-3099(07)70108-1.

[124] Brynildsrud OB, Pepperell CS, Suffys P, Grandjean L, Monteserin J, Debech N, Bohlin J, Alfsnes K, Pettersson JOH, Kirkeleite I *et al.* Global expansion of *Mycobacterium tuberculosis* lineage 4 shaped by colonial migration and local adaptation: **4**(10):eaat5869. ISSN 2375-2548. doi:10.1126/sciadv.aat5869.

[125] Gagneux S. Host–pathogen coevolution in human tuberculosis: **367**(1590):850–859. ISSN 0962-8436, 1471-2970. doi:10.1098/rstb.2011.0316.

[126] Holt KE, McAdam P, Thai PVK, Thuong NTT, Ha DTM, Lan NN, Lan NH, Nhu NTQ, Hai HT, Ha VTN *et al.* Frequent transmission of the mycobacterium tuberculosis beijing lineage and positive selection for the EsxW beijing variant in vietnam: **50**(6):849–856. ISSN 1061-4036, 1546-1718. doi:10.1038/s41588-018-0117-9.

[127] Comas I, Coscolla M, Luo T, Borrell S, Holt KE, Kato-Maeda M, Parkhill J, Malla B, Berg S, Thwaites G *et al.* Out-of-africa migration and neolithic coexpansion of mycobacterium tuberculosis with modern humans: **45**(10):1176–1182. ISSN 1061-4036, 1546-1718. doi:10.1038/ng.2744.

[128] Bos KI, Harkins KM, Herbig A, Coscolla M, Weber N, Comas I, Forrest SA, Bryant JM, Harris SR, Schuenemann VJ *et al.* Pre-columbian mycobacterial genomes reveal seals as a source of new world human tuberculosis: **514**(7523):494–497. ISSN 1476-4687. doi:10.1038/nature13591.

[129] Stavrum R, PrayGod G, Range N, Faurholt-Jepsen D, Jeremiah K, Faurholt-Jepsen M, Krarup H, Aabye MG, Changalucha J, Friis H *et al.* Increased level of acute phase reactants in patients infected with modern mycobacterium tuberculosis genotypes in mwanza, tanzania: **14**(1):309. ISSN 1471-2334. doi:10.1186/1471-2334-14-309.

[130] de Jong B, Hill P, Aiken A, Awine T, Antonio M, Adetifa I, Jackson-Sillah D, Fox A, DeRiemer K, Gagneux S *et al.* Progression to active tuberculosis, but not transmission, varies by *Mycobacterium tuberculosis* lineage in the gambia: **198**(7):1037–1043. ISSN 0022-1899, 1537-6613. doi:10.1086/591504.

[131] Borrell S, Tudó G, Rey E, González-Martín J, Español M, March F, Coll P, Orcau A, Caylà J, Jansà J *et al.* Tuberculosis transmission patterns among spanish-born and foreign-born populations in the city of barcelona: **16**(6):568–574. ISSN 1198743X. doi:10.1111/j.1469-0691.2009.02886.x.

[132] Coscolla M. Biological and epidemiological consequences of MTBC diversity. In Gagneux S, ed., *Strain Variation in the Mycobacterium tuberculosis Complex: Its Role in Biology, Epidemiology and Control*, pages 95–116. Springer International Publishing. ISBN 978-3-319-64371-7: doi:10.1007/978-3-319-64371-7_5.

[133] Chavadi S, Wooff E, Coldham NG, Sritharan M, Hewinson RG, Gordon SV and Wheeler PR. Global effects of inactivation of the pyruvate kinase gene in the mycobacterium tuberculosis complex: **191**(24):7545–7553. ISSN 0021-9193. doi:10.1128/JB.00619-09.

[134] Sanoussi CN, Affolabi D, Rigouts L, Anagonou S and de Jong B. Genotypic characterization directly applied to sputum improves the detection of mycobacterium africanum west african 1, under-represented in positive cultures: **11**(9):e0005900. doi:10.1371/journal.pntd.0005900. Publisher: Public Library of Science.

[135] Cowley D, Govender D, February B, Wolfe M, Steyn L, Evans J, Wilkinson R and Nicol M. Recent and rapid emergence of w-beijing strains of *Mycobacterium tuberculosis* in cape town, south africa: **47**(10):1252–1259. ISSN 1058-4838, 1537-6591. doi:10.1086/592575.

[136] Comas I, Homolka S, Niemann S and Gagneux S. Genotyping of genetically monomorphic bacteria: DNA sequencing in mycobacterium tuberculosis highlights the limitations of current methodologies: **4**(11):e7815. ISSN 1932-6203. doi:10.1371/journal.pone.0007815.

[137] CRyPTIC Consortium {and} the 100 GP, Allix-Béguec C, Arandjelovic I, Bi L, Beckert P, Bonnet M, Bradley P, Cabibbe AM, Cancino-Muñoz I, Caulfield MJ *et al.* Prediction of susceptibility to first-line tuberculosis drugs by DNA sequencing: **379**(15):1403–1415. ISSN 0028-4793. doi:10.1056/NEJMoa1800474.

[138] Starks AM, Avilés E, Cirillo DM, Denkinger CM, Dolinger DL, Emerson C, Gallarda J, Hanna D, Kim PS, Liwski R *et al.* Collaborative effort for a centralized worldwide tuberculosis relational sequencing data platform: Figure 1.: **61**:S141–S146. ISSN 1058-4838, 1537-6591. doi:10.1093/cid/civ610.

[139] Coll F, McNerney R, Preston MD, Guerra-Assunção JA, Warry A, Hill-Cawthorne G, Mallard K, Nair M, Miranda A, Alves A *et al.* Rapid determination of anti-tuberculosis drug resistance from whole-genome sequences: **7**(1):51. ISSN 1756-994X. doi:10.1186/s13073-015-0164-0.

[140] Papaventsis D, Casali N, Kontsevaya I, Drobniewski F, Cirillo D and Nikolayevskyy V. Whole genome sequencing of mycobacterium tuberculosis for detection of drug resistance: a systematic review: **23**(2):61–68. ISSN 1198743X. doi:10.1016/j.cmi.2016.09.008.

[141] Cohen KA, Manson AL, Desjardins CA, Abeel T and Earl AM. Deciphering drug resistance in mycobacterium tuberculosis using whole-genome sequencing: progress, promise, and challenges: **11**(1):45. ISSN 1756-994X. doi:10.1186/s13073-019-0660-8.

[142] Mears J, Vynnycky E, Lord J, Borgdorff MW, Cohen T, Crisp D, Innes JA, Lilley M, Maguire H, McHugh TD *et al.* The prospective evaluation of the TB strain typing service in england: a mixed methods study: **71**(8):734–741. ISSN 0040-6376, 1468-3296. doi:10.1136/thoraxjnl-2014-206480.

[143] Meehan CJ, Moris P, Kohl TA, Pečerska J, Akter S, Merker M, Utpatel C, Beckert P, Gehre F, Lempens P *et al.* The relationship between transmission time and clustering methods in mycobacterium tuberculosis epidemiology: **37**:410–416. ISSN 23523964. doi:10.1016/j.ebiom.2018.10.013.

[144] Nikolayevskyy V, Niemann S, Anthony R, van Soolingen D, Tagliani E, Ködmön C, van der Werf M and Cirillo D. Role and value of whole genome sequencing in studying tuberculosis transmission: **25**(11):1377–1382. ISSN 1198743X. doi:10.1016/j.cmi.2019.03.022.

[145] Schurch AC, Kremer K, Daviena O, Kiers A, Boeree MJ, Siezen RJ and van Soolingen D. High-resolution typing by integration of genome sequencing data in a large tuberculosis cluster: **48**(9):3403–3406. ISSN 0095-1137. doi:10.1128/JCM.00370-10.

[146] Stucki D, Ballif M, Egger M, Furrer H, Altpeter E, Battegay M, Droz S, Bruderer T, Coscolla M, Borrell S *et al.* Standard genotyping overestimates transmission

of mycobacterium tuberculosis among immigrants in a low-incidence country: **54**(7):1862–1870. ISSN 0095-1137, 1098-660X. doi:10.1128/JCM.00126-16.

[147] Yang C, Lu L, Warren JL, Wu J, Jiang Q, Zuo T, Gan M, Liu M, Liu Q, DeRiemer K *et al.* Internal migration and transmission dynamics of tuberculosis in shanghai, china: an epidemiological, spatial, genomic analysis: **18**(7):788–795. ISSN 14733099. doi:10.1016/S1473-3099(18)30218-4.

[148] Arandjelović I, Merker M, Richter E, Kohl TA, Savić B, Soldatović I, Wirth T, Vuković D and Niemann S. Longitudinal outbreak of multidrug-resistant tuberculosis in a hospital setting, serbia: **25**(3):555–558. ISSN 1080-6040, 1080-6059. doi:10.3201/eid2503.181220.

[149] Jajou R, de Neeling A, van Hunen R, de Vries G, Schimmel H, Mulder A, Anthony R, van der Hoek W and van Soolingen D. Epidemiological links between tuberculosis cases identified twice as efficiently by whole genome sequencing than conventional molecular typing: A population-based study: **13**(4):e0195413. ISSN 1932-6203. doi:10.1371/journal.pone.0195413.

[150] Walker TM, Ip CL, Harrell RH, Evans JT, Kapatai G, Dedicoat MJ, Eyre DW, Wilson DJ, Hawkey PM, Crook DW *et al.* Whole-genome sequencing to delineate mycobacterium tuberculosis outbreaks: a retrospective observational study: **13**(2):137–146. ISSN 14733099. doi:10.1016/S1473-3099(12)70277-3.

[151] Guerra-Assunção J, Crampin A, Houben R, Mzembe T, Mallard K, Coll F, Khan P, Banda L, Chiwaya A, Pereira R *et al.* Large-scale whole genome sequencing of m. tuberculosis provides insights into transmission in a high prevalence area: **4**:e05166. ISSN 2050-084X. doi:10.7554/eLife.05166.

[152] Nikolayevskyy V, Kranzer K, Niemann S and Drobniewski F. Whole genome sequencing of mycobacterium tuberculosis for detection of recent transmission and tracing outbreaks: A systematic review: **98**:77–85. ISSN 14729792. doi:10.1016/j.tube.2016.02.009.

[153] Menardo F, Duchêne S, Brites D and Gagneux S. The molecular clock of mycobacterium tuberculosis: **15**(9):e1008067. ISSN 1553-7374. doi:10.1371/journal.ppat.1008067.

[154] Didelot X, Fraser C, Gardy J and Colijn C. Genomic infectious disease epidemiology in partially sampled and ongoing outbreaks: page msw075. ISSN 0737-4038, 1537-1719. doi:10.1093/molbev/msw275.

[155] Didelot X, Croucher NJ, Bentley SD, Harris SR and Wilson DJ. Bayesian inference of ancestral dates on bacterial phylogenetic trees: **46**(22):e134–e134. ISSN 0305-1048, 1362-4962. doi:10.1093/nar/gky783.

[156] Campbell F, Cori A, Ferguson N and Jombart T. Bayesian inference of transmission chains using timing of symptoms, pathogen genomes and contact data: **15**(3):e1006930. ISSN 1553-7358. doi:10.1371/journal.pcbi.1006930.

[157] Coll F, McNerney R, Guerra-Assunção JA, Glynn JR, Perdigão J, Viveiros M, Portugal I, Pain A, Martin N and Clark TG. A robust SNP barcode for typing

mycobacterium tuberculosis complex strains: **5**(1):4812. ISSN 2041-1723. doi:10.1038/ncomms5812.

[158] Brown T, Nikolayevskyy V, Velji P and Drobniewski F. Associations between *Mycobacterium tuberculosis* strains and phenotypes: **16**(2):272–280. ISSN 1080-6040, 1080-6059. doi:10.3201/eid1602.091032.

[159] Ford CB, Shah RR, Maeda MK, Gagneux S, Murray MB, Cohen T, Johnston JC, Gardy J, Lipsitch M and Fortune SM. Mycobacterium tuberculosis mutation rate estimates from different lineages predict substantial differences in the emergence of drug-resistant tuberculosis: **45**(7):784–790. ISSN 1061-4036, 1546-1718. doi:10.1038/ng.2656.

[160] Bainomugisa A, Pandey S, Donnan E, Simpson G, Foster J, Lavu E, Hiasihri S, McBryde ES, Moke R, Vincent S *et al.* Cross-border movement of highly drug-resistant mycobacterium tuberculosis from papua new guinea to australia through torres strait protected zone, 2010–2015: **25**(3):406–415. doi:10.3201/eid2503.181003.

[161] Stucki D, Malla B, Hostettler S, Huna T, Feldmann J, Yeboah-Manu D, Borrell S, Fenner L, Comas I, Coscollà M *et al.* Two new rapid SNP-typing methods for classifying mycobacterium tuberculosis complex into the main phylogenetic lineages: **7**(7):e41253. ISSN 1932-6203. doi:10.1371/journal.pone.0041253.

[162] Carcelén M, Abascal E, Herranz M, Santantón S, Zenteno R, Ruiz Serrano MJ, Bouza E, Pérez-Lago L and García-de Viedma D. Optimizing and accelerating the assignation of lineages in mycobacterium tuberculosis using novel alternative single-tube assays: **12**(11):e0186956. ISSN 1932-6203. doi:10.1371/journal.pone.0186956.

[163] Pérez-Lago L, Martínez Lirola M, Herranz M, Comas I, Bouza E and García-de Viedma D. Fast and low-cost decentralized surveillance of transmission of tuberculosis based on strain-specific PCRs tailored from whole genome sequencing data: a pilot study: **21**(3):249.e1–249.e9. ISSN 1198743X. doi:10.1016/j.cmi.2014.10.003.

[164] Stucki D, Ballif M, Bodmer T, Coscolla M, Maurer AM, Droz S, Butz C, Borrell S, Längle C, Feldmann J *et al.* Tracking a tuberculosis outbreak over 21 years: Strain-specific single-nucleotide polymorphism typing combined with targeted whole-genome sequencing: **211**(8):1306–1316. ISSN 1537-6613, 0022-1899. doi:10.1093/infdis/jiu601.

[165] Pérez-Lago L, Campos-Herrero MI, Cañas F, Copado R, Sante L, Pino B, Lecuona M, Gil , Martín C, Muñoz P *et al.* A mycobacterium tuberculosis beijing strain persists at high rates and extends its geographic boundaries 20 years after importation: **9**(1):4687. ISSN 2045-2322. doi:10.1038/s41598-019-40525-6.

[166] Abascal E, Herranz M, Acosta F, Agapito J, Cabibbe AM, Monteserin J, Ruiz Serrano MJ, Gijón P, Fernández-González F, Lozano N *et al.* Screening of inmates transferred to spain reveals a peruvian prison as a reservoir of persistent mycobacterium tuberculosis MDR strains and mixed infections: **10**(1):2704. ISSN 2045-2322. doi:10.1038/s41598-020-59373-w.

[167] Galarza M, Fasabi M, Levano KS, Castillo E, Barreda N, Rodriguez M and Guio H. High-resolution melting analysis for molecular detection of multidrug resistance tuberculosis in peruvian isolates: **16**(1):260. ISSN 1471-2334. doi:10.1186/s12879-016-1615-y.

[168] Juma SP, Maro A, Pholwat S, Mpagama SG, Gratz J, Liyoyo A, Houpt ER, Kibiki GS, Mmbaga BT and Heysell SK. Underestimated pyrazinamide resistance may compromise outcomes of pyrazinamide containing regimens for treatment of drug susceptible and multi-drug-resistant tuberculosis in tanzania: **19**(1):129. ISSN 1471-2334. doi:10.1186/s12879-019-3757-1.

[169] Feuerriegel S, Schleusener V, Beckert P, Kohl TA, Miotto P, Cirillo DM, Cabibbe AM, Niemann S and Fellenberg K. PhyResSE: a web tool delineating mycobacterium tuberculosis antibiotic resistance and lineage from whole-genome sequencing data: **53**(6):1908–1914. ISSN 0095-1137, 1098-660X. doi:10.1128/JCM.00025-15.

[170] Churchyard G, Kim P, Shah NS, Rustomjee R, Gandhi N, Mathema B, Dowdy D, Kasmar A and Cardenas V. What we know about tuberculosis transmission: An overview: **216**:S629–S635. ISSN 0022-1899, 1537-6613. doi:10.1093/infdis/jix362.

[171] Dowdy DW, Grant AD, Dheda K, Nardell E, Fielding K and Moore DAJ. Designing and evaluating interventions to halt the transmission of tuberculosis: **216**:S654–S661. ISSN 0022-1899. doi:10.1093/infdis/jix320.

[172] Lönnroth K, Corbett E, Golub J, Godfrey-Faussett P, Uplekar M, Weil D and Raviglione M. Systematic screening for active tuberculosis: rationale, definitions and key considerations: **17**(3):289–298. ISSN 1815-7920. doi:10.5588/ijtld.12.0797.

[173] Ho J, Fox GJ and Marais BJ. Passive case finding for tuberculosis is not enough: **5**(4):374–378. ISSN 2212-554X. doi:10.1016/j.ijmyco.2016.09.023.

[174] Marks GB, Nguyen NV, Nguyen PTB, Nguyen TA, Nguyen HB, Tran KH, Nguyen SV, Luu KB, Tran DTT, Vo QTN *et al.* Community-wide screening for tuberculosis in a high-prevalence setting: **381**(14):1347–1357. ISSN 1533-4406. doi:10.1056/NEJMoa1902129.

[175] Gardy JL, Johnston JC, Sui SJH, Cook VJ, Shah L, Brodkin E, Rempel S, Moore R, Zhao Y, Holt R *et al.* Whole-genome sequencing and social-network analysis of a tuberculosis outbreak: **364**(8):730–739. ISSN 0028-4793, 1533-4406. doi:10.1056/NEJMoa1003176.

[176] Guthrie JL, Strudwick L, Roberts B, Allen M, McFadzen J, Roth D, Jorgensen D, Rodrigues M, Tang P, Hanley B *et al.* Whole genome sequencing for improved understanding of *Mycobacterium tuberculosis* transmission in a remote circumpolar region: **147**:e188. ISSN 0950-2688, 1469-4409. doi:10.1017/S0950268819000670.

[177] Zignol M, Cabibbe AM, Dean AS, Glaziou P, Alikhanova N, Ama C, Andres S, Barbova A, Borbe-Reyes A, Chin DP *et al.* Genetic sequencing for surveillance

of drug resistance in tuberculosis in highly endemic countries: a multi-country population-based surveillance study: **18**(6):675–683. ISSN 14733099. doi: 10.1016/S1473-3099(18)30073-2.

[178] Shea J, Halse TA, Lapierre P, Shudt M, Kohlerschmidt D, Van Roey P, Limberger R, Taylor J, Escuyer V and Musser KA. Comprehensive whole-genome sequencing and reporting of drug resistance profiles on clinical cases of mycobacterium tuberculosis in new york state: **55**(6):1871–1882. ISSN 1098-660X. doi:10.1128/JCM.00298-17.

[179] Jajou R, van der Laan T, de Zwaan R, Kamst M, Mulder A, de Neeling A, Anthony R and van Soolingen D. WGS more accurately predicts susceptibility of mycobacterium tuberculosis to first-line drugs than phenotypic testing: **74**(9):2605–2616. ISSN 0305-7453, 1460-2091. doi:10.1093/jac/dkz215.

[180] Xu Y, Cancino-Muñoz I, Torres-Puente M, Villamayor LM, Borrás R, Borrás-Máñez M, Bosque M, Camarena JJ, Colomer-Roig E, Colomina J *et al.* High-resolution mapping of tuberculosis transmission: Whole genome sequencing and phylogenetic modelling of a cohort from valencia region, spain: **16**(10):e1002961. ISSN 1549-1676. doi:10.1371/journal.pmed.1002961.

[181] Ezewudo M, Borens A, Chiner-Oms Miotto P, Chindelevitch L, Starks AM, Hanna D, Liwski R, Zignol M, Gilpin C *et al.* Integrating standardized whole genome sequence analysis with a global mycobacterium tuberculosis antibiotic resistance knowledgebase: **8**(1):15382. ISSN 2045-2322. doi:10.1038/s41598-018-33731-1.

[182] Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies: **30**(9):1312–1313. ISSN 1460-2059, 1367-4803. doi:10.1093/bioinformatics/btu033.

[183] Parikh R, Mathai A, Parikh S, Sekhar GC and Thomas R. Understanding and using sensitivity, specicity and predictive values: **56**(1):6.

[184] Alonso Rodríguez N, Andrés S, Bouza E, Herranz M, Serrano MJR, de Viedma DG, Palenque E, Chaves F, Iñigo J, Cías R *et al.* Transmission permeability of tuberculosis involving immigrants, revealed by a multicentre analysis of clusters: **15**(5):435–442. ISSN 1198743X. doi:10.1111/j.1469-0691.2008.02670.x.

[185] Guthrie JL, Kong C, Roth D, Jorgensen D, Rodrigues M, Hoang L, Tang P, Cook V, Johnston J and Gardy JL. Molecular epidemiology of tuberculosis in british columbia, canada: A 10-year retrospective study: **66**(6):849–856. ISSN 1058-4838, 1537-6591. doi:10.1093/cid/cix906.

[186] López MG, Dogba JB, Torres-Puente M, Goig GA, Moreno-Molina M, Villamayor LM, Cadmus S and Comas I. Tuberculosis in liberia: high multidrug-resistance burden, transmission and diversity modelled by multiple importation events: **6**(1). ISSN 2057-5858. doi:10.1099/mgen.0.000325.

[187] Cancino-Muñoz I, Moreno-Molina M, Furió V, Goig GA, Torres-Puente M, Chiner-Oms Villamayor LM, Sanz F, Guna-Serrano MR and Comas I. Cryptic resistance mutations associated with misdiagnoses of multidrug-resistant tuberculosis: **220**(2):316–320. ISSN 0022-1899, 1537-6613. doi:10.1093/infdis/jiz104.

[188] Lönnroth K, Migliori GB, Abubakar I, D'Ambrosio L, de Vries G, Diel R, Douglas P, Falzon D, Gaudreau MA, Goletti D *et al.* Towards tuberculosis elimination: an action framework for low-incidence countries: **45**(4):928–952. ISSN 0903-1936, 1399-3003. doi:10.1183/09031936.00214014.

[189] Hatherell HA, Colijn C, Stagg HR, Jackson C, Winter JR and Abubakar I. Interpreting whole genome sequencing for investigating tuberculosis transmission: a systematic review: **14**(1):21. ISSN 1741-7015. doi:10.1186/s12916-016-0566-x.

[190] Folkvardsen DB, Norman A, Andersen , Michael Rasmussen E, Jelsbak L and Lillebaek T. Genomic epidemiology of a major mycobacterium tuberculosis outbreak: Retrospective cohort study in a low-incidence setting using sparse time-series sampling: **216**(3):366–374. ISSN 0022-1899, 1537-6613. doi:10.1093/infdis/jix298.

[191] van Soolingen D. Whole-genome sequencing of mycobacterium tuberculosis as an epidemiological marker: **2**(4):251–252. ISSN 22132600. doi:10.1016/S2213-2600(14)70049-9.

[192] Wood DE and Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments: **15**(3):R46. ISSN 1465-6906. doi:10.1186/gb-2014-15-3-r46.

[193] Feuerriegel S, Schleusener V, Beckert P, Kohl TA, Miotto P, Cirillo DM, Cabibbe AM, Niemann S and Fellenberg K. PhyResSE: a web tool delineating mycobacterium tuberculosis antibiotic resistance and lineage from whole-genome sequencing data: **53**(6):1908–1914. ISSN 0095-1137. doi:10.1128/JCM.00025-15.

[194] Chiner-Oms Sánchez-Busó L, Corander J, Gagneux S, Harris SR, Young D, González-Candelas F and Comas I. Genomic determinants of speciation and spread of the mycobacterium tuberculosis complex: page 15.

[195] Leigh JW and Bryant D. POPART:full-feature software for haplotype network reconstruction: **6**:1110–1116.

[196] Duchêne S, Holt KE, Weill FX, Le Hello S, Hawkey J, Edwards DJ, Fourment M and Holmes EC. Genome-scale rates of evolutionary change in bacteria: **2**(11). ISSN 2057-5858. doi:10.1099/mgen.0.000094.

[197] Volz EM and Frost SDW. Scalable relaxed clock phylogenetic dating: **3**(2). ISSN 2057-1577. doi:10.1093/ve/vex025.

[198] Merker M, Barbier M, Cox H, Rasigade JP, Feuerriegel S, Kohl TA, Diel R, Borrell S, Gagneux S, Nikolayevskyy V *et al.* Compensatory evolution drives multidrug-resistant tuberculosis in central asia: **7**. ISSN 2050-084X. doi:10.7554/eLife.38200.

[199] Rodrigo T, Cayl{\textbackslash}a JA, Garc{\textbackslash}ia de Olalla P, Gald{\textbackslash}os-Tangüis H, Jans{\textbackslash}a JM, Miranda P and Brugal T. Characteristics of tuberculosis patients who generate secondary cases: **1**(4):352–357. ISSN 1027-3719.

[200] Comas I and Gardy JL. TB transmission: Closing the gaps: **34**:4–5. ISSN 23523964. doi:10.1016/j.ebiom.2018.07.020.

[201] Stimson J, Gardy J, Mathema B, Crudu V, Cohen T and Colijn C. Beyond the SNP threshold: Identifying outbreak clusters using inferred transmissions: **36**(3):587–603. ISSN 0737-4038, 1537-1719. doi:10.1093/molbev/msy242.

[202] Glynn JR, Guerra-Assunção JA, Houben RMGJ, Sichali L, Mzembe T, Mwaungulu LK, Mwaungulu JN, McNerney R, Khan P, Parkhill J *et al.* Whole genome sequencing shows a low proportion of tuberculosis disease is attributable to known close contacts in rural malawi: **10**(7):e0132840. ISSN 1932-6203. doi:10.1371/journal.pone.0132840.

[203] Tostmann A, Kik S, Kalisvaart N, Sebek M, Verver S, Boeree M and van Soolingen D. Tuberculosis transmission by patients with smear-negative pulmonary tuberculosis in a large cohort in the netherlands: **47**(9):1135–1142. ISSN 1058-4838, 1537-6591. doi:10.1086/591974.

[204] Klinkenberg D, Backer JA, Didelot X, Colijn C and Wallinga J. Simultaneous inference of phylogenetic and transmission trees in infectious disease outbreaks: **13**(5):e1005495. ISSN 1553-7358. doi:10.1371/journal.pcbi.1005495.

[205] Hall M, Woolhouse M and Rambaut A. Epidemic reconstruction in a phylogenetics framework: Transmission trees as partitions of the node set: **11**(12):e1004613. ISSN 1553-7358. doi:10.1371/journal.pcbi.1004613.

[206] De Maio N, Wu CH and Wilson DJ. SCOTTI: Efficient reconstruction of transmission within outbreaks with the structured coalescent: **12**(9):e1005130. ISSN 1553-7358. doi:10.1371/journal.pcbi.1005130.

[207] Campbell F, Strang C, Ferguson N, Cori A and Jombart T. When are pathogen genome sequences informative of transmission events?: **14**(2):e1006885. ISSN 1553-7374. doi:10.1371/journal.ppat.1006885.

[208] Esmail H, Dodd PJ and Houben RMGJ. Tuberculosis transmission during the subclinical period: could unrelated cough play a part?: **6**(4):244–246. ISSN 22132600. doi:10.1016/S2213-2600(18)30105-X.

[209] Houben RMGJ, Esmail H, Emery JC, Joslyn LR, McQuaid CF, Menzies NA, Sanz J, Shrestha S, White RG, Yang C *et al.* Spotting the old foe—revisiting the case definition for TB: **7**(3):199–201. ISSN 22132600. doi:10.1016/S2213-2600(19)30038-4.

[210] Ford CB, Lin PL, Chase MR, Shah RR, Iartchouk O, Galagan J, Mohaideen N, Ioerger TR, Sacchettini JC, Lipsitch M *et al.* Use of whole genome sequencing to estimate the mutation rate of mycobacterium tuberculosis during latent infection: **43**(5):482–486. ISSN 1061-4036, 1546-1718. doi:10.1038/ng.811.

[211] Bryant JM, Schürch AC, van Deutekom H, Harris SR, de Beer JL, de Jager V, Kremer K, van Hijum SAFT, Siezen RJ, Borgdorff M *et al.* Inferring patient to patient transmission of mycobacterium tuberculosisfrom whole genome sequencing data: **13**(1):110. ISSN 1471-2334. doi:10.1186/1471-2334-13-110.

[212] Merker M, Blin C, Mona S, Duforet-Frebourg N, Lecher S, Willery E, Blum MGB, Rüsch-Gerdes S, Mokrousov I, Aleksic E *et al.* Evolutionary history and global

spread of the mycobacterium tuberculosis beijing lineage: **47**(3):242–249. ISSN 1061-4036, 1546-1718. doi:10.1038/ng.3195.

[213] Luo T, Comas I, Luo D, Lu B, Wu J, Wei L, Yang C, Liu Q, Gan M, Sun G *et al.* Southern east asian origin and coexpansion of *Mycobacterium tuberculosis* beijing family with han chinese: **112**(26):8136–8141. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1424063112.

[214] Eldholm V, Monteserin J, Rieux A, Lopez B, Sobkowiak B, Ritacco V and Balloux F. Four decades of transmission of a multidrug-resistant mycobacterium tuberculosis outbreak strain: **6**(1):7119. ISSN 2041-1723. doi:10.1038/ncomms8119.

[215] Kay GL, Sergeant MJ, Zhou Z, Chan JZM, Millard A, Quick J, Szikossy I, Pap I, Spigelman M, Loman NJ *et al.* Eighteenth-century genomes show that mixed infections were common at time of peak tuberculosis in europe: **6**(1):6717. ISSN 2041-1723. doi:10.1038/ncomms7717.

[216] Bjorn-Mortensen K, Soborg B, Koch A, Ladefoged K, Merker M, Lillebaek T, Andersen AB, Niemann S and Kohl TA. Tracing mycobacterium tuberculosis transmission by whole genome sequencing in a high incidence setting: a retrospective population-based study in east greenland: **6**(1):33180. ISSN 2045-2322. doi:10.1038/srep33180.

[217] Liu Q, Ma A, Wei L, Pang Y, Wu B, Luo T, Zhou Y, Zheng HX, Jiang Q, Gan M *et al.* China's tuberculosis epidemic stems from historical expansion of four strains of mycobacterium tuberculosis: **2**(12):1982–1992. ISSN 2397-334X. doi:10.1038/s41559-018-0680-6.

[218] Duchene S, Duchene DA, Geoghegan JL, Dyson ZA, Hawkey J and Holt KE. Inferring demographic parameters in bacterial genomic data using bayesian and hybrid phylogenetic methods: **18**(1):95. ISSN 1471-2148. doi:10.1186/s12862-018-1210-5.

[219] Rutaihwa LK, Menardo F, Stucki D, Gygli SM, Ley SD, Malla B, Feldmann J, Borrell S, Beisel C, Middelkoop K *et al.* Multiple introductions of mycobacterium tuberculosis lineage 2–beijing into africa over centuries: **7**:112. ISSN 2296-701X. doi:10.3389/fevo.2019.00112.

[220] Cox H, Hughes J, Black J and Nicol MP. Precision medicine for drug-resistant tuberculosis in high-burden countries: is individualised treatment desirable and feasible?: **18**(9):e282–e287. ISSN 14733099. doi:10.1016/S1473-3099(18)30104-X.

[221] Van Deun A, Aung KJM, Bola V, Lebeke R, Hossain MA, de Rijk WB, Rigouts L, Gumusboga A, Torrea G and de Jong BC. Rifampin drug resistance tests for tuberculosis: challenging the gold standard: **51**(8):2633–2640. ISSN 0095-1137. doi:10.1128/JCM.00553-13.

[222] Ho J, Jelfs P and Sintchencko V. Phenotypically occult multidrug-resistant mycobacterium tuberculosis: dilemmas in diagnosis and treatment: **68**(12):2915–2920. ISSN 0305-7453, 1460-2091. doi:10.1093/jac/dkt284.

[223] Romanowski K, Balshaw RF, Benedetti A, Campbell JR, Menzies D, Ahmad Khan F and Johnston JC. Predicting tuberculosis relapse in patients treated with the standard 6-month regimen: an individual patient data meta-analysis: **74**(3):291–297. ISSN 0040-6376, 1468-3296. doi:10.1136/thoraxjnl-2017-211120.

[224] Metcalfe JZ, Streicher E, Theron G, Colman RE, Allender C, Lemmer D, Warren R and Engelthaler DM. Cryptic microheteroresistance explains *Mycobacterium tuberculosis* phenotypic resistance: **196**(9):1191–1201. ISSN 1073-449X, 1535-4970. doi:10.1164/rccm.201703-0556OC.

[225] Ridzon R. Risk factors for rifampin mono-resistant tuberculosis: **157**:4.

[226] Manson AL, Abeel T, Galagan JE, Sundaramurthi JC, Salazar A, Gehrmann T, Shanmugam SK, Palaniyandi K, Narayanan S, Swaminathan S *et al.* Mycobacterium tuberculosis whole genome sequences from southern india suggest novel resistance mechanisms and the need for region-specific diagnostics: **64**(11):1494–1501. ISSN 1058-4838, 1537-6591. doi:10.1093/cid/cix169.

[227] World Health Organization and World Health Organization. *Companion handbook to the WHO guidelines for the programmatic management of drug-resistant tuberculosis.* ISBN 978-92-4-154880-9. OCLC: 890467392.

[228] van Soolingen D, Hermans PW, de Haas PE, Soll DR and van Embden JD. Occurrence and stability of insertion sequences in mycobacterium tuberculosis complex strains: evaluation of an insertion sequence-dependent DNA polymorphism as a tool in the epidemiology of tuberculosis.: **29**(11):2578–2586. ISSN 0095-1137, 1098-660X. doi:10.1128/JCM.29.11.2578-2586.1991.

[229] Li H and Durbin R. Fast and accurate short read alignment with burrows-wheeler transform: **25**(14):1754–1760. ISSN 1367-4803, 1460-2059. doi:10.1093/bioinformatics/btp324.

[230] Guerra-Assunção JA, Houben RMGJ, Crampin AC, Mzembe T, Mallard K, Coll F, Khan P, Banda L, Chiwaya A, Pereira RPA *et al.* Recurrence due to relapse or reinfection with *Mycobacterium tuberculosis* : A whole-genome sequencing approach in a large, population-based cohort with a high HIV infection prevalence and active follow-up: **211**(7):1154–1163. ISSN 0022-1899, 1537-6613. doi:10.1093/infdis/jiu574.

[231] Palomino JC, Martin A, Camacho M, Guerra H, Swings J and Portaels F. Resazurin microtiter assay plate: Simple and inexpensive method for detection of drug resistance in mycobacterium tuberculosis: page 3.

[232] Colman RE, Schupp JM, Hicks ND, Smith DE, Buchhagen JL, Valafar F, Crudu V, Romancenco E, Noroc E, Jackson L *et al.* Detection of low-level mixed-population drug resistance in mycobacterium tuberculosis using high fidelity amplicon sequencing: **10**(5):e0126626. ISSN 1932-6203. doi:10.1371/journal.pone.0126626.

[233] Li J, Munsiff SS, Driver CR and Sackoff J. Relapse and acquired rifampin resistance in HIV-infected patients with tuberculosis treated with rifampin- or rifabutin-based regimens in new york city, 1997-2000: **41**(1):83–91. ISSN 1058-4838, 1537-6591. doi:10.1086/430377.

[234] Meyssonnier V, Bui TV, Veziris N, Jarlier V and Robert J. Rifampicin mono-resistant tuberculosis in france: a 2005–2010 retrospective cohort analysis: **14**(1). ISSN 1471-2334. doi:10.1186/1471-2334-14-18.

[235] Comas I, Borrell S, Roetzer A, Rose G, Malla B, Kato-Maeda M, Galagan J, Niemann S and Gagneux S. Whole-genome sequencing of rifampicin-resistant mycobacterium tuberculosis strains identifies compensatory mutations in RNA polymerase genes: **44**(1):106–110. ISSN 1061-4036, 1546-1718. doi:10.1038/ng.1038.

[236] Farhat MR, Shapiro BJ, Kieser KJ, Sultana R, Jacobson KR, Victor TC, Warren RM, Streicher EM, Calver A, Sloutsky A *et al.* Genomic analysis identifies targets of convergent positive selection in drug-resistant mycobacterium tuberculosis: **45**(10):1183–1189. ISSN 1061-4036, 1546-1718. doi:10.1038/ng.2747.

[237] Kandler JL, Mercante AD, Dalton TL, Ezewudo MN, Cowan LS, Burns SP and Metchock B. Validation of novel mycobacterium tuberculosis isoniazid resistance mutations not detectable by common molecular tests: **62**(10):16.

[238] Homolka S, Meyer CG, Hillemann D, Owusu-Dabo E, Adjei O, Horstmann RD, Browne EN, Chinbuah A, Osei I, Gyapong J *et al.* Unequal distribution of resistance-conferring mutations among mycobacterium tuberculosis and mycobacterium africanum strains from ghana: **300**(7):489–495. ISSN 14384221. doi:10.1016/j.ijmm.2010.04.019.

[239] World Health Organization. *Global tuberculosis report 2018*. ISBN 978-92-4-156564-6. OCLC: 1079428104.

[240] Hershberg R, Lipatov M, Small PM, Sheffer H, Niemann S, Homolka S, Roach JC, Kremer K, Petrov DA, Feldman MW *et al.* High functional diversity in mycobacterium tuberculosis driven by genetic drift and human demography: **6**(12):e311. ISSN 1545-7885. doi:10.1371/journal.pbio.0060311.

[241] Comas I, Hailu E, Kiros T, Bekele S, Mekonnen W, Gumi B, Tschopp R, Ameni G, Hewinson RG, Robertson BD *et al.* Population genomics of mycobacterium tuberculosis in ethiopia contradicts the virgin soil hypothesis for human tuberculosis in sub-saharan africa: **25**(24):3260–3266. ISSN 09609822. doi:10.1016/j.cub.2015.10.061.

[242] Otchere ID, Coscollá M, Sánchez-Busó L, Asante-Poku A, Brites D, Loiseau C, Meehan C, Osei-Wusu S, Forson A, Laryea C *et al.* Comparative genomics of mycobacterium africanum lineage 5 and lineage 6 from ghana suggests distinct ecological niches: **8**(1). ISSN 2045-2322. doi:10.1038/s41598-018-29620-2.

[243] Kamerbeek J, Schouls L, Kolk A, van Agterveld M, van Soolingen D, Kuijper S, Bunschoten A, Molhuizen H, Shaw R, Goyal M *et al.* Simultaneous detection and strain differentiation of mycobacterium tuberculosis for diagnosis and epidemiology.: **35**(4):907–914. ISSN 0095-1137. doi:10.1128/JCM.35.4.907-914.1997.

[244] Supply P, Allix C, Lesjean S, Cardoso-Oelemann M, Rusch-Gerdes S, Willery E, Savine E, de Haas P, van Deutekom H, Roring S *et al.* Proposal for standardization

of optimized mycobacterial interspersed repetitive unit-variable-number tandem repeat typing of mycobacterium tuberculosis: **44**(12):4498–4510. ISSN 0095-1137. doi:10.1128/JCM.01392-06.

[245] Ramirez MV, Cowart KC, Campbell PJ, Morlock GP, Sikes D, Winchell JM and Posey JE. Rapid detection of multidrug-resistant mycobacterium tuberculosis by use of real-time PCR and high-resolution melt analysis: **48**(11):4003–4009. ISSN 0095-1137. doi:10.1128/JCM.00812-10.

[246] Yin X, Zheng L, Liu Q, Lin L, Hu X, Hu Y and Wang Q. High-resolution melting curve analysis for rapid detection of rifampin resistance in mycobacterium tuberculosis: a meta-analysis: **51**(10):3294–3299. ISSN 0095-1137. doi:10.1128/JCM.01264-13.

[247] Sharma K, Sharma M, Singh S, Modi M, Sharma A, Ray P and Varma S. Real-time PCR followed by high-resolution melting curve analysis: A rapid and pragmatic approach for screening of multidrug-resistant extrapulmonary tuberculosis: **106**:56–61. ISSN 14729792. doi:10.1016/j.tube.2017.07.002.

[248] Pholwat S, Stroup S, Gratz J, Trangan V, Foongladda S, Kumburu H, Juma SP, Kibiki G and Houpt E. Pyrazinamide susceptibility testing of mycobacterium tuberculosis by high resolution melt analysis: **94**(1):20–25. ISSN 14729792. doi:10.1016/j.tube.2013.10.006.

[249] Landolt P, Stephan R and Scherrer S. Development of a new high resolution melting (HRM) assay for identification and differentiation of mycobacterium tuberculosis complex samples: **9**(1). ISSN 2045-2322. doi:10.1038/s41598-018-38243-6.

[250] Borrell S, Trauner A, Brites D, Rigouts L, Loiseau C, Coscolla M, Niemann S, Jong BD, Yeboah-Manu D, Kato-Maeda M *et al.* Reference set of *Mycobacterium tuberculosis* clinical strains: A tool for research and product development. doi:10.1101/399709.

[251] Fenner L, Malla B, Ninet B, Dubuis O, Stucki D, Borrell S, Huna T, Bodmer T, Egger M and Gagneux S. "pseudo-beijing": Evidence for convergent evolution in the direct repeat region of mycobacterium tuberculosis: **6**(9):e24737. ISSN 1932-6203. doi:10.1371/journal.pone.0024737.

[252] Untergasser A, Nijveen H, Rao X, Bisseling T, Geurts R and Leunissen JA. Primer3plus, an enhanced web interface to primer3: **35**:W71–W74. ISSN 0305-1048, 1362-4962. doi:10.1093/nar/gkm306.

[253] Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K and Madden TL. BLAST+: architecture and applications: **10**(1):421. ISSN 1471-2105. doi:10.1186/1471-2105-10-421.

[254] Bonfield JK and Whitwham A. Gap5—editing the billion fragment sequence assembly: **26**(14):1699–1703. ISSN 1460-2059, 1367-4803. doi:10.1093/bioinformatics/btq268.

[255] For the Tuberculosis Research Unit, Wampande EM, Hatzios SK, Achan B, Mupere E, Nsereko M, Mayanja HK, Eisenach K, Boom WH, Gagneux S *et al.* A single-nucleotide-polymorphism real-time PCR assay for genotyping of mycobacterium

tuberculosis complex in peri-urban kampala: **15**(1). ISSN 1471-2334. doi:10.1186/s12879-015-1121-7.

[256] Lopes JS, Marques I, Soares P, Nebenzahl-Guimaraes H, Costa J, Miranda A, Duarte R, Alves A, Macedo R, Duarte TA *et al.* SNP typing reveals similarity in mycobacterium tuberculosis genetic diversity between portugal and northeast brazil: **18**:238–246. ISSN 15671348. doi:10.1016/j.meegid.2013.04.028.

[257] Samper S, Gavín P, Millán-Lou MI, Iglesias MJ, Jiménez MS, Couvin D and Rastogi N. Mycobacterium tuberculosis genotypes and predominant clones among the multidrug-resistant isolates in spain 1998–2005: **55**:117–126. ISSN 15671348. doi:10.1016/j.meegid.2017.08.003.

[258] Chihota VN, Niehaus A, Streicher EM, Wang X, Sampson SL, Mason P, Källenius G, Mfinanga SG, Pillay M, Klopper M *et al.* Geospatial distribution of mycobacterium tuberculosis genotypes in africa: **13**(8):e0200632. ISSN 1932-6203. doi:10.1371/journal.pone.0200632.

[259] Anthwal D, Gupta RK, Bhalla M, Bhatnagar S, Tyagi JS and Haldar S. Direct detection of rifampin and isoniazid resistance in sputum samples from tuberculosis patients by high-resolution melt curve analysis: **55**(6):1755–1766. ISSN 0095-1137, 1098-660X. doi:10.1128/JCM.02104-16.

[260] Singhania A, Wilkinson RJ, Rodrigue M, Haldar P and O'Garra A. The value of transcriptomics in advancing knowledge of the immune response and diagnosis in tuberculosis: **19**(11):1159–1168. ISSN 1529-2908. doi:10.1038/s41590-018-0225-9.

[261] Yuen CM, Amanullah F, Dharmadhikari A, Nardell EA, Seddon JA, Vasilyeva I, Zhao Y, Keshavjee S and Becerra MC. Turning off the tap: stopping tuberculosis transmission through active case-finding and prompt effective treatment: **386**(10010):2334–2343. ISSN 0140-6736. doi:10.1016/S0140-6736(15)00322-0.

[262] Calligaro GL, Zijenah LS, Peter JG, Theron G, Buser V, McNerney R, Bara W, Bandason T, Govender U, Tomasicchio M *et al.* Effect of new tuberculosis diagnostic technologies on community-based intensified case finding: a multicentre randomised controlled trial: **17**(4):441–450. ISSN 1473-3099. doi:10.1016/S1473-3099(16)30384-X.

[263] Goig GA, Torres-Puente M, Mariner-Llicer C, Villamayor LM, Chiner-Oms Gil-Brusola A, Borrás R and Comas I. Towards next-generation diagnostics for tuberculosis: identification of novel molecular targets by large-scale comparative genomics: ISSN 1367-4803. doi:10.1093/bioinformatics/btz729.

[264] Torres JN, Paul LV, Rodwell TC, Victor TC, Amallraja AM, Elghraoui A, Goodmanson AP, Ramirez-Busby SM, Chawla A, Zadorozhny V *et al.* Novel katG mutations causing isoniazid resistance in clinical m. tuberculosis isolates: **4**(7):e42. ISSN 2222-1751. doi:10.1038/emi.2015.42.

[265] Bloemberg GV, Keller PM, Stucki D, Trauner A, Borrell S, Latshang T, Coscolla M, Rothe T, Hömke R, Ritter C *et al.* Acquired resistance to bedaquiline and delamanid

in therapy for tuberculosis: **373**(20):1986–1988. ISSN 0028-4793. doi:10.1056/NEJMc1505196.

[266] de Vos M, Ley SD, Wiggins KB, Derendinger B, Dippenaar A, Grobbelaar M, Reuter A, Dolby T, Burns S, Schito M *et al.* Bedaquiline microheteroresistance after cessation of tuberculosis treatment: **380**(22):2178–2180. ISSN 0028-4793. doi:10.1056/NEJMc1815121.

[267] Andres S, Merker M, Heyckendorf J, Kalsdorf B, Rumetshofer R, Indra A, Hofmann-Thiel S, Hoffmann H, Lange C, Niemann S *et al.* Bedaquiline-resistant tuberculosis: Dark clouds on the horizon: ISSN 1073-449X. doi:10.1164/rccm.201909-1819LE.

[268] Pérez-Lago L, Herranz M, Comas I, Ruiz-Serrano MJ, López Roa P, Bouza E and García-de Viedma D. Ultrafast assessment of the presence of a high-risk mycobacterium tuberculosis strain in a population: **54**(3):779–781. ISSN 0095-1137, 1098-660X. doi:10.1128/JCM.02851-15.

[269] Luo T, Yang C, Peng Y, Lu L, Sun G, Wu J, Jin X, Hong J, Li F, Mei J *et al.* Whole-genome sequencing to detect recent transmission of mycobacterium tuberculosis in settings with a high burden of tuberculosis: **94**(4):434–440. ISSN 14729792. doi:10.1016/j.tube.2014.04.005.

[270] Smit PW, Vasankari T, Aaltonen H, Haanperä M, Casali N, Marttila H, Marttila J, Ojanen P, Ruohola A, Ruutu P *et al.* Enhanced tuberculosis outbreak investigation using whole genome sequencing and IGRA: **45**(1):276–279. ISSN 0903-1936, 1399-3003. doi:10.1183/09031936.00125914.

[271] Yang C, Luo T, Shen X, Wu J, Gan M, Xu P, Wu Z, Lin S, Tian J, Liu Q *et al.* Transmission of multidrug-resistant mycobacterium tuberculosis in shanghai, china: a retrospective observational study using whole-genome sequencing and epidemiological investigation: **17**(3):275–284. ISSN 14733099. doi:10.1016/S1473-3099(16)30418-2.

[272] Casali N, Broda A, Harris SR, Parkhill J, Brown T and Drobniewski F. Whole genome sequence analysis of a large isoniazid-resistant tuberculosis outbreak in london: A retrospective observational study: **13**(10):e1002137. ISSN 1549-1676. doi:10.1371/journal.pmed.1002137.

[273] Arnold A, Witney AA, Vergnano S, Roche A, Cosgrove CA, Houston A, Gould KA, Hinds J, Riley P, Macallan D *et al.* XDR-TB transmission in london: Case management and contact tracing investigation assisted by early whole genome sequencing: **73**(3):210–218. ISSN 01634453. doi:10.1016/j.jinf.2016.04.037.

[274] Fiebig L, Kohl TA, Popovici O, Mühlenfeld M, Indra A, Homorodean D, Chiotan D, Richter E, Rüsch-Gerdes S, Schmidgruber B *et al.* A joint cross-border investigation of a cluster of multidrug-resistant tuberculosis in austria, romania and germany in 2014 using classic, genotyping and whole genome sequencing methods: lessons learnt: **22**(2):30439. Type: doi:https://doi.org/10.2807/1560-7917.ES.2017.22.2.30439.

[275] Black AT, Hamblion EL, Buttivant H, Anderson SR, Stone M, Casali N, Drobniewski F, Nwoguh F, Marshall BG and Booth L. Tracking and responding to an outbreak

of tuberculosis using MIRU-VNTR genotyping and whole genome sequencing as epidemiological tools: **40**(2):e66–e73. ISSN 1741-3842, 1741-3850. doi:10.1093/pubmed/fdx075.

[276] Norheim G, Seterelv S, Arnesen TM, Mengshoel AT, Tønjum T, Rønning JO and Eldholm V. Tuberculosis outbreak in an educational institution in norway: **55**(5):1327–1333. ISSN 0095-1137, 1098-660X. doi:10.1128/JCM.01152-16.

[277] Gautam SS, Mac Aogáin M, Cooley LA, Haug G, Fyfe JA, Globan M and O'Toole RF. Molecular epidemiology of tuberculosis in tasmania and genomic characterisation of its first known multi-drug resistant case: **13**(2):e0192351. ISSN 1932-6203. doi:10.1371/journal.pone.0192351.

[278] Séraphin MN, Didelot X, Nolan DJ, May JR, Khan MSR, Murray ER, Salemi M, Morris JG and Lauzardo M. Genomic investigation of a mycobacterium tuberculosis outbreak involving prison and community cases in florida, united states: **99**(4):867–874. ISSN 0002-9637, 1476-1645. doi:10.4269/ajtmh.17-0700.

[279] Conceição EC, Guimarães AEdS, Lopes ML, Furlaneto IP, Rodrigues YC, da Conceição ML, Barros WA, Cardoso NC, Sharma A, Lima LNGC *et al.* Analysis of potential household transmission events of tuberculosis in the city of belem, brazil: **113**:125–129. ISSN 14729792. doi:10.1016/j.tube.2018.09.011.

[280] Jajou R, de Neeling A, Rasmussen EM, Norman A, Mulder A, van Hunen R, de Vries G, Haddad W, Anthony R, Lillebaek T *et al.* A predominant variable-number tandem-repeat cluster of mycobacterium tuberculosis isolates among asylum seekers in the netherlands and denmark, deciphered by whole-genome sequencing: **56**(2):e01100–17. ISSN 0095-1137, 1098-660X. doi:10.1128/JCM.01100-17.

[281] Alaridah N, Hallbäck ET, Tångrot J, Winqvist N, Sturegård E, Florén-Johansson K, Jönsson B, Tenland E, Welinder-Olsson C, Medstrand P *et al.* Transmission dynamics study of tuberculosis isolates with whole genome sequencing in southern sweden: **9**(1):4931. ISSN 2045-2322. doi:10.1038/s41598-019-39971-z.

[282] Dixit A, Freschi L, Vargas R, Calderon R, Sacchettini J, Drobniewski F, Galea JT, Contreras C, Yataco R, Zhang Z *et al.* Whole genome sequencing identifies bacterial factors affecting transmission of multidrug-resistant tuberculosis in a high-prevalence setting: **9**(1):5602. ISSN 2045-2322. doi:10.1038/s41598-019-41967-8.

[283] Diel R, Kohl TA, Maurer FP, Merker M, Meywald Walter K, Hannemann J, Nienhaus A, Supply P and Niemann S. Accuracy of whole-genome sequencing to determine recent tuberculosis transmission: an 11-year population-based study in hamburg, germany: **54**(5):1901154. ISSN 0903-1936, 1399-3003. doi:10.1183/13993003.01154-2019.

Additional Data

Study	Study Year	Aim of the study	Study Region	TB burden setting
[175]	2011	Comparison between MIRU-VNTR and WGS for detecting genetic transmission	British Columbia, Canada	LI
[269]	2013	Use of WGS to corroborate three transmission clusters detected by previous investigations	Shangai, China	HI
[27]	2014	To perform a population-based study to detect transmission by WGS	Oxfordshire, UK	LI
[28]	2015	To perform a population-based study to detect transmission by WGS	Quebec, Canada	LI
[270]	2015	Use of WGS to corroborate one transmission clusters detected by previous investigations	Turku, Finland	LI
[202]	2015	Use WGS in order to determine SNP difference between close contacts TB patients	Karonga District, Malawi	HI
[40]	2016	Comparison between MIRU-VNTR and WGS for detecting genetic transmission	Switzerland	LI
[271]	2016	Comparison between MIRU-VNTR and WGS for detecting genetic transmission	Shangai, China	HI
[272]	2016	Use of WGS to corroborate one transmission clusters detected by previous investigations	London, UK	MI
[273]	2016	Use of WGS in order to improve an XDR-TB cluster investigation	London, UK	MI
[274]	2017	Use of WGS to corroborate one MDR transmission cluster between three different countries	Austria, Romania and Germany	LI, HI and LI
[275]	2017	Use of WGS to corroborate one transmission clusters detected by previous investigations	Southampton, UK	LI
[276]	2017	Use of WGS to corroborate one transmission clusters detected by previous investigations	Oslo, Norway	LI
[86]	2018	Use of WGS in order to improve a MDR-TB cluster investigation	Republic of Singapore	HI
[277]	2018	To perform a population-based study to detect transmission by WGS	Tasmania	LI
[147]	2018	To perform a population-based study to detect transmission by WGS	Shanghai, China	HI
[149]	2018	To perform a population-based study to detect transmission by WGS	The Netherlands	LI
[278]	2018	Use of WGS to corroborate transmission clusters detected by previous investigations	Florida, US	LI
[279]	2018	Use of WGS to corroborate transmission clusters detected by previous investigations	Belem, Brazil	HI
[280]	2018	Use of WGS to corroborate one transmission cluster between two different countries	The Netherlands and Denmark	LI
[281]	2019	To perform a population-based study to detect transmission by WGS	Scania, Sweden	LI
[160]	2019	To perform a population-based study to detect transmission by WGS	Papua New Guinea	HI
[176]	2019	To perform a population-based study to detect transmission by WGS	Yukon, Canada	LI
[282]	2019	Use of WGS to detect transmission clusters within MDR-TB cases	Lima, Peru	HI
[148]	2019	Use of WGS to detect transmission clusters within MDR-TB cases	Belgrade, Serbia	MI
[283]	2019	To perform a population-based study to detect transmission by WGS	Hamburg, Germany	LI

Table 10.1: Tuberculosis transmission studies that use whole genome sequencing as an primary or alternative tool for detecting transmision clusters (Part 1). Abbreviations; LI, Low incidence; MI, Middle incidence; HI, High incidence; WGS, Whole genome sequencing.

Study	Other genotyping method used	Epidemiological intervention	Study years	Isolates analized	MTBC lineage identified	SNP threshold applying for transmission	Clusters detected by WGS	Clustering rate
[175]	RFLPs, MIRU-VNTR	-	2006-2008	36	-	-	2	-
[269]	RFLPs, MIRU-VNTR	Contact tracing	2009-2010	32	L2	≤5 SNPs	4	32%
[27]	-	Contact tracing	2007-2012	247	-	≤12 SNPs	13	16%
[28]	-	-	1991-2013	163	L4	≤12 SNPs	11	92%
[270]	MIRU-VNTR	Contact tracing	2012-2013	12	L4	≤5 SNPs	1	-
[202]	-	Contact tracing	1997-2010	374	L3	≤10 SNPs	124	33%
[40]	MIRU-VNTR	-	2008-2012	90	L4	≤12 SNPs	17	47%
[271]	MIRU-VNTR	-	2009-2012	324	L2	≤12 SNPs	38	32%
[272]	MIRU-VNTR	Contact tracing	1998-2012	344	L4	≤12 SNPs	1	-
[273]	-	Contact tracing	2013-2015	6	-	≤5 SNPs	1	-
[274]	MIRU-VNTR	-	2009-2014	12	-	≤12 SNPs	3	-
[275]	MIRU-VNTR	Contact tracing	2011	17	-	≤5 SNPs	1	-
[276]	MIRU-VNTR	Contact tracing	2009-2014	22	L2	≤12 SNPs	1	-
[86]	-	Contact tracing	2012-2016	10	-	≤12 SNPs	1	-
[277]	-	-	2014-2016	29	L3	≤5 SNPs	2	33%
[147]	MIRU-VNTR	Contact tracing	2009-2015	218	L2	≤10 SNPs	44	68%
[149]	MIRU-VNTR	-	2016	535	L4	≤12 SNPs	29	14%
[278]	MIRU-VNTR	Contact tracing	2009-2015	21	-	≤5 SNPs	3	-
[279]	MIRU-VNTR	Contact tracing	1998-2011	63	L4	≤12 SNPs	26	-
[280]	RFLPs, MIRU-VNTR	-	1993-2016	40	L3	≤12 SNPs	5	42%
[281]	MIRU-VNTR	Contact tracing	2004-2014	93	L4	≤12 SNPs	18	56%
[160]	MIRU-VNTR	-	2010-2015	104	L2	≤8 SNPs	17	71%
[176]	MIRU-VNTR	Contact tracing	2005-2014	1316	-	≤5 SNPs	6	88%
[282]	MIRU-VNTR	-	2009-2012	61	L4	≤5 SNPs	6	54%
[148]	-	-	2008-2014	103	L4	≤5 SNPs	12	61%
[283]	MIRU-VNTR	Contact tracing	2005-2015	1171	-	≤5 SNPs	87	32%

Table 10.2: Tuberculosis transmission studies that use whole genome sequencing as an primary or alternative tool for detecting transmision clusters (Part 2).

Hospital	Province
Hospital General de Alicante	Alicante
Hospital de la Marina Baixa	Alicante
Hospital San Juan de Alicante	Alicante
Hospital Arnau de Vilanova	Alicante
Hospital de Dénia	Alicante
Hospital Vega Baja de Orihuela	Alicante
Hospital General Universitario de Elche	Alicante
Hospital Público Virgen de los Lirios	Alicante
Hospital General de Castellón	Castellón
Hospital de la Ribera	Valencia
Hospital Lluís Alcanyís de Xàtiva	Valencia
Hospital Francesc De Borja de Gandia	Valencia
Hospital Clínico Universitario de Valencia	Valencia
Hospital de Sagunto	Valencia
Hospital General de Valencia	Valencia
Hospital Universitario y Politécnico de Valencia	Valencia
Centro de Especialidades de Valencia	Valencia
Hospital Universitario Doctor Peset	Valencia

Table 10.3: Hospitals involved for the completion of this thesis

Additional Data Table 2. List of loci associated with drug resistance used to predict resistance profile.

Genomic position	Wild-type allele	Mutation allele	Genomic Region	Gene Name	Gene Alias	Gene Direction	Amino acid Change	Codon Change	Antibiotic	SNP Confidence*
6575	C	T	coding	Rv0005	*gyrB*	+	Arg446Cys	cgt/tgt	FQ	Low confidence
6620	G	C	coding	Rv0005	*gyrB*	+	Asp461His	gac/cac	FQ	High confidence
6620	G	A	coding	Rv0005	*gyrB*	+	Asp461Asn	gac/aac	FQ	High confidence
6621	A	C	coding	Rv0005	*gyrB*	+	Asp461Ala	gac/gcc	FQ	High confidence
6734	A	G	coding	Rv0005	*gyrB*	+	Asn499Asp	aac/gac	FQ	High confidence
6735	A	C	coding	Rv0005	*gyrB*	+	Asn499Thr	aac/acc	FQ	High confidence
6736	C	G	coding	Rv0005	*gyrB*	+	Asn499Lys	aac/aag	FQ	High confidence
6737	A	C	coding	Rv0005	*gyrB*	+	Thr500Pro	acc/ccc	FQ	High confidence
6738	C	A	coding	Rv0005	*gyrB*	+	Thr500Asn	acc/aac	FQ	High confidence
6741	A	T	coding	Rv0005	*gyrB*	+	Glu501Val	gaa/gta	FQ	High confidence
6742	A	T	coding	Rv0005	*gyrB*	+	Glu501Asp	gaa/gat	FQ	High confidence
6749	G	A	coding	Rv0005	*gyrB*	+	Ala504Thr	gcg/acg	FQ	Low confidence
6750	C	T	coding	Rv0005	*gyrB*	+	Ala504Val	gcg/gtg	FQ	High confidence
7563	G	T	coding	Rv0006	*gyrA*	+	Gly88Cys	ggc/tgc	FQ	High confidence
7564	G	C	coding	Rv0006	*gyrA*	+	Gly88Ala	ggc/gcc	FQ	High confidence
7566	G	A	coding	Rv0006	*gyrA*	+	Asp89Asn	gac/aac	FQ	Low confidence
7570	C	T	coding	Rv0006	*gyrA*	+	Ala90Val	gcg/gtg	FQ	High confidence
7572	T	C	coding	Rv0006	*gyrA*	+	Ser91Pro	tcg/ccg	FQ	High confidence
7581	G	C	coding	Rv0006	*gyrA*	+	Asp94His	gac/cac	FQ	High confidence
7581	G	A	coding	Rv0006	*gyrA*	+	Asp94Asn	gac/aac	FQ	High confidence
7581	G	T	coding	Rv0006	*gyrA*	+	Asp94Tyr	gac/tac	FQ	High confidence
7582	A	G	coding	Rv0006	*gyrA*	+	Asp94Gly	gac/ggc	FQ	High confidence
7582	A	C	coding	Rv0006	*gyrA*	+	Asp94Ala	gac/gcc	FQ	High confidence
7582	A	T	coding	Rv0006	*gyrA*	+	Asp94Val	gac/gtc	FQ	Low confidence
575729	C	T	coding	Rv0486	*mshA*	+	Gln128STOP	cag/tag	ETH	Low confidence
576164	C	T	coding	Rv0486	*mshA*	+	Arg273Cys	cgc/tgc	ETH	Low confidence
576242	G	T	coding	Rv0486	*mshA*	+	Gly299Cys	ggc/tgc	ETH	Low confidence
576338	C	T	coding	Rv0486	*mshA*	+	Gln331STOP	cag/tag	ETH	Low confidence
576414	G	A	coding	Rv0486	*mshA*	+	Gly356Asp	ggc/gac	ETH	Low confidence
576429	A	C	coding	Rv0486	*mshA*	+	Glu361Ala	gag/gcg	ETH	Low confidence
760314	G	T	coding	Rv0667	*rpoB*	+	Val170Phe	gtc/ttc	RIF	High confidence
761004	A	G	coding	Rv0667	*rpoB*	+	Thr400Ala	acc/gcc	RIF	Low confidence
761093	G	C	coding	Rv0667	*rpoB*	+	Gln429His	cag/cac	RIF	Low confidence
761095	T	C	coding	Rv0667	*rpoB*	+	Leu430Pro	ctg/ccg	RIF	Low confidence
761095	T	G	coding	Rv0667	*rpoB*	+	Leu430Arg	ctg/cgg	RIF	Low confidence
761098	G	T	coding	Rv0667	*rpoB*	+	Ser431Ile	agc/atc	RIF	Low confidence
761098	G	C	coding	Rv0667	*rpoB*	+	Ser431Thr	agc/acc	RIF	Low confidence
761100	C	A	coding	Rv0667	*rpoB*	+	Gln432Lys	caa/aaa	RIF	Low confidence
761101	A	T	coding	Rv0667	*rpoB*	+	Gln432Leu	caa/cta	RIF	High confidence
761101	A	C	coding	Rv0667	*rpoB*	+	Gln432Pro	caa/cca	RIF	Low confidence
761108	G	T	coding	Rv0667	*rpoB*	+	Met434Ile	atg/att	RIF	Low confidence
761109	G	T	coding	Rv0667	*rpoB*	+	Asp435Tyr	gac/tac	RIF	High confidence
761110	A	G	coding	Rv0667	*rpoB*	+	Asp435Gly	gac/ggc	RIF	High confidence
761110	A	T	coding	Rv0667	*rpoB*	+	Asp435Val	gac/gtc	RIF	High confidence

761111	C	G	coding	Rv0667	*rpoB*	+	Asp435Glu	gac/gag	RIF	Low confidence
761120	C	G	coding	Rv0667	*rpoB*	+	Asn438Lys	aac/aag	RIF	Low confidence
761128	C	T	coding	Rv0667	*rpoB*	+	Ser441Leu	tcg/ttg	RIF	Low confidence
761128	C	G	coding	Rv0667	*rpoB*	+	Ser441Trp	tcg/tgg	RIF	Low confidence
761139	C	A	coding	Rv0667	*rpoB*	+	His445Asn	cac/aac	RIF	High confidence
761139	C	G	coding	Rv0667	*rpoB*	+	His445Asp	cac/gac	RIF	High confidence
761139	C	T	coding	Rv0667	*rpoB*	+	His445Tyr	cac/tac	RIF	High confidence
761140	A	C	coding	Rv0667	*rpoB*	+	His445Pro	cac/ccc	RIF	High confidence
761140	A	G	coding	Rv0667	*rpoB*	+	His445Arg	cac/cgc	RIF	High confidence
761140	A	T	coding	Rv0667	*rpoB*	+	His445Leu	cac/ctc	RIF	Low confidence
761141	C	A	coding	Rv0667	*rpoB*	+	His445Gln	cac/caa	RIF	Low confidence
761154	T	G	coding	Rv0667	*rpoB*	+	Ser450Ala	tcg/gcg	RIF	Low confidence
761155	C	G	coding	Rv0667	*rpoB*	+	Ser450Trp	tcg/tgg	RIF	High confidence
761155	C	T	coding	Rv0667	*rpoB*	+	Ser450Leu	tcg/ttg	RIF	High confidence
761161	T	C	coding	Rv0667	*rpoB*	+	Leu452Pro	ctg/ccg	RIF	High confidence
761277	A	T	coding	Rv0667	*rpoB*	+	Ile491Phe	atc/ttc	RIF	High confidence
781687	A	G	coding	Rv0682	*rpsL*	+	Lys43Arg	aag/agg	RIF	High confidence
781821	A	C	coding	Rv0682	*rpsL*	+	Lys88Gln	aag/cag	RIF	Low confidence
781822	A	G	coding	Rv0682	*rpsL*	+	Lys88Arg	aag/agg	RIF	High confidence
801268	T	C	coding	Rv0701	*rplC*	+	Cys154Arg	tgt/cgt	LZD	Low confidence
1472337	C	T	ribosomal	MTB000019	*rrs*	+	---	-	STR	Low confidence
1472358	C	T	ribosomal	MTB000019	*rrs*	+	---	-	STR	Low confidence
1472359	A	C	ribosomal	MTB000019	*rrs*	+	---	-	STR	Low confidence
1472362	C	T	ribosomal	MTB000019	*rrs*	+	---	-	STR	Low confidence
1472750	C	A	ribosomal	MTB000019	*rrs*	+	---	-	STR	Low confidence
1472751	A	G	ribosomal	MTB000019	*rrs*	+	---	-	STR	Low confidence
1472752	A	T	ribosomal	MTB000019	*rrs*	+	---	-	STR	Low confidence
1473246	A	G	ribosomal	MTB000019	*rrs*	+	---	-	AMK, KAN, CM	High confidence
1473247	C	T	ribosomal	MTB000019	*rrs*	+	---	-	AMK, KAN, CM	High confidence
1473329	G	T	ribosomal	MTB000019	*rrs*	+	---	-	AMK, KAN, CM	High confidence
1475956	G	T	ribosomal	MTB000020	*rrl*	+	---	-	LZD	Low confidence
1476471	G	T	ribosomal	MTB000020	*rrl*	+	---	-	LZD	Low confidence
1673423	G	T	intergenic	Rv1483	*fabG1*	+	---	-	INH	Low confidence
1673424	A	G	intergenic	Rv1483	*fabG1*	+	---	-	INH	Low confidence
1673425	C	T	intergenic	Rv1483	*fabG1*	+	---	-	INH	High confidence
1673432	T	A	intergenic	Rv1483	*fabG1*	+	---	-	INH	High confidence
1673432	T	C	intergenic	Rv1483	*fabG1*	+	---	-	INH	High confidence
1674481	T	G	coding	Rv1484	*inhA*	+	Ser94Ala	tcg/gcg	INH, ETH	High confidence
1674782	T	C	coding	Rv1484	*inhA*	+	Ile194Thr	atc/acc	INH, ETH	Low confidence
1833909	A	C	coding	Rv1630	*rpsA*	+	Asp123Ala	gac/gcc	PZA	Low confidence
1834325	G	A	coding	Rv1630	*rpsA*	+	Val262Met	gtg/atg	PZA	Low confidence
1917946	C	T	coding	Rv1694	*tlyA*	+	Arg3STOP	cga/tga	CM	Low confidence
1917979	C	T	coding	Rv1694	*tlyA*	+	Arg14Trp	cgg/tgg	CM	Low confidence
1917991	C	T	coding	Rv1694	*tlyA*	+	Arg18STOP	cga/tga	CM	Low confidence
1918003	C	T	coding	Rv1694	*tlyA*	+	Gln22STOP	cag/tag	CM	Low confidence
1918139	C	A	coding	Rv1694	*tlyA*	+	Ala67Glu	gcg/gag	CM	Low confidence
1918144	A	G	coding	Rv1694	*tlyA*	+	Lys69Glu	aaa/gaa	CM	Low confidence
1918211	C	A	coding	Rv1694	*tlyA*	+	Ala91Glu	gca/gaa	CM	Low confidence

1918292	T	C	coding	Rv1694	*tlyA*	+	Leu118Pro	ctg/ccg	CM	Low confidence
1918322	T	A	coding	Rv1694	*tlyA*	+	Val128Glu	gtg/gag	CM	Low confidence
1918388	T	C	coding	Rv1694	*tlyA*	+	Leu150Pro	ctg/ccg	CM	Low confidence
1918487	C	T	coding	Rv1694	*tlyA*	+	Pro183Leu	ccg/ctg	CM	Low confidence
1918489	C	A	coding	Rv1694	*tlyA*	+	Gln184Lys	cag/aag	CM	Low confidence
1918489	C	T	coding	Rv1694	*tlyA*	+	Gln184STOP	cag/tag	CM	Low confidence
1918494	T	G	coding	Rv1694	*tlyA*	+	Phe185Leu	ttt/ttg	CM	Low confidence
1918651	G	A	coding	Rv1694	*tlyA*	+	Glu238Lys	gag/aag	CM	Low confidence
2102240	G	A	coding	Rv1854c	*ndh*	-	Arg268His	cgc/cac	INH, ETH	Low confidence
2102715	A	G	coding	Rv1854c	*ndh*	-	Thr110Ala	acc/gcc	INH, ETH	Low confidence
2155167	C	G	coding	Rv1908c	*katG*	-	Ser315Arg	agc/agg	INH	Low confidence
2155167	C	A	coding	Rv1908c	*katG*	-	Ser315Arg	agc/aga	INH	Low confidence
2155168	G	C	coding	Rv1908c	*katG*	-	Ser315Thr	agc/acc	INH	High confidence
2155168	G	A	coding	Rv1908c	*katG*	-	Ser315Asn	agc/aac	INH	High confidence
2155168	G	T	coding	Rv1908c	*katG*	-	Ser315Ile	agc/atc	INH	High confidence
2155169	A	G	coding	Rv1908c	*katG*	-	Ser315Gly	agc/ggc	INH	High confidence
2155206	C	G	coding	Rv1908c	*katG*	-	Ser302Arg	agc/agg	INH	Low confidence
2155212	G	C	coding	Rv1908c	*katG*	-	Trp300Cys	tgg/tgc	INH	Low confidence
2155214	T	G	coding	Rv1908c	*katG*	-	Trp300Gly	tgg/ggg	INH	High confidence
2155222	G	T	coding	Rv1908c	*katG*	-	Gly297Val	ggc/gtc	INH	Low confidence
2155289	A	C	coding	Rv1908c	*katG*	-	Thr275Pro	acc/ccc	INH	High confidence
2155699	A	G	coding	Rv1908c	*katG*	-	Asn138Ser	aac/agc	INH	Low confidence
2288683	T	C	coding	Rv2043c	*pncA*	-	STOP187Arg	tga/cga	PZA	High confidence
2288697	T	C	coding	Rv2043c	*pncA*	-	Leu182Ser	ttg/tcg	PZA	High confidence
2288703	T	C	coding	Rv2043c	*pncA*	-	Val180Ala	gtc/gcc	PZA	High confidence
2288703	T	G	coding	Rv2043c	*pncA*	-	Val180Gly	gtc/ggc	PZA	High confidence
2288704	G	T	coding	Rv2043c	*pncA*	-	Val180Phe	gtc/ttc	PZA	Low confidence
2288718	T	C	coding	Rv2043c	*pncA*	-	Met175Thr	atg/acg	PZA	High confidence
2288719	A	G	coding	Rv2043c	*pncA*	-	Met175Val	atg/gtg	PZA	High confidence
2288727	T	C	coding	Rv2043c	*pncA*	-	Leu172Pro	ctg/ccg	PZA	Low confidence
2288730	C	T	coding	Rv2043c	*pncA*	-	Ala171Val	gcg/gtg	PZA	Low confidence
2288740	A	C	coding	Rv2043c	*pncA*	-	Thr168Pro	acc/ccc	PZA	High confidence
2288752	T	C	coding	Rv2043c	*pncA*	-	Ser164Pro	tcg/ccg	PZA	High confidence
2288754	T	C	coding	Rv2043c	*pncA*	-	Val163Ala	gtg/gcg	PZA	High confidence
2288757	G	A	coding	Rv2043c	*pncA*	-	Gly162Asp	ggt/gat	PZA	Low confidence
2288761	G	C	coding	Rv2043c	*pncA*	-	Ala161Pro	gcg/ccg	PZA	High confidence
2288764	A	C	coding	Rv2043c	*pncA*	-	Thr160Pro	aca/cca	PZA	High confidence
2288766	T	G	coding	Rv2043c	*pncA*	-	Leu159Arg	ctg/cgg	PZA	Low confidence
2288772	T	C	coding	Rv2043c	*pncA*	-	Val157Ala	gtg/gcg	PZA	High confidence
2288772	T	G	coding	Rv2043c	*pncA*	-	Val157Gly	gtg/ggg	PZA	High confidence
2288778	T	G	coding	Rv2043c	*pncA*	-	Val155Gly	gtg/ggg	PZA	High confidence
2288779	G	A	coding	Rv2043c	*pncA*	-	Val155Met	gtg/atg	PZA	High confidence
2288782	A	G	coding	Rv2043c	*pncA*	-	Arg154Gly	agg/ggg	PZA	Low confidence
2288805	C	T	coding	Rv2043c	*pncA*	-	Ala146Val	gcg/gtg	PZA	High confidence
2288805	C	A	coding	Rv2043c	*pncA*	-	Ala146Glu	gcg/gag	PZA	Low confidence
2288806	G	C	coding	Rv2043c	*pncA*	-	Ala146Pro	gcg/ccg	PZA	High confidence
2288806	G	A	coding	Rv2043c	*pncA*	-	Ala146Thr	gcg/acg	PZA	Low confidence
2288817	C	A	coding	Rv2043c	*pncA*	-	Thr142Lys	acg/aag	PZA	High confidence

2288817	C	T	coding	Rv2043c	*pncA*	-	Thr142Met	acg/atg	PZA	High confidence
2288818	A	G	coding	Rv2043c	*pncA*	-	Thr142Ala	acg/gcg	PZA	High confidence
2288820	A	C	coding	Rv2043c	*pncA*	-	Gln141Pro	cag/ccg	PZA	Low confidence
2288821	C	T	coding	Rv2043c	*pncA*	-	Gln141STOP	cag/tag	PZA	Low confidence
2288823	G	C	coding	Rv2043c	*pncA*	-	Arg140Pro	cgc/ccc	PZA	High confidence
2288826	T	G	coding	Rv2043c	*pncA*	-	Val139Gly	gtg/ggg	PZA	High confidence
2288827	G	C	coding	Rv2043c	*pncA*	-	Val139Leu	gtg/ctg	PZA	High confidence
2288828	T	G	coding	Rv2043c	*pncA*	-	Cys138Trp	tgt/tgg	PZA	High confidence
2288830	T	C	coding	Rv2043c	*pncA*	-	Cys138Arg	tgt/cgt	PZA	High confidence
2288832	A	G	coding	Rv2043c	*pncA*	-	His137Arg	cat/cgt	PZA	High confidence
2288832	A	C	coding	Rv2043c	*pncA*	-	His137Pro	cat/cct	PZA	High confidence
2288833	C	G	coding	Rv2043c	*pncA*	-	His137Asp	cat/gat	PZA	High confidence
2288836	G	A	coding	Rv2043c	*pncA*	-	Asp136Asn	gat/aat	PZA	Low confidence
2288836	G	T	coding	Rv2043c	*pncA*	-	Asp136Tyr	gat/tat	PZA	Low confidence
2288838	C	A	coding	Rv2043c	*pncA*	-	Thr135Asn	acc/aac	PZA	High confidence
2288839	A	C	coding	Rv2043c	*pncA*	-	Thr135Pro	acc/ccc	PZA	High confidence
2288841	C	T	coding	Rv2043c	*pncA*	-	Ala134Val	gcc/gtc	PZA	High confidence
2288844	T	C	coding	Rv2043c	*pncA*	-	Ile133Thr	att/act	PZA	Low confidence
2288847	G	C	coding	Rv2043c	*pncA*	-	Gly132Ala	ggt/gct	PZA	High confidence
2288847	G	A	coding	Rv2043c	*pncA*	-	Gly132Asp	ggt/gat	PZA	High confidence
2288848	G	T	coding	Rv2043c	*pncA*	-	Gly132Cys	ggt/tgt	PZA	High confidence
2288848	G	A	coding	Rv2043c	*pncA*	-	Gly132Ser	ggt/agt	PZA	Low confidence
2288853	T	C	coding	Rv2043c	*pncA*	-	Val130Ala	gtg/gcg	PZA	High confidence
2288853	T	G	coding	Rv2043c	*pncA*	-	Val130Gly	gtg/ggg	PZA	High confidence
2288857	G	T	coding	Rv2043c	*pncA*	-	Asp129Tyr	gat/tat	PZA	High confidence
2288859	T	G	coding	Rv2043c	*pncA*	-	Val128Gly	gtc/ggc	PZA	Low confidence
2288868	T	G	coding	Rv2043c	*pncA*	-	Val125Gly	gtc/ggc	PZA	High confidence
2288869	G	T	coding	Rv2043c	*pncA*	-	Val125Phe	gtc/ttc	PZA	High confidence
2288874	G	C	coding	Rv2043c	*pncA*	-	Arg123Pro	cgc/ccc	PZA	High confidence
2288880	G	C	coding	Rv2043c	*pncA*	-	Arg121Pro	cgg/ccg	PZA	High confidence
2288883	T	A	coding	Rv2043c	*pncA*	-	Leu120Gln	ctg/cag	PZA	High confidence
2288883	T	C	coding	Rv2043c	*pncA*	-	Leu120Pro	ctg/ccg	PZA	High confidence
2288883	T	G	coding	Rv2043c	*pncA*	-	Leu120Arg	ctg/cgg	PZA	Low confidence
2288885	G	A	coding	Rv2043c	*pncA*	-	Trp119STOP	tgg/tga	PZA	Low confidence
2288886	G	C	coding	Rv2043c	*pncA*	-	Trp119Ser	tgg/tcg	PZA	High confidence
2288886	G	A	coding	Rv2043c	*pncA*	-	Trp119STOP	tgg/tag	PZA	High confidence
2288887	T	C	coding	Rv2043c	*pncA*	-	Trp119Arg	tgg/cgg	PZA	High confidence
2288887	T	G	coding	Rv2043c	*pncA*	-	Trp119Gly	tgg/ggg	PZA	High confidence
2288895	T	G	coding	Rv2043c	*pncA*	-	Leu116Arg	ctg/cgg	PZA	High confidence
2288895	T	C	coding	Rv2043c	*pncA*	-	Leu116Pro	ctg/ccg	PZA	High confidence
2288902	A	C	coding	Rv2043c	*pncA*	-	Thr114Pro	acg/ccg	PZA	Low confidence
2288920	G	A	coding	Rv2043c	*pncA*	-	Gly108Arg	gga/aga	PZA	High confidence
2288920	G	C	coding	Rv2043c	*pncA*	-	Gly108Arg	gga/cga	PZA	High confidence
2288928	G	A	coding	Rv2043c	*pncA*	-	Gly105Asp	ggc/gac	PZA	High confidence
2288930	C	A	coding	Rv2043c	*pncA*	-	Ser104Arg	agc/aga	PZA	High confidence
2288931	G	T	coding	Rv2043c	*pncA*	-	Ser104Ile	agc/atc	PZA	High confidence
2288933	C	G	coding	Rv2043c	*pncA*	-	Tyr103STOP	tac/tag	PZA	High confidence
2288934	A	G	coding	Rv2043c	*pncA*	-	Tyr103Cys	tac/tgc	PZA	High confidence

2288934	A	C	coding	Rv2043c	*pncA*	-	Tyr103Ser	tac/tcc	PZA	Low confidence
2288935	T	G	coding	Rv2043c	*pncA*	-	Tyr103Asp	tac/gac	PZA	High confidence
2288938	G	A	coding	Rv2043c	*pncA*	-	Ala102Thr	gcg/acg	PZA	High confidence
2288944	A	G	coding	Rv2043c	*pncA*	-	Thr100Ala	acc/gcc	PZA	High confidence
2288944	A	C	coding	Rv2043c	*pncA*	-	Thr100Pro	acc/ccc	PZA	High confidence
2288945	C	A	coding	Rv2043c	*pncA*	-	Tyr99STOP	tac/taa	PZA	High confidence
2288952	G	C	coding	Rv2043c	*pncA*	-	Gly97Ala	ggt/gct	PZA	High confidence
2288952	G	A	coding	Rv2043c	*pncA*	-	Gly97Asp	ggt/gat	PZA	High confidence
2288953	G	A	coding	Rv2043c	*pncA*	-	Gly97Ser	ggt/agt	PZA	Low confidence
2288954	G	A	coding	Rv2043c	*pncA*	-	Lys96Asn	aag/aac	PZA	High confidence
2288955	A	G	coding	Rv2043c	*pncA*	-	Lys96Arg	aag/agg	PZA	High confidence
2288955	A	C	coding	Rv2043c	*pncA*	-	Lys96Thr	aag/acg	PZA	High confidence
2288956	A	C	coding	Rv2043c	*pncA*	-	Lys96Gln	aag/cag	PZA	High confidence
2288956	A	G	coding	Rv2043c	*pncA*	-	Lys96Glu	aag/gag	PZA	Low confidence
2288957	C	G	coding	Rv2043c	*pncA*	-	Tyr95STOP	tac/tag	PZA	High confidence
2288960	C	A	coding	Rv2043c	*pncA*	-	Phe94Leu	ttc/tta	PZA	High confidence
2288960	C	G	coding	Rv2043c	*pncA*	-	Phe94Leu	ttc/ttg	PZA	High confidence
2288961	T	G	coding	Rv2043c	*pncA*	-	Phe94Cys	ttc/tgc	PZA	High confidence
2288961	T	C	coding	Rv2043c	*pncA*	-	Phe94Ser	ttc/tcc	PZA	High confidence
2288962	T	C	coding	Rv2043c	*pncA*	-	Phe94Leu	ttc/ctc	PZA	High confidence
2288971	G	T	coding	Rv2043c	*pncA*	-	Glu91STOP	gag/tag	PZA	High confidence
2288982	C	T	coding	Rv2043c	*pncA*	-	Thr87Met	acg/atg	PZA	Low confidence
2288988	T	G	coding	Rv2043c	*pncA*	-	Leu85Arg	ctg/cgg	PZA	High confidence
2288988	T	C	coding	Rv2043c	*pncA*	-	Leu85Pro	ctg/ccg	PZA	Low confidence
2288997	A	G	coding	Rv2043c	*pncA*	-	His82Arg	cat/cgt	PZA	High confidence
2288998	C	G	coding	Rv2043c	*pncA*	-	His82Asp	cat/gat	PZA	High confidence
2289000	T	G	coding	Rv2043c	*pncA*	-	Phe81Cys	ttc/tgc	PZA	High confidence
2289000	T	C	coding	Rv2043c	*pncA*	-	Phe81Ser	ttc/tcc	PZA	High confidence
2289001	T	G	coding	Rv2043c	*pncA*	-	Phe81Val	ttc/gtc	PZA	High confidence
2289016	A	C	coding	Rv2043c	*pncA*	-	Thr76Pro	act/cct	PZA	Low confidence
2289028	T	C	coding	Rv2043c	*pncA*	-	Cys72Arg	tgc/cgc	PZA	High confidence
2289029	T	A	coding	Rv2043c	*pncA*	-	His71Gln	cat/caa	PZA	High confidence
2289030	A	G	coding	Rv2043c	*pncA*	-	His71Arg	cat/cgt	PZA	High confidence
2289031	C	T	coding	Rv2043c	*pncA*	-	His71Tyr	cat/tat	PZA	Low confidence
2289036	C	T	coding	Rv2043c	*pncA*	-	Pro69Leu	cca/cta	PZA	Low confidence
2289038	G	C	coding	Rv2043c	*pncA*	-	Trp68Cys	tgg/tgc	PZA	High confidence
2289038	G	T	coding	Rv2043c	*pncA*	-	Trp68Cys	tgg/tgt	PZA	High confidence
2289039	G	A	coding	Rv2043c	*pncA*	-	Trp68STOP	tgg/tag	PZA	High confidence
2289039	G	C	coding	Rv2043c	*pncA*	-	Trp68Ser	tgg/tcg	PZA	Low confidence
2289040	T	C	coding	Rv2043c	*pncA*	-	Trp68Arg	tgg/cgg	PZA	Low confidence
2289040	T	G	coding	Rv2043c	*pncA*	-	Trp68Gly	tgg/ggg	PZA	Low confidence
2289043	T	C	coding	Rv2043c	*pncA*	-	Ser67Pro	tcg/ccg	PZA	High confidence
2289050	T	G	coding	Rv2043c	*pncA*	-	Tyr64STOP	tat/tag	PZA	High confidence
2289052	T	G	coding	Rv2043c	*pncA*	-	Tyr64Asp	tat/gat	PZA	Low confidence
2289054	A	G	coding	Rv2043c	*pncA*	-	Asp63Gly	gac/ggc	PZA	Low confidence
2289057	C	T	coding	Rv2043c	*pncA*	-	Pro62Leu	ccg/ctg	PZA	High confidence
2289061	A	C	coding	Rv2043c	*pncA*	-	Thr61Pro	aca/cca	PZA	Low confidence
2289068	C	A	coding	Rv2043c	*pncA*	-	Phe58Leu	ttc/tta	PZA	High confidence

2289068	C	G	coding	Rv2043c	*pncA*	-	Phe58Leu	ttc/ttg	PZA	High confidence
2289070	T	C	coding	Rv2043c	*pncA*	-	Phe58Leu	ttc/ctc	PZA	High confidence
2289071	C	G	coding	Rv2043c	*pncA*	-	His57Gln	cac/ag	PZA	High confidence
2289072	A	G	coding	Rv2043c	*pncA*	-	His57Arg	cac/cgc	PZA	High confidence
2289072	A	C	coding	Rv2043c	*pncA*	-	His57Pro	cac/ccc	PZA	High confidence
2289073	C	G	coding	Rv2043c	*pncA*	-	His57Asp	cac/gac	PZA	Low confidence
2289073	C	T	coding	Rv2043c	*pncA*	-	His57Tyr	cac/tac	PZA	Low confidence
2289081	C	G	coding	Rv2043c	*pncA*	-	Pro54Arg	ccg/cgg	PZA	High confidence
2289081	C	A	coding	Rv2043c	*pncA*	-	Pro54Gln	ccg/cag	PZA	High confidence
2289081	C	T	coding	Rv2043c	*pncA*	-	Pro54Leu	ccg/ctg	PZA	Low confidence
2289082	C	T	coding	Rv2043c	*pncA*	-	Pro54Ser	ccg/tcg	PZA	High confidence
2289089	C	A	coding	Rv2043c	*pncA*	-	His51Gln	cac/caa	PZA	High confidence
2289090	A	G	coding	Rv2043c	*pncA*	-	His51Arg	cac/cgc	PZA	High confidence
2289090	A	C	coding	Rv2043c	*pncA*	-	His51Pro	cac/ccc	PZA	High confidence
2289091	C	T	coding	Rv2043c	*pncA*	-	His51Tyr	cac/tac	PZA	High confidence
2289096	A	C	coding	Rv2043c	*pncA*	-	Asp49Ala	gac/gcc	PZA	Low confidence
2289096	A	G	coding	Rv2043c	*pncA*	-	Asp49Gly	gac/ggc	PZA	Low confidence
2289097	G	A	coding	Rv2043c	*pncA*	-	Asp49Asn	gac/aac	PZA	High confidence
2289100	A	G	coding	Rv2043c	*pncA*	-	Lys48Glu	aag/gag	PZA	High confidence
2289100	A	T	coding	Rv2043c	*pncA*	-	Lys48STOP	aag/tag	PZA	High confidence
2289103	A	C	coding	Rv2043c	*pncA*	-	Thr47Pro	acc/ccc	PZA	High confidence
2289103	A	G	coding	Rv2043c	*pncA*	-	Thr47Ala	acc/gcc	PZA	Low confidence
2289108	T	G	coding	Rv2043c	*pncA*	-	Val45Gly	gtg/ggg	PZA	Low confidence
2289111	T	G	coding	Rv2043c	*pncA*	-	Val44Gly	gtc/ggc	PZA	High confidence
2289133	G	T	coding	Rv2043c	*pncA*	-	Glu37top	gaa/taa	PZA	High confidence
2289138	T	C	coding	Rv2043c	*pncA*	-	Leu35Pro	ctg/ccg	PZA	High confidence
2289140	C	G	coding	Rv2043c	*pncA*	-	Tyr34STOP	tac/tag	PZA	High confidence
2289150	T	G	coding	Rv2043c	*pncA*	-	Ile31Ser	atc/agc	PZA	High confidence
2289159	C	A	coding	Rv2043c	*pncA*	-	Ala28Asp	gcc/gac	PZA	Low confidence
2289162	T	C	coding	Rv2043c	*pncA*	-	Leu27Pro	ctg/ccg	PZA	High confidence
2289171	G	A	coding	Rv2043c	*pncA*	-	Gly24Asp	ggc/gac	PZA	High confidence
2289180	T	G	coding	Rv2043c	*pncA*	-	Val21Gly	gta/gga	PZA	High confidence
2289186	T	C	coding	Rv2043c	*pncA*	-	Leu19Pro	ctg/ccg	PZA	High confidence
2289193	G	A	coding	Rv2043c	*pncA*	-	Gly17Ser	ggc/agc	PZA	High confidence
2289200	C	A	coding	Rv2043c	*pncA*	-	Cys14STOP	tgc/tga	PZA	High confidence
2289201	G	A	coding	Rv2043c	*pncA*	-	Cys14Tyr	tgc/tac	PZA	Low confidence
2289202	T	C	coding	Rv2043c	*pncA*	-	Cys14Arg	tgc/cgc	PZA	High confidence
2289203	C	G	coding	Rv2043c	*pncA*	-	Phe13Leu	ttc/ttg	PZA	High confidence
2289204	T	C	coding	Rv2043c	*pncA*	-	Phe13Ser	ttc/tcc	PZA	Low confidence
2289206	C	G	coding	Rv2043c	*pncA*	-	Asp12Glu	gac/gag	PZA	High confidence
2289207	A	C	coding	Rv2043c	*pncA*	-	Asp12Ala	gac/gcc	PZA	High confidence
2289208	G	A	coding	Rv2043c	*pncA*	-	Asp12Asn	gac/aac	PZA	Low confidence
2289213	A	G	coding	Rv2043c	*pncA*	-	Gln10Arg	cag/cgg	PZA	Low confidence
2289213	A	C	coding	Rv2043c	*pncA*	-	Gln10Pro	cag/ccg	PZA	Low confidence
2289214	C	A	coding	Rv2043c	*pncA*	-	Gln10Lys	cag/aag	PZA	High confidence
2289216	T	C	coding	Rv2043c	*pncA*	-	Val9Ala	gtg/gcg	PZA	Low confidence
2289216	T	G	coding	Rv2043c	*pncA*	-	Val9Gly	gtg/ggg	PZA	Low confidence
2289218	C	A	coding	Rv2043c	*pncA*	-	Asp8Glu	gac/gaa	PZA	High confidence

2289219	A	C	coding	Rv2043c	*pncA*	-	Asp8Ala	gac/gcc	PZA	High confidence
2289219	A	G	coding	Rv2043c	*pncA*	-	Asp8Gly	gac/ggc	PZA	Low confidence
2289220	G	A	coding	Rv2043c	*pncA*	-	Asp8Asn	gac/aac	PZA	High confidence
2289222	T	G	coding	Rv2043c	*pncA*	-	Val7Gly	gtc/ggc	PZA	High confidence
2289222	T	A	coding	Rv2043c	*pncA*	-	Val7Asp	gtc/gac	PZA	Low confidence
2289223	G	T	coding	Rv2043c	*pncA*	-	Val7Phe	gtc/ttc	PZA	High confidence
2289225	T	C	coding	Rv2043c	*pncA*	-	Ile6Thr	atc/acc	PZA	High confidence
2289231	T	G	coding	Rv2043c	*pncA*	-	Leu4Ser	ttg/tcg	PZA	High confidence
2289231	T	G	coding	Rv2043c	*pncA*	-	Leu4Trp	ttg/tgg	PZA	Low confidence
2289234	C	A	coding	Rv2043c	*pncA*	-	Ala3Glu	gcg/gag	PZA	High confidence
2289235	G	C	coding	Rv2043c	*pncA*	-	Ala3Pro	gcg/ccg	PZA	Low confidence
2289239	G	A	coding	Rv2043c	*pncA*	-	Met1Ile	atg/ata	PZA	High confidence
2289239	G	T	coding	Rv2043c	*pncA*	-	Met1Ile	atg/att	PZA	Low confidence
2289240	T	A	coding	Rv2043c	*pncA*	-	Met1Lys	atg/aag	PZA	High confidence
2289240	T	C	coding	Rv2043c	*pncA*	-	Met1Thr	atg/acg	PZA	High confidence
2289248	T	C	intergenic	Rv2043c	*pncA*	-	---	-	PZA	High confidence
2289248	T	G	intergenic	Rv2043c	*pncA*	-	---	-	PZA	High confidence
2289252	A	G	intergenic	Rv2043c	*pncA*	-	---	-	PZA	High confidence
2289252	A	C	intergenic	Rv2043c	*pncA*	-	---	-	PZA	High confidence
2289252	A	T	intergenic	Rv2043c	*pncA*	-	---	-	PZA	High confidence
2715342	G	A	intergenic	Rv2416c	*eis*	-	---	-	KAN	Low confidence
2715346	C	T	intergenic	Rv2416c	*eis*	-	---	-	KAN	Low confidence
2726136	C	T	intergenic	Rv2428	*ahpC*	+	---	-	INH	Low confidence
2726145	G	A	intergenic	Rv2428	*ahpC*	+	---	-	INH	Low confidence
3073808	C	G	coding	Rv2764c	*thyA*	-	Arg222Gly	cgc/ggc	PAS	Low confidence
4241078	A	G	coding	Rv3793	*embC*	+	Ile406Val	atc/gtc	EMB	Low confidence
4243221	C	T	intergenic	Rv3794	*embA*	+	---	-	EMB	Low confidence
4243225	C	A	intergenic	Rv3794	*embA*	+	---	-	EMB	Low confidence
4243242	G	A	coding	Rv3794	*embA*	+	Asp4Asn	gac/aac	EMB	Low confidence
4243245	G	A	coding	Rv3794	*embA*	+	Gly5Ser	ggt/agt	EMB	Low confidence
4243833	G	A	coding	Rv3794	*embA*	+	Ala201Thr	gcg/acg	EMB	Low confidence
4244193	G	A	coding	Rv3794	*embA*	+	Gly321Ser	ggc/agc	EMB	Low confidence
4244281	G	A	coding	Rv3794	*embA*	+	Gly350Asp	ggc/gac	EMB	Low confidence
4244617	C	T	coding	Rv3794	*embA*	+	Ala462Val	gcg/gtg	EMB	Low confidence
4245730	A	C	coding	Rv3794	*embA*	+	Asp833Ala	gac/gcc	EMB	Low confidence
4246734	T	G	coding	Rv3795	*embB*	+	Leu74Arg	ctg/cgg	EMB	Low confidence
4247402	T	G	coding	Rv3795	*embB*	+	Ser297Ala	tcg/gcg	EMB	Low confidence
4247429	A	G	coding	Rv3795	*embB*	+	Met306Val	atg/gtg	EMB	High confidence
4247429	A	C	coding	Rv3795	*embB*	+	Met306Leu	atg/ctg	EMB	High confidence
4247430	T	C	coding	Rv3795	*embB*	+	Met306Thr	atg/acg	EMB	High confidence
4247431	G	A	coding	Rv3795	*embB*	+	Met306Ile	atg/ata	EMB	High confidence
4247431	G	C	coding	Rv3795	*embB*	+	Met306Ile	atg/atc	EMB	High confidence
4247431	G	T	coding	Rv3795	*embB*	+	Met306Ile	atg/att	EMB	High confidence
4247469	A	C	coding	Rv3795	*embB*	+	Tyr319Ser	tat/tct	EMB	Low confidence
4247495	G	T	coding	Rv3795	*embB*	+	Asp328Tyr	gat/tat	EMB	Low confidence
4247496	A	G	coding	Rv3795	*embB*	+	Asp328Gly	gat/ggt	EMB	Low confidence
4247507	T	C	coding	Rv3795	*embB*	+	Trp332Arg	tgg/cgg	EMB	Low confidence
4247513	T	C	coding	Rv3795	*embB*	+	Tyr334His	tac/cac	EMB	Low confidence

4247573	G	A	coding	Rv3795	*embB*	+	Asp354Asn	gac/aac	EMB	Low confidence
4247717	C	G	coding	Rv3795	*embB*	+	Leu402Val	ctg/gtg	EMB	Low confidence
4247723	C	T	coding	Rv3795	*embB*	+	Pro404Ser	ccg/tcg	EMB	Low confidence
4247729	G	A	coding	Rv3795	*embB*	+	Gly406Ser	ggc/agc	EMB	High confidence
4247729	G	T	coding	Rv3795	*embB*	+	Gly406Cys	ggc/tgc	EMB	High confidence
4247730	G	C	coding	Rv3795	*embB*	+	Gly406Ala	ggc/gcc	EMB	High confidence
4247730	G	A	coding	Rv3795	*embB*	+	Gly406Asp	ggc/gac	EMB	High confidence
4247863	C	G	coding	Rv3795	*embB*	+	Ile450Met	atc/atg	EMB	Low confidence
4247873	G	A	coding	Rv3795	*embB*	+	Ala454Thr	gcg/acg	EMB	Low confidence
4248002	C	A	coding	Rv3795	*embB*	+	Gln497Lys	cag/aag	EMB	Low confidence
4248003	A	G	coding	Rv3795	*embB*	+	Gln497Arg	cag/cgg	EMB	High confidence
4248747	G	A	coding	Rv3795	*embB*	+	Gly745Asp	ggc/gac	EMB	Low confidence
4249518	A	G	coding	Rv3795	*embB*	+	His1002Arg	cac/cgc	EMB	Low confidence
4326087	C	A	coding	Rv3854c	*ethA*	-	Arg463Ser	cgt/agt	ETH	Low confidence
4326236	G	A	coding	Rv3854c	*ethA*	-	Gly413Asp	ggt/gat	ETH	Low confidence
4326300	A	G	coding	Rv3854c	*ethA*	-	Thr392Ala	acg/gcg	ETH	Low confidence
4326320	G	A	coding	Rv3854c	*ethA*	-	Gly385Asp	ggc/gac	ETH	Low confidence
4326333	G	C	coding	Rv3854c	*ethA*	-	Ala381Pro	gcc/ccc	ETH	Low confidence
4326449	C	A	coding	Rv3854c	*ethA*	-	Thr342Lys	acg/aag	ETH	Low confidence
4326461	T	G	coding	Rv3854c	*ethA*	-	Ile338Ser	atc/agc	ETH	Low confidence
4326738	C	T	coding	Rv3854c	*ethA*	-	Gln246STOP	cag/tag	ETH	Low confidence
4326807	G	A	coding	Rv3854c	*ethA*	-	Glu223Lys	gag/aag	ETH	Low confidence
4326917	C	A	coding	Rv3854c	*ethA*	-	Thr186Lys	acg/aag	ETH	Low confidence
4327224	T	G	coding	Rv3854c	*ethA*	-	Tyr84Asp	tac/gac	ETH	Low confidence
4327301	A	C	coding	Rv3854c	*ethA*	-	Asp58Ala	gac/gcc	ETH	Low confidence
4327307	A	C	coding	Rv3854c	*ethA*	-	Asp56Ala	gac/gcc	ETH	Low confidence
4327322	C	T	coding	Rv3854c	*ethA*	-	Pro51Leu	ccc/ctc	ETH	Low confidence
4327346	G	A	coding	Rv3854c	*ethA*	-	Gly43Asp	ggc/gac	ETH	Low confidence
4327347	G	T	coding	Rv3854c	*ethA*	-	Gly43Cys	ggc/tgc	ETH	Low confidence
4407604	C	A	coding	Rv3919c	*gidB*	-	Ala200Glu	gcg/gag	STR	Low confidence
4407790	C	T	coding	Rv3919c	*gidB*	-	Ala138Val	gcg/gtg	STR	Low confidence
4407824	C	T	coding	Rv3919c	*gidB*	-	Gln127STOP	caa/taa	STR	Low confidence
4407931	T	C	coding	Rv3919c	*gidB*	-	Leu91Pro	cta/cca	STR	Low confidence
4407940	T	C	coding	Rv3919c	*gidB*	-	Val88Ala	gta/gca	STR	Low confidence
4407992	G	A	coding	Rv3919c	*gidB*	-	Gly71Arg	gga/aga	STR	Low confidence
4408009	T	G	coding	Rv3919c	*gidB*	-	Val65Gly	gtc/ggc	STR	Low confidence
4408102	G	C	coding	Rv3919c	*gidB*	-	Gly34Ala	ggg/gcg	STR	Low confidence

Additional Data Table 3. Significant variant frequencies across all isolates. All values are given in percentage

Genomic Position	Wild-type Allele	Variant Allele	Jan 2014	March 2014	June 2014	Aug 2014	Oct 2014	Dec 2014	Jan 2015	April 2015	June 2015	Nov 2015	Dec 2015	June 2016	Oct 2016	Jan 2017
6742	A	C	0	0	0	0	0	0	0	0	0	0	0	100	100	100
36471	C	G	0	0	57.14	0	0	0	0	0	0	0	60	0	0	0
80564	C	G	0	0	0	0	0	0	0	0	14.58	0	0	0	0	0
130660	T	A	100	0	0	0	0	0	0	0	0	0	0	0	0	0
208321	C	G	0	0	0	0	0	0	0	0	0	85.71	0	0	0	0
232974	A	C	0	0	0	0	0	0	0	93.33	85.35	28.22	0	0	0	0
580797	A	G	0	0	0	0	0	0	0	0	0	66.67	0	0	33.33	0
623273	G	A	0	0	0	0	100	0	0	0	0	0	0	0	0	0
650379	C	T	0	0	0	0	0	0	0	0	0	0	0	0	0	0
699980	G	C	0	0	0	15	0	55.56	0	0	0	0	0	0	0	0
725302	C	T	0	20.93	0	0	0	0	0	0	0	0	0	0	0	0
761152	T	A	0	0	0	0	11.82	0	0	7.25	0	0	0	0	0	6.41
761277	A	T	100	100	99.24	99.56	99.07	99.32	100	100	98.63	98.97	100	98.97	100	99.32
851731	T	G	0	0	0	0	23.48	0	0	0	0	0	0	0	0	0
852638	C	G	0	0	0	0	0	50.81	71.91	0	0	0	0	0	0	0
854253	C	G	0	0	0	30.77	0	0	0	0	27.27	0	0	0	0	0
916546	G	T	0	0	0	0	0	19.05	0	0	0	0	0	0	0	0
939197	G	C	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1002273	G	A	100	100	100	100	100	100	100	100	100	100	100	98.37	100	100
1131770	C	T	0	0	0	0	14.97	0	0	0	0	0	0	0	0	0
1157076	C	A	100	100	99.25	98.97	100	100	99.07	100	99.35	100	100	100	100	100
1230842	T	C	0	0	0	0	0	0	0	0	0	0	0	99.19	99.64	100
1474752	T	C	100	100	100	98.68	95.83	97.83	100	95.65	100	100	98.04	100	100	99.49
1502314	C	T	99.48	100	100	100	100	100	100	100	100	98.35	100	99.05	100	100
1514788	G	T	0	25.66	0	0	0	0	0	0	0	0	0	0	0	0
1519823	A	G	0	0	0	0	0	0	0	0	100	75	0	0	0	0
1524571	A	T	0	0	0	100	0	0	0	0	100	100	100	0	75	0
1543413	C	G	0	0	0	0	0	0	0	0	37.50	0	0	0	0	0
1917986	T	G	0	0	0	0	0	0	0	0	0	0	0	99.08	98.07	100
1960078	C	A	0	0	0	0	0	0	0	84.47	0	0	0	0	0	0
1997457	G	C	0	0	0	36.36	0	0	0	0	20	20	0	0	27.78	0
2074547	G	T	0	0	0	0	0	0	0	0	0	0	0	0	100	0
2094917	T	A	0	0	0	0	0	0	0	0	0	0	0	0	21.43	0
2154827	C	-	99.00	0	0	0	67	41	38	0	8	71	93	72	98	97
2155295	C	G	0	98.06	40.74	97.14	24.39	52.21	65.08	95.89	82.95	28.32	0	0	0	0
2180818	T	C	0	0	0	0	0	0	0	0	0	80	0	0	0	0
2183360	C	T	0	0	0	0	0	0	0	0	0	0	0	49.07	0	0
2262857	C	T	100	100	100	100	100	100	100	100	0	100	100	100	100	100
2300456	G	A	0	0	0	0	0	0	0	0	0	51.97	0	0	0	0
2304044	T	A	0	0	0	0	0	0	0	0	0	19.85	0	0	0	0

2314650	T	C	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2339605	A	G	100	100	90.91	100	100	100	0	100	100	100	100	0	100	0
2363682	C	A	0	0	0	100	0	0	100	0	0	100	0	0	100	0
2370345	C	T	0	0	0	0	0	0	52	0	0	0	0	0	0	0
2482657	G	C	0	0	21.05	36.17	0	0	0	0	0	0	0	0	0	0
2525723	G	C	20	0	0	0	0	0	0	0	0	17.39	0	0	0	0
2527757	C	T	0	0	23.27	36.18	0	0	0	0	0	0	0	0	0	0
2534563	G	C	30	0	75	0	27.27	0	0	0	0	33.33	0	0	0	0
2536917	T	C	0	16.39	0	0	0	0	0	0	0	0	0	0	0	0
2542966	C	T	0	0	0	0	0	60	0	0	0	0	0	0	0	0
2551675	A	C	0	0	0	0	0	0	0	0	0	0	100	0	0	0
2726112	C	T	0	0	19.77	35.34	0	25.83	0	96.72	0	0	0	21.78	0	0
2726139	C	T	0	0	13.79	67.91	0	26	0	0	0	0	0	0	0	0
2726141	C	A/T	0	5.71	0	0	0	0	45.19	0	0	0	0	0	0	0
2726145	G	A	0	0	0	0	0	0	0	0	0	7.52	64.93	0	0	0
2726153	G	A	0	0	0	0	0	0	0	0	0	0	7	0	0	0
2726767	A	C	94.94	0	0	0	0	0	0	0	0	0	0	0	0	0
2786787	T	G	0	0	0	0	0	0	0	0	0	0	75	0	0	0
2794585	C	A	100	100	100	100	99.32	100	100	100	100	100	100	100	100	100
2879778	G	T	100	100	97.22	99.02	100	98.41	100	100	100	100	98.15	92.86	98.91	97.87
2895066	C	G	0	0	0	0	0	0	0	96.30	82.96	30.65	0	0	0	0
2895750	G	A	0	0	0	0	75	40.58	41.73	0	16.34	66.83	97.47	100	100	100
3007115	C	A	0	0	0	66.67	0	0	0	0	0	0	0	0	0	0
3007143	T	A	0	0	100	0	0	0	0	0	0	0	0	0	100	0
3007407	C	G	0	0	0	0	0	0	0	0	30.25	0	0	0	0	0
3008814	C	T	15.85	0	0	0	0	0	0	0	0	0	0	0	0	0
3008839	G	T	0	34.78	0	0	0	0	0	0	0	0	0	0	0	0
3045144	C	T	0	0	0	0	0	0	0	0	0	64.52	81.82	100	100	100
3096287	G	C	100	100	0	100	0	100	0	0	0	0	0	0	0	0
3096295	G	C	0	0	100	100	0	0	0	75	0	80	100	75	0	100
3160000	A	C	71.43	0	60	87.50	0	100	80	50	0	71.43	85.71	0	87.50	0
3173107	G	A	100	100	100	100	100	98.23	100	100	98.65	99.40	100	100	100	100
3247866	C	A	0	0	0	0	0	0	0	0	0	60	0	0	0	0
3247867	A	G	80	0	0	0	0	0	0	0	0	60	0	0	0	0
3247868	A	G	100	0	0	0	0	42.86	0	0	0	0	0	0	0	0
3247869	A	G	100	0	0	0	0	66.67	0	0	60	0	0	0	0	0
3349257	T	C	0	0	0	0	16.28	0	0	0	0	0	0	0	0	0
3415194	G	A	0	0	0	0	0	0	22.22	0	16.13	16.13	0	30.77	13.16	0
3498811	A	G	0	0	0	0	0	0	0	0	20.67	0	0	0	0	0
3580637	T	C	62.50	0	0	0	50	50	0	0	0	0	0	0	0	0
3596631	G	A	0	0	0	0	0	0	0	0	14.97	0	0	0	0	0
3664945	C	T	99.17	100	99.40	100	98.90	99.12	100	96.84	100	100	99.37	100	100	100
3854063	G	T	0	0	0	0	0	0	0	0	0	0	0	0	30	0
3862473	A	G	0	0	0	71.43	77.78	0	0	0	0	0	0	0	0	0
3878567	G	C	0	28.57	0	0	0	0	0	0	0	0	0	0	16.67	23.53

3915436	C	T	0	0	0	0	0	57.14	61.90	0	0	0	0	0	0	0
4026874	T	C	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4117169	C	G	0	60	0	50	50	66.67	0	37.50	0	54.55	55.56	50	0	75
4247429	A	G	0	0	0	0	0	0	0	0	0	0	10.26	0	0	0
4247495	G	T	0	0	0	0	0	0	0	0	0	54.97	88.82	100	100	100
4255385	C	G	0	0	0	0	0	0	0	0	0	62.84	86.62	100	100	100
4269341	C	T	0	0	0	0	0	0	0	97.03	86.79	23.53	0	0	0	0
4338596	T	G	0	0	0	50	40	0	50	0	0	0	50	0	21.43	0
4353414	G	C	100	0	0	100	0	0	0	0	0	0	0	0	0	0
4358702	A	G	71.43	0	0	0	100	0	0	0	0	0	0	0	0	0
4373008	T	C	0	0	0	0	42.86	0	0	0	0	60	0	0	33.33	0

Resumen en castellano

Introducción

La tuberculosis (TB) es una enfermedad infecciosa causada por bacterias patógenas pertenecientes al complejo de *Mycobacterium tuberculosis* (MTBC). Aunque es curable y prevenible, la TB es una de las diez principales causas de muerte en todo el mundo según la Organización Mundial de la Salud (OMS). Se estima que aproximadamente una cuarta parte de la población mundial tiene TB latente, que se caracteriza por no presentar los síntomas típicos de TB y por una enfermedad presumiblemente no transmisible.

El MTBC es un grupo de bacterias de lento crecimiento pertenecientes al género *Mycobacterium*. Aunque la mayoría de las especies de *Mycobacterium* (más de 170 especies) son organismos de vida libre que no causan enfermedad, algunas de ellas pueden causar infecciones en humanos. Los principales agentes causantes de la tuberculosis en humanos son las micobacterias *Mycobacterium tuberculosis* y *Mycobacterium africanum*. Otras micobacterias pertenecientes al MTBC, como por ejemplo, *M.bovis*, *M. caprae*, *M. microti*, *M. pinnipedii*, *M. orygis*, *M. suricattae* y *M.mungi*, son conocidas por causar infecciones a otros huéspedes animales. Además del MTBC, existen otras micobacterias no tuberculosas que causan la enfermedad, dentro de estas se incluyen las especies *M.avium* y *M.kansasii*, entre otras. Una característica fisiológica importante del género *Mycobacterium* es la estructura de su pared celular. Esta pared consiste en una bicapa lipídica hecha de ácidos grasos de cadena larga (ácidos micólicos), que proporcionan una barrera impermeable que confiere resistencia a diferentes compuestos nocivos y evita la deshidratación celular. Gracias a esta característica, las micobacterias son ácido-alcohol resistentes, por lo que necesitan un método de tinción especial llamado tinción de Ziehl-Neelsen. Todas las micobacterias son bacilos aeróbicos, con forma de bastón y que no forman esporas. Su tamaño varía de 0.2 a 0.6μm por 1.0 a 10μm, formando colonias que varían morfológicamente en textura y color entre especies.

Situación actual de la tuberculosis

En el 2018, la OMS estimó que 10 millones de personas se infectaron con TB. De estos, 1,45 millones de personas murieron a causa de esta enfermedad (484,000 casos relacionados con tuberculosis fármaco resistente, y 251,000 casos por una coinfección con VIH). Las regiones más afectadas son el sudeste asiático y el continente africano, con el 44En 2017, el sistema de vigilancia español reportó 4,483 casos de TB (tasa de incidencia equivalente a 9.6 casos por cada 100,000 habitantes), siendo Galicia, Catalunya y Asturias las regiones con la tasas más altas con 19.6, 12.9 y 10.8 casos por cada 100,000 personas, respectivamente. Particularmente, en la Comunidad Valenciana, se notificaron 424 casos de TB con una incidencia de 8,6 por cada 100.000 habitantes. Aunque el número de casos de TB ha disminuido desde 1990, la mayoría de los países del mundo todavía tienen una tasa de incidencia superior a 10 casos por cada 100.000 habitantes. Con el fin de reducir de reducir la incidencia de TB a nivel global, la OMS ha desarrollado una iniciativa llamada “The End TB Strategy”. Sus principales objetivos son reducir el número de muertes por TB en un 90% y disminuir la incidencia de TB en un 80% para 2030. Dichas estrategias destacan, entre otras, la importancia de un diagnóstico precoz y rápido de TB, pruebas precisas de sensibilidad a los fármacos administrados, aplicación de tratamientos apropiados y un control efectivo de la transmisión. Además, enfatiza el uso de la investigación para el desarrollo de nuevas pruebas diagnósticas, medicamentos, vacunas.

Infección y transmisión de la tuberculosis

La infección de TB comienza cuando el bacilo ingresa a los pulmones por inhalación y alcanza el espacio alveolar. Una vez allí, se encuentra con los macrófagos alveolares residentes, en donde la bacteria se replica. Una vez que se establece la infección, existen dos escenarios diferentes según la gravedad de la enfermedad: 1) el sistema inmunitario del huésped puede controlar el crecimiento bacteriano y, sin llegar a eliminar la infección inicial, el individuo infectado no desarrolla los síntomas de la enfermedad. Esta condición asintomática se conoce como infección de TB latente (LTBI). Sin embargo, 2) Entre el 5 y 15% de las personas infectadas, el sistema inmunitario falla y desarrolla una enfermedad activa, presentando síntomas leves, moderados o severos, o incluso desarrollando cavitaciones pulmonares. Los factores que desencadenan el progreso de la infección (de infecciones latentes a activas) y los cambios metabólicos bacterianos subyacentes siguen sin estar claros y podrían involucrar factores bacterianos, así como, elementos inmunológicos y clínicos de huésped. Los métodos más comunes para la detección de infección por TB activa son la inspección de esputos para la identificación de bacilos mediante microscopio, la siembra del cultivo. El diagnóstico de LTBI se realiza mediante la positividad de la ensayos basados en respuestas inmunológicas del huésped, como son la prueba cutánea de la tuberculina (TST), y pruebas de sangre de liberación del interferón gamma (IGRA). Durante muchos años, se ha establecido una dicotomía entre los estados clínicos de TB activa y latente para diferenciar entre aquellos pacientes que están infectados y pueden permanecer asintomáticos durante años o toda la vida, y aquellos que desarrollan síntomas típicos de TB (tos continua, fiebres, sudores nocturnos, fatiga, dolor de pecho). Sin embargo, existe evidencia científica que sugiere que la infección de TB se refleja mejor por la existencia de un amplio espectro de diferentes estados de infección, que están mediados por la naturaleza heterogénea y dinámica de la bacteria, además, de las respuestas inmunitarias del huésped en los granulomas. Una combinación de manifestaciones clínicas de la enfermedad, evidencia microbiológica y resultados de pruebas inmunológicas (como TST e IGRA) pueden ayudar a definir estos estados de la enfermedad. Por ejemplo, la tuberculosis subclínica se caracteriza por presentar resultados positivos de esputo o cultivo y ensayos inmunológicos, y a veces radiografías de tórax, pero sin desarrollar síntomas típicos de TB. Estos casos de TB son difíciles de detectar principalmente porque las personas asintomáticas no acuden a un centro de atención médica. El diagnóstico de estos casos solo es posible a las investigaciones de contacto, por lo que representan desafíos importantes para los sistemas de salud pública. Asimismo, aún se debate si estos pacientes asintomáticos pueden transmitir la enfermedad.

La TB se transmite por el aire a través de aerosoles, por lo que, el sitio de infección más común es el sistema respiratorio. Existen muchos factores asociados con la transmisión de TB, los cuales incluyen características biológicas relacionadas la bacteria y el huésped, así como factores sociales a nivel de población (por ejemplo, comportamientos culturales). Los factores del huésped están relacionados con la infecciosidad de la enfermedad, por lo tanto, las personas con síntomas más graves (por ejemplo, con más cavidades pulmonares profundas), tienen más probabilidades de transmitir TB. Otros determinantes que pueden alterar el nivel de infección incluyen el tabaquismo, alcoholismo, y la desnutrición, así como presentar comorbilidades como infecciones por VIH y diabetes. Además de esto, el retraso en el diagnóstico e inicio del tratamiento de TB, aumentan la probabilidad de transmisión. Con respecto a los factores biológicos bacterianos, se ha descrito que existen diferencias de transmisión entre los diferentes linajes de MTBC. De hecho, se cree que algunos de estos linajes son ”generalistas” o ”especialistas”, según su distribución geográfica. Si esta distribución desigual se debe a contingencias históricas o factores biológicos relacionados con la bacteria, es aún controversial.

¿Cómo medir la transmisión?

Medir la transmisión de la tuberculosis es complejo debido a la historia natural de la bacteria, ya que no todos los pacientes infectados desarrollan la enfermedad, y solo unos pocos, la transmiten. Además, no existe una herramienta única para evaluar la transmisión, ya que algunas técnicas miden la presencia de infección (como TST e IGRA), mientras que otras se limitan evaluar solo los casos de tuberculosis activa (por ejemplo, genotipado molecular). Globalmente, las intervenciones epidemiológicas para detectar y controlar la transmisión de TB se centran principalmente en la identificación de casos activos junto con un estudio epidemiológico de contactos. Esta estrategia de búsqueda pasiva de casos, supone que una persona con síntomas de TB buscará un centro de salud para recibir un tratamiento. Por otra parte, el objetivo principal del estudio de contactos es reducir el tiempo necesario para detectar y tratar un caso, mediante la identificación de casos secundarios entre pacientes con tuberculosis activa, reduciendo así la transmisión. El estudio de contactos combina la aplicación de encuestas epidemiológicas junto con evidencia clínica (por ejemplo, pruebas TST/IGRA o radiografías torácicas), en contactos cercanos de pacientes con TB activa. Se ha demostrado que los estudios de contactos estiman mejor la prevalencia de tuberculosis activa y latente, y durante muchos años, ha sido la práctica estándar utilizada como método de intervención para el control de la enfermedad en países desarrollados. Sin embargo, su coste-efectividad en los programas nacionales de TB aún se desconoce.

Desde principios de la década de los 90's, se han desarrollado técnicas moleculares para investigar la transmisión de la tuberculosis. Estas herramientas de genotipado han mejorado nuestro entendimiento de la dinámica de transmisión de TB al revelar que algunos casos pertenecen a un mismo grupo de transmisión. Debido a su aplicación rápida y replicable en los sistemas de vigilancia de tuberculosis, las técnicas moleculares combinadas con estrategias epidemiológicas han ayudado a resolver investigaciones de TB, así como a identificar y establecer factores de riesgo asociados a la transmisión de la enfermedad. Sin embargo, se ha reportado que estos métodos moleculares sobreestiman el número de casos que forman parte de algunos grupos de transmisión. Debido a esto, su implementación en los sistemas de salud pública es escasa.

Recientemente, la secuenciación de genomas completos (WGS, por sus siglas en inglés) ha comenzado a emplearse como otra herramienta molecular para detectar la transmisión de TB. En principio, la WGS proporciona una mayor resolución que los métodos moleculares tradicionales para identificar fuentes de infección y delinear las redes de transmisión. Además, la técnica de WGS se está abaratando y ofrece una alternativa rentable para investigar la transmisión de la TB, ya que el porcentaje de concordancia con las investigaciones epidemiológicas es mayor que los descritos con las herramientas anteriores.

Tratamiento de la tuberculosis

El tratamiento de la tuberculosis tiene como objetivo curar a todos los pacientes con enfermedad activa o latente para detener la transmisión o al menos minimizarla. Así, el propósito de la terapia antituberculosa es reducir el número de bacilos activos dentro del paciente; erradicar las poblaciones bacterianas infectantes para prevenir un episodio de recaída y el desarrollo de TB-MDR durante la terapia.

El tratamiento estándar para pacientes diagnosticados con tuberculosis dura al menos 6 meses. La OMS recomienda la administración de rifampicina, isoniazida, pirazinamida y etambutol todos los días durante 2 meses, seguido de 4 meses solo con isoniazida y rifampicina. Este régimen conocido también como tratamiento de primera línea, cuesta alrededor de 20US$ y su éxito clínico es del 85% en todos los casos de TB recientemente diagnosticados. Sin embargo, es una terapia de larga duración y algunos pacientes no la toleran bien, por lo que, algunos casos terminan desarrollando resistencia a uno o más fármacos. En 2017, se estimó que el 4.1% de todos los nuevos casos de

TB a nivel mundial tenían TB resistente a uno o múltiples fármacos, y que aumentaba hasta un 19% en aquellos casos que fueron tratados previamente contra tuberculosis.

El tratamiento para casos MDR es más largo, con más efectos secundarios asociados, y más costosos. Recientemente, la OMS recomendó el uso de un nuevo régimen farmacológico que consiste en la administración de medicamentos de segunda línea durante 6-9 meses. Estas terapias incluyen el uso de bedaquilina, fluoroquinolonas, etionamida y clofazimina, en combinación con fármacos efectivos de primera línea. Cabe resaltar que estos fármacos son totalmente orales. En el caso de pacientes resistentes a las fluoroquinolonas, deben llevar un tratamiento más largo (hasta 20 meses), que incluya una combinación de medicamentos de segunda línea aprobados por la OMS, como por ejemplo, algunos agentes aminoglucósidos inyectables.

Además de casos MDR-TB, existen pacientes que presentan resistencias a fármacos de segunda línea, estos casos son conocidos como tuberculosis extremadamente resistente (XDR-TB, por sus siglas en inglés). Las cepas XDR presentan resistencia a isoniazida, rifampicina, a cualquier fluoroquinolona y al menos a un agente inyectable. En 2018, la OMS informó que el 6.2% de todos los casos MDR-TB fueron XDR-TB. Al igual que el tratamiento de MDR, la OMS ha aprobado recientemente un régimen de medicamentos más corto para tratar estos casos. Esta terapia consiste en el uso de bedaquilina, pretomanida y linezolid durante 6-9 meses. En todos los casos, el éxito del tratamiento depende del grado de resistencia a los medicamentos, la gravedad de la enfermedad y el estado del sistema inmunitario del paciente. Se recomienda la monitorización de la resistencia a los medicamentos y el seguimiento del paciente durante todo el tiempo de tratamiento.

Diagnóstico de cepas resistentes

Poco después del uso de medicamentos contra la tuberculosis en los años 40, surgieron cepas fármaco resistentes. Como consecuencia, en los años 60, se describió un método fenotípico para detectar poblaciones de *M. tuberculosis* resistentes llamado método de proporciones. Desde entonces, este método ha sido la prueba estándar para la detección de resistencias antimicrobianas. El método de proporciones utiliza diferentes concentraciones en serie de los fármacos y compara el crecimiento bacteriano entre cepas sensibles y resistentes. Se puede realizar tanto en medios de crecimiento líquidos como a sólidos. El método de proporciones líquido más común es el sistema automatizado BACTEC MGIT 960 (Becton Dickinson, EE. UU.). Sin embargo, este ensayo introduce errores en las pruebas de sensibilidad y los resultados deben usarse con cuidado. Además de esto, se ha demostrado que algunas mutaciones relacionadas con resistencia a rifampicina (también conocidas como mutaciones en disputa) no son detectadas el sistema. Igualmente, se han reportado que algunas cepas sensibles y resistentes presentan concentraciones mínimas inhibitorias (CMIs) similares para algunos de fármacos de primera y segunda línea. Por esta razón, el Comité Europeo de Pruebas Antimicrobianas (EUCAST) recomendó una revisión de estas distribuciones para algunos fármacos, con el fin de definir los puntos de corte clínicos de cada medicamento.

La secuenciación de cepas MDR y XDR, ha llevado a descubrir genes diana asociados con diferentes resistencias. Así, diferentes pruebas genotípicas (basados en la amplificación ácidos nucleicos) para la identificación de distintas mutaciones puntuales, como los polimorfismos de un solo nucleótido (SNP, por sus siglas en inglés), deleciones e inserciones, se han implementado como pruebas clínicas rutinarias para el diagnósticos de resistencias. Estos ensayos son más rápidos que los métodos fenotípicos, pudiendo detectar una resistencia particular en un par de horas o días. Además, requieren menos experiencia técnica, así como, una menor infraestructura para su realización. Sin embargo, tienden a tener valores de sensibilidad y / o especificidad más bajos que los sistemas basados en cultivos fenotípicos. Entre los kits de detección comerciales disponibles, el ensayo GeneXpert MTB/RIF (Cepheid, EE. UU.), y GenoType MTBDRplus (Hain Lifescience, Alemania) han sido aprobados por la OMS desde 2008. El ensayo Xpert MTB/RIF consiste en una metodología basada en PCR a tiempo real para la detección de ADN de *M. tuberculosis*, así como mutaciones relacionadas con resistencia a rifampicina. Por otro lado, el GenoType MTBDRplus es un ensayo tipo sonda que identifica casos MDR y XDR mediante la detección de mutaciones específicas relacionadas con la resistencia a fármacos de primera y segunda línea (excepto pirazinamida). Recientemente, el ensayo Xpert Ultra (una versión actualizada del Xpert MTB/RIF) ha sido probado y demostrado que mejora la precisión del diagnóstico de TB y puede usarse como una prueba inicial para diagnosticar la tuberculosis pulmonar. A pesar del despliegue masivo de ensayos Xpert (ambas versiones), solo unos pocos países tienen acceso a estas técnicas. En su lugar, el método fenotípico de proporciones sigue siendo la técnica estándar para la detección de resistencias.

A pesar de que los ensayos genotípicos muestran un alto porcentaje de concordancia en comparación con los fenotípicos, todavía hay un porcentaje significativo de cepas resistentes que se clasifican como ”sensibles” por estas pruebas moleculares, especialmente aquellas relacionadas con resistencias a fármacos de segunda línea. Esto se debe a que las sondas genotípicas tienen un número limitado de mutaciones que pueden detectar, ya que solo se interrogan las variantes más comunes que confieren resistencia fenotípica. La aplicación de nuevas tecnologías como WGS podría ayudar a resolver esta limitación; dado que es posible identificar y anotar todas las mutaciones presentes en el genoma, así como identificar nuevas mutaciones relacionadas con la resistencia a fármacos.

Aplicaciones de WGS en tuberculosis

Hoy en día, la WGS se está convirtiendo en una herramienta esencial en el campo de la TB, no solo en las áreas de investigación básica sino también en áreas de diagnóstico y la salud pública. Actualmente, se tiene la capacidad de secuenciar el genoma completo de cientos de cepas de MTBC al mismo tiempo. Las principales aplicaciones de WGS en tuberculosis son: 1) mejorar la predicción de la sensibilidad de los fármacos antituberculosos; 2) detección rápida de grupos de transmisión; 3) vigilancia genómica-epidemiológica de los diferentes genotipos que circulan en una zona específica; y 4) la identificación y clasificación de cepas en linajes pertenecientes al MTBC.

El MTBC es genéticamente monomórfico con muy poca diversidad. Utilizando la tecnología de WGS en una colección global de cepas de MTBC, se describió que la máxima distancia genética entre dos cepas de diferentes linajes son 2.200 SNPs, lo que corresponde al 0.05% del genoma. Los análisis de WGS se basan principalmente en la detección de SNP específicos y/o pequeñas deleciones o inserciones genómicas (llamados INDELS) utilizando “pipelines” bioinformáticos personalizados. Estos “pipelines” consisten en tres pasos principales. A pesar de un protocolo bien definido, no existe un “pipeline” estándar para el análisis de MTBC mediante WGS.

Diversidad de *Mycobacterium tuberculosis* Complex

El MTBC se clasifica en 8 linajes adaptados al ser humano (*M.tuberculosis* y *M.africanum*), así como a algunos animales (*M.bovis*, entre otros). Los ocho linajes filogenéticos de MTBC adaptados al ser humano están diseminados geográficamente, aunque en general, los linajes 2 y 4 son los más distribuidos a nivel mundial. Por el contrario, los linajes 5 y 6 están restringidos a las regiones de África occidental. Esto ha llevado a la hipótesis de que algunos linajes son ”especialistas” con un nicho estrecho para una población humana específica, mientras que los linajes distribuidos globalmente se consideran ”generalistas” que infectan a una gama más amplia de poblaciones. Respecto al origen del MTBC, existe evidencia científica basada en datos de WGS que respalda que el posible origen geográfico del MTBC fue en África hace unos 6000 años.

Además de las características ecológicas, algunos linajes de MTBC muestran diferencias genéticas que tienen un impacto clínico y epidemiológico, lo que resulta en un fenotipo más virulento. Este fenotipo más virulento está relacionado con la severidad de la enfermedad y su tasa de transmisión, haciéndolo exitoso en algunas poblaciones humanas. Sin embargo, el éxito de estos genotipos depende en gran medida, del contexto socioeconómico de cada país, así como la presencia de otros genotipos, y los antecedentes genéticos humanos presentes en cada población.

Predicción de sensibilidad a fármacos

La predicción de la sensibilidad a fármacos mediante la WGS, se basa en la presencia o ausencia de mutaciones específicas relacionadas resistencias a lo largo del genoma de *M.tuberculosis*, por lo que es necesario tener catálogos de mutaciones de alta calidad que ayuden a predecir resistencias con alta confianza. Debido a esto, se han formado diferentes consorcios internacionales que tienen como objetivo expandir la lista actual de mutaciones de alta confianza (ReSeqTB), así como, reemplazar las pruebas fenotípicas (CRyPTIC). Actualmente, podemos predecir perfiles de resistencia a fármacos de primera línea en ausencia de datos fenotípicos, demostrando que se puede reemplazar el cultivo microbiológico.

Desafortunadamente, un pequeño porcentaje de aislados resistentes poseen mutaciones poco comunes que se clasifican como sensibles, dando como resultado un falso negativo, y como consecuencia, comprometiendo el tratamiento contra TB. Una situación similar ocurre con los casos XDR, en donde la mayoría de las variantes relacionadas con resistencia a fármacos de segunda línea aún se desconocen. Esto se debe a que las pruebas fenotípicas disponibles no están bien estandarizadas, lo que genera discrepancias entre diferentes laboratorios. Se necesita el desarrollo de una alternativa o estandarización de los métodos fenotípicos actuales para validar estas mutaciones, así como para aumentar el número de mutaciones de alta confianza en la lista global.

El uso de WGS como marcador epidemiológico

Se ha demostrado que WGS tiene un mayor poder discriminatorio para identificar y desenredar cadenas de transmisión. De hecho, varios estudios han implementado el uso de WGS como marcador de apoyo para ayudar a las investigaciones epidemiológicas. Sin embargo, la mayoría se basan principalmente en estudios retrospectivos, a pesar de que el uso de WGS es extremadamente útil para comprender mejor la dinámica de transmisión en una población, y también para evaluar el impacto de las intervenciones epidemiológicas en el control de la tuberculosis.

Utilizando información de investigaciones epidemiológicas y datos de WGS, se ha propuesto que una distancia genética máxima de 12 SNPs entre dos cepas indica una transmisión reciente, y un umbral de $\leq$ 5 SNP para eventos muy recientes. Así mismo, el tiempo estimado de infección se basa en la baja tasa de mutación del MTBC (0.04-2.2 SNP por genoma, por año). Aunque estos umbrales de SNPs pueden resolver brotes de transmisión, se calibraron en países de baja incidencia, por lo que, aún se desconoce si estas distancias genéticas se pueden aplicar en otros entornos de TB.

Dada su mayor resolución, la WGS se puede usar para inferir enlaces individuales dentro de un grupo de transmisión, en otras palabras, identificar quién infecta a quién. Para ello, se han desarrollado muchos algoritmos basados en modelos matemáticos. Sin embargo, la mayoría de estos se crearon pensando en patógenos con mayor tasa de mutación, y sus aplicaciones en la epidemiología de TB es generalmente limitada a brotes conocidos en lugar de probarla en toda la población.

Otras aplicaciones

Al tener acceso a todo el genoma del MTBC, podemos identificar y extraer SNPs filogenéticos que pueden clasificar cepas de manera rápida. La clasificación de aislados es importante para describir la diversidad bacteriana y comprender su estructura poblacional a escalas local y global. En un contexto epidemiológico, la clasificación de cepas es importante porque permite rastrear brotes específicos que se están extendiendo en entornos locales, así como entre países. Así, se han desarrollado diferentes métodos moleculares basados en PCR, para identificar SNP filogenéticos, como una herramienta alternativa y asequible para el genotipado de cepas del MTBC.

Aunque la técnica de tipificación de SNP no tiene la resolución requerida para definir y resolver grupos de transmisión, se han desarrollado algunas alternativas para identificar brotes de TB en entornos específicos y en tiempo real. Otras aplicaciones de tipificación de SNP incluyen la detección de mutaciones específicas relacionadas con resistencia a fármacos. Debido a sus múltiples aplicaciones, fáciles y rápidas de realizar, estas metodologías continúan desempeñando un papel importante en la investigación y el control de la TB, especialmente en los laboratorios de bajos y medianos recursos, donde la WGS aún está lejos de ser una herramienta esencial.

Implementación de WGS en los sistemas de salud

A pesar del gran avance en el uso de WGS en la investigación de TB, países como el Reino Unido y Holanda, la han implementado como parte de diagnóstico de rutina. Hay una serie de problemas técnicos y económicos con respecto a su implementación como práctica habitual. Estos problemas involucran la infraestructura de laboratorio y computación requerida. Además, es necesario que bioinformáticos especializados desarrollen "pipelines" estandarizados y fáciles de usar. Se espera que en los próximos años, otros países introduzcan WGS en sus sistemas de salud pública.

En países de bajos y medianos ingresos, dicha implementación parece más lejana. Algunas soluciones a corto plazo incluyen herramientas de análisis en línea, que sean rápidas y que los resultados sean fácil de interpretar de forma remota. Existen diferentes herramientas como PhyResSE y TBprofiler que pueden ayudar a cubrir este problema. Ambos “pipelines” realizan un análisis de WGS rápido y completo, lo que incluye con una predicción de resistencias y clasificación filogenética de aislados de MTBC. Sin embargo, la principal limitación a estas alternativas es la necesidad de los archivos de secuenciación de los aislados como entrada, y lo que es más importante, la mala conexión a Internet que tienen algunos lugares. Es necesario apoyo internacional para una implementación sostenible de WGS en el de diagnóstico de la tuberculosis.

Propósito de la tesis

Aunque el uso de WGS en el campo de la TB es cada vez más común, su aplicación como marcador epidemiológico y de diagnóstico aún es escaso. Incluso en regiones económicamente desarrolladas como España, es muy poca su integración en los sistemas de salud pública. La Comunidad Valenciana es un área de baja incidencia de TB en la que, los métodos de diagnóstico actuales y las intervenciones epidemiológicas, son suficientes para mantener esta tasa baja de la enfermedad a lo largo del tiempo, pero con un ritmo lento en la disminución de la incidencia. Gracias a que La Comunidad Valenciana es una región donde la mayoría de los casos de TB son aportados por individuos nacidos en el país, podemos estudiar y explorar diferentes factores asociados de la tuberculosis en las personas locales. Además, los datos sobre la cantidad de transmisión continua de la enfermedad en diferentes entornos son limitados, y los pocos estudios publicados se basan principalmente en marcadores moleculares con baja resolución. Sugerimos que la aplicación de WGS mejorará nuestra comprensión y conocimiento sobre las características clínicas y epidemiológicas de la tuberculosis en la región. Además, creemos que las lecciones y/o métodos aprendidos en esta tesis se pueden extrapolar a otros entornos de TB.

En esta tesis, utilizamos WGS para caracterizar genómicamente una gran colección de aislados clínicos de MTBC recolectados durante tres años (2014-2016) en la Comunidad Valenciana. Primero, realizamos un estudio epidemiológico donde estimamos la tasa de transmisión genómica e identificamos los factores de riesgo asociados (capítulos 3 y 4). También evaluamos el uso de WGS para predecir la resistencia a fármacos en la población estudiada (capítulo 3), y utilizamos WGS para identificar nuevas mutaciones asociadas con resistencia para ayudar a guiar y personalizar el tratamiento de TB en un paciente complicado (capítulo 5). Finalmente, estudiamos la diversidad genómica global de MTBC para proponer una metodología nueva, eficiente y rápida para el genotipado de aislados clínicos (capítulo 6). Cabe resaltar que nuestros resultados se compartieron y compararon con el sistema de salud local. Hasta donde sabemos, este es el primer proyecto regional y probablemente nacional de este tipo. Esperamos que este estudio basado en la población sirva como precursor en el uso de WGS como herramienta rutinaria en el sistema de vigilancia de la salud pública, y también pueda extrapolarse a nivel nacional, así como a países de bajos y medianos ingresos.

Objetivos

Los objetivos principales de la tesis se centran en el uso de WGS aplicada a la vigilancia epidemiológica de la tuberculosis, por lo tanto, los objetivos específicos son:

- Caracterizar por WGS los aislados clínicos de MTBC recogidos durante el período de estudio. Específicamente, estimar y predecir la resistencia a fármacos, así como la transmisión en función de los datos de secuenciación (**capítulo 3**).

- Identificar las características clínicas y epidemiológicas asociadas con la transmisión genómica (**capítulo 3**).

- Comparar la transmisión de tuberculosis detectada por WGS y por las investigaciones epidemiológicas de rutina (estudio de contactos) realizadas por el sistema de salud local (**capítulo 3**).
- Evaluar la dinámica de transmisión dentro de los brotes genómicos detectados, mediante la inferencia de posibles transmisores (incluidos los casos índice), y la identificación de factores de riesgo asociados con ellos (**capítulo 4**).
- Utilizar los datos de WGS para personalizar el tratamiento de la TB, especialmente en aquellos pacientes con perfiles fenotípicos dudosos (**capítulo 5**).
- Desarrollar técnicas rápidas y asequibles basadas en PCR, para caracterizar aislados de MTBC a partir de SNP filogenéticos específicos derivados de datos de WGS (**capítulo 6**).
- Validar y aplicar nuestros ensayos moleculares de tipificación de SNP en dos diferentes entornos de incidencia de TB (**capítulo 6**).

Esquema de la tesis

Esta tesis está compuesta por 4 capítulos principales, 3 de ellos ya han sido publicados en revistas científicas de alto impacto (capítulos 3 a 6).

En el **capítulo 3**, utilizamos los datos de WGS obtenidos de 785 aislamientos clínicos para describir la población con tuberculosis. Más específicamente, evaluamos los factores de riesgo asociados con la transmisión de la enfermedad. También clasificamos y predecimos perfiles de resistencia de todos los casos de TB disponibles. Además, detectamos la transmisión genómica y la comparamos con la detectada por el sistema de vigilancia local. Finalmente, estimamos los valores de sensibilidad y especificidad de WGS utilizando los métodos de diagnóstico de TB de rutina como referencia.

En el **capítulo 4**, combinamos modelos matemáticos con datos de WGS para inferir si el caso índice más probable se muestrea o no, dentro de un grupo de transmisión genómica. Además, estimamos cuándo estos posibles transmisores infectaron a otras personas. En otras palabras, cuando ocurrieron eventos probables de transmisión. Una vez que se identificaron los transmisores, buscamos factores de riesgo específicamente asociados con ellos.

En el **capítulo 5**, utilizamos datos de WGS para identificar e informar un caso MDR-TB mal identificado en un paciente con tuberculosis con una presunta infección ”totalmente sensible”, durante 9 años. Primero, identificamos que unas mutaciones poco comunes asociadas con resistencia eran las responsables del estado MDR. Además, descubrimos que dichas variantes no se detectaron mediante los métodos clínicos de rutina, lo que explica por qué la cepa infectante se identificó como sensible. En este capítulo, destacamos la importancia de WGS para la predicción de resistencias para proporcionar un tratamiento farmacológico adecuado.

Finalmente, en el **capítulo 5**, desarrollamos dos ensayos moleculares rápidos y asequibles de tipificación basados en PCR, para clasificar aislados clínicos de MTBC en los principales linajes filogenéticos de MTBC, así como las sublinajes del linaje 4. Después de la validación, aplicamos nuestras pruebas moleculares en una colección clínica de 491 muestras de MTBC, demostrando valores de alta sensibilidad y especificidad.

Resultados y Discusión

Comparación de la tasa de transmisión genómica en diferentes entornos

En la Comunidad Valenciana, los individuos nacidos en España con tuberculosis son los principales contribuyentes a la incidencia general de la enfermedad, a diferencia de la situación en otras regiones de baja incidencia, donde la mayoría de los casos de tuberculosis son aportados por extranjeros. Al mismo tiempo, se sabe que en estas regiones, la transmisión se asocia a individuos autóctonos. Dada la gran contribución de casos locales en la Comunidad Valenciana, no es sorprendente que también las tasas de transmisión genómica sean más altas en comparación con otras regiones de baja incidencia. Para ilustrar este punto y poner la transmisión de la Comunidad Valenciana en contexto, utilizamos bases de datos comparables de regiones con diferentes tasas de incidencia de tuberculosis, como el Reino Unido y Malawi. El Reino Unido tiene una incidencia de 8 por cada 100,000 habitantes, lo que similar a la Comunidad Valenciana pero contribuido principalmente por extranjeros (72%). Por otro lado, Malawi es un país con tuberculosis endémica con una incidencia de 181 por cada 100,000 personas. Utilizando solo los casos autóctonos de cada país y un umbral de 12 SNP para delinear la transmisión genómica, observamos que en la Comunidad Valenciana, así como en Malawi, casi la mitad de sus respectivos casos locales están involucrados en la transmisión genómica (47.4% en la Comunidad Valenciana contra 49.3% en Malawi), independientemente de la incidencia de TB. En contraste, solo el 32% de los pacientes con TB nacidos en el Reino Unido están en transmisión. Adicionalmente, observamos que varios casos locales de la Comunidad Valenciana tenían entre 15 y 50 SNPs de diferencia entre ellos, mientras que en el Reino Unido, no existen tales diferencias en la gente local. Dado que esta distancia genética refleja eventos de transmisión que tienen décadas de antigüedad, este resultado sugiere que en la en la Comunidad Valenciana, los eventos de contagio más antiguos todavía contribuyen a los casos actuales. Cabe resaltar que en Malawi se muestra un patrón similar al de la Comunidad Valenciana, pero más exacerbado. Este análisis muestra que para llegar a una situación como en el Reino Unido, los esfuerzos para detener la transmisión de TB son clave.

Sugerimos que los altos valores de nuestros casos autóctonos son la consecuencia de una transmisión continua que no se detuvo durante las últimas décadas, y que ahora se refleja en la incidencia local de TB. Por el contrario, en el Reino Unido se observa que los esfuerzos aplicados en el pasado sobre el control de la transmisión han sido exitosos, donde solo la transmisión muy reciente está contribuyendo a la incidencia local de TB. Estos datos sugieren que la situación del Reino Unido es similar en Holanda, Canadá, Alemania y EE. UU., donde la mayoría de los casos se son contribuidos por reactivaciones en extranjeros. Igualmente, estos resultados sugieren que el control de la TB en la Comunidad Valenciana y probablemente en España, se está rezagado con respecto a otros países de baja incidencia. Esto no solo se refleja en los diferentes patrones de transmisión, sino también, en que la incidencia de TB en la población autóctona (6.7 por 100,000 personas) sigue siendo mucho más alta que en la población local del Reino Unido (3.5 por 100,000 personas), por ejemplo. Dado que el control real de la TB en la Comunidad Valenciana está cumpliendo con los objetivos propuestos por la OMS para reducir la incidencia de TB, es posible que los esfuerzos para controlar la transmisión sean eficientes, pero se necesita más tiempo para alcanzar los resultados observados en otros países de baja incidencia, como el Reino Unido.

Uso de WGS para identificar transmisión reciente

El término de transmisión reciente se usa para definir los eventos de contagio que ocurrieron en un corto período de tiempo, típicamente revelados por estudios de contacto. Generalmente se define como transmisión reciente aquellos casos de contacto que desarrollen la enfermedad en un periodo de 2 a 5 años después del caso índice. Cuando se utiliza WGS para definir transmisión, esto se traduce a un diferencia media genética de 0-5 SNPs entre dos casos, ya que se supone que la bacteria acumula 0.3-0-5 SNPs/año. Aunque utilizar umbrales de hasta 12 SNPs es bien aceptados para definir grupos de transmisión, el porcentaje de concordancia con las investigaciones epidemiológicas disminuye y, por lo tanto, muchos eventos de transmisión genómica no se puede validar fácilmente. Por lo tanto, diseñar medidas de control efectivas no solo es relevante para medir la transmisión reciente sino también la contribución de eventos de transmisión más antiguos. Mientras que en el Reino Unido esos eventos de transmisión más antiguos no existen, en la Comunidad Valenciana, así como en Malawi, el análisis de distancias genéticas muestra un patrón continuo de distancias que no se ajustan a un umbral estricto de 12 SNPs, y reflejan unos constantes flujos de eventos de transmisión, tanto recientes como antiguos. Con esto sugerimos que, en la Comunidad Valenciana, el uso de umbrales de SNPs para definir transmisión es útil para revelar eventos que han sucedido muy recientemente, pero se pierde la imagen completa de cómo la transmisión está contribuyendo a la incidencia anual de TB. Además, es probable que esta situación sea más común entre otras regiones epidemiológicas, como se muestra en Malawi.

Como se esperaba en la Comunidad Valenciana, usando un umbral de 5 SNPs, equivalente a 5 años, los casos epidemiológicamente relacionados identificados por el sistema de vigilancia de salud pública son detectados como eventos de transmisión muy recientes por WGS. Esto está relacionado con el hecho de que, las investigaciones epidemiológicas buscan contactos recientes y no transmisiones más antiguas. Sin embargo, incluso en ese período de 5 años, las investigaciones locales no detectaron el 60% de los casos de transmisión. Esto sugiere que, si bien las investigaciones locales son muy buenas para rastrear algunos contactos cercanos (miembros de la familia, lugar de trabajo), se pierden muchos casos de contacto que probablemente ocurren fuera de los límites de los cuestionarios. En ese sentido, algunas investigaciones han incorporado información sobre contactos sociales que han ayudado a identificar casos de transmisión adicionales en algunas regiones específicas. La aplicación de estas estrategias podría ayudar a mejorar la identificación transmisión en la Comunidad Valenciana.

Nuestros datos también sugieren que la transmisión puede ocurre entre contactos casuales o incluso antes de desarrollar síntomas (como se muestra en el capítulo 2). La identificación de esos individuos es más complicada. En esos casos, la implementación de algún tipo de búsqueda activa de casos puede ayudar a identificar estos transmisores. Por ejemplo, mediante la implementación de pruebas selectivas basadas en la comunidad en grupos de riesgo o pruebas de detección a gran escala en puntos críticos de transmisión, según lo informado por WGS.

Transmisión de la tuberculosis durante estados subclínicos

El control de la transmisión es clave para disminuir la incidencia de la tuberculosis. Con respecto a esto, es obligatorio comprender la compleja y dinámica transmisión de la TB y los factores de riesgo asociados a esta. Usando WGS combinado con datos epidemiológicos, investigamos la dinámica de la transmisión dentro de una fracción de brotes genómicos utilizando un método filogenético bayesiano llamado TransPhylo (**capítulo 4**). Los resultados fueron estimaciones que incluyen la probabilidad de que en un brote de transmisión el caso índice sea muestreado, cuál de todos los casos era el caso índice y cuándo ocurrió un evento de transmisión. Sorpresivamente, descubrimos que en algunos individuos, la transmisión ocurre antes de desarrollar los síntomas, probablemente durante la enfermedad subclínica. Este resultado respalda la idea de la existencia de diferentes estados de infección antes de desarrollar TB activa más allá de la clásica dicotomía de la enfermedad latente/activa. Igualmente, mostramos por primera vez que las cepas de MTBC son transmisibles durante algunos de estos estados de infección recientemente reconocidos. Sin embargo, se necesitan más estudios basados en WGS para evaluar la cantidad de transmisión subclínica en diferentes entornos clínicos.

Estos hallazgos proporcionan nuevas ideas de cómo se transmite la enfermedad, y sobre la relación patógeno-huésped durante la infección que finalmente conduce a la transmisión. Sugerimos que algunos métodos basados en tecnologías ómicas, pueden vincular el tiempo de transmisión con las el nivel de infección del huésped (por ejemplo, utilizando transcriptomas del huésped) tienen el potencial de aumentar la identificación de casos de TB en riesgo de transmitir la enfermedad durante las diferentes etapas de la infección de TB.

Búsqueda activa de casos para detener la transmisión

Como se ha demostrado en esta tesis, un gran porcentaje de transmisión (la mayor parte relacionada con pacientes autóctonos) es ignorada por los sistemas de vigilancia locales. Por lo que es necesario el desarrollo de nuevas herramientas de diagnóstico y nuevas intervenciones epidemiológicas para detener esta transmisión, y la disminuir la incidencia de tuberculosis. Probablemente, la estrategia con efectos a corto plazo más importante es la de cambiar la búsqueda de casos, de una manera pasiva a una activa. Esta búsqueda activa tiene como objetivo encontrar casos de tuberculosis antes de que el paciente busque atención médica. De esta manera, las intervenciones pensadas a nivel de comunidad incluyen el monitoreo de personas en áreas concurridas en busca de tuberculosis. Se han demostrado que esta búsqueda activa aumenta la detección de TB en individuos asintomáticos con esputos negativos y, sin aparente carga bacilar, tanto en regiones de baja y alta incidencia. A pesar de las mejoras epidemiológicas que brinda esta intervención, se requieren estudios sobre su rendimiento y rentabilidad en la salud pública, así como herramientas de detección rápida de tuberculosis. El desarrollo de pruebas moleculares capaces de detectar niveles muy bajos de ADN de *M. tuberculosis* podría identificar más casos de TB que la microscopía de esputo. Dada la baja prevalencia de la enfermedad de TB en la población general, necesitamos encontrar estrategias rentables para encontrar casos de manera activa. En ese sentido, la transmisión revelada por WGS puede ayudar a diseñar y guiar las intervenciones epidemiológicas.

Predicción de resistencias

Aunque la Comunidad Valenciana no es considerada una región con alta incidencia de MDR-TB, el uso de herramientas precisas y rápidas capaces de predecir un perfil de resistencias es esencial para el manejo adecuado del paciente. Estudios recientes han demostrado que la prevalencia de cepas resistentes puede predecirse mediante la vigilancia genómica, especialmente en fármacos de primera línea. Los altos valores de especificidad y sensibilidad demuestran que la WGS es una herramienta de predicción confiable para detectar resistencias de primera línea, al menos en la Comunidad Valenciana (**capítulo 3**). Los pocos resultados discrepantes que identificamos podrían deberse a problemas fenotípicos relacionados con las técnicas usadas para medir la sensibilidad, o por la presencia de mutaciones desconocidas. En este sentido, la predicción por WGS depende de la existencia de catálogos de mutaciones de alta confianza. Aunque existen herramientas *online* que contienen varios catálogos de mutaciones asociados con resistencia, aún existen variantes desconocidas para fármacos de primera y particularmente de segunda línea, así como mutaciones de resistencia a los nuevos antibióticos. A pesar de estas limitaciones, existe suficiente confianza para predecir la sensibilidad a fármacos de primera línea, como lo demostró un estudio que incluyó 10,000 aislados clínicos de diferentes partes del mundo. Esto ha llevado a algunos países a eliminar las pruebas fenotípicas y reemplazarlo por la predicción de WGS. Sin embargo, los ensayos fenotípicos son necesarios en aquellos casos donde la inspección por WGS no es concluyente.

Tratamiento personalizado de tuberculosis basado en WGS

En esta tesis utilizamos la WGS para descubrir y describir nuevas mutaciones relacionadas con la resistencia a la isoniazida (**capítulo 5**) en un paciente aparentemente “sensible”. Estas variantes no fueron detectadas por las pruebas fenotípicas automatizados, y así, los aislados que tenían estas mutaciones se consideraron sensibles. El análisis fenotípico y genómico de estos aislados reveló que estas mutaciones confieren resistencia fenotípica de ”bajo nivel”, lo que resulta en un resultado difícil de interpretar. Sabiendo esto, utilizamos la predicción de WGS de manera prospectiva para guiar el tratamiento del paciente. Aunque el caso adquirió XDR-TB, un seguimiento cercano del paciente utilizando la información genómica, clínica y microbiológica llevó a los médicos a tratar al paciente con éxito. Hasta donde sabemos, este es uno de los primeros casos que utilizan WGS para personalizar el tratamiento de un individuo utilizando información genómica. En el contexto de este paciente, esto fue especialmente relevante ya que las pruebas fenotípicas para algunos medicamentos de primera y segunda línea no fueron concluyentes. Estos resultados destacan la capacidad de la WGS para pronosticar de forma rápida y precisa la resistencia a los medicamentos, al menos en nuestra región de estudio. Además, también es notable que la WGS pueda resolver casos fenotípicamente indeterminados, en este sentido, los tratamientos personalizados son imprescindibles para salvar la vida de los pacientes. Esto es realmente relevante para los países con alta incidencia de MDR-TB, donde se encuentran en situaciones de manejo más complicadas y donde muchas veces, el tratamiento de segunda línea es empírico y no está bien estandarizado. Por esta razón, es importante expandir los catálogos de mutaciones para estos fármacos, y más importante aún, para aquellos incluidos en los nuevos regímenes recomendados por la OMS. Nuestro conocimiento sobre las bases genéticas de la resistencia a esas drogas está lejos de ser completo. Combinado con el hecho de que no hay pruebas comerciales fenotípicas disponibles, el monitoreo y seguimiento de pacientes para detectar la adquisición de resistencia será extremadamente desafiante.

Clasificación de aislados

Como se indicó a lo largo de esta tesis, la WGS es una herramienta valiosa para identificar SNPs que podrían utilizarse para diferentes aplicaciones en estudios de población microbiana. Además, la WGS puede usarse como una herramienta de genotipado para clasificar cepas de MTBC en diferentes linajes. Aunque esta tecnología se ha vuelto más barata en los últimos años, se necesita una infraestructura específica y personal calificado para poner en marcha este método, estas condiciones no siempre están disponibles en las regiones de bajos y medianos ingresos. Sin embargo, el conocimiento obtenido de WGS podría traducirse para desarrollar métodos de laboratorio más simples con el fin de obtener información específica y precisa de una manera más, fácil, barata y rápida. En el capítulo 6, se desarrollaron dos herramientas moleculares para clasificar aislados de MTBC en los principales linajes, utilizando reactivos basados en PCR e instrumentos básicos de laboratorio. Uno de los objetivos era tener un método de genotipado alternativo y asequible que pudiera ser útil en países con recursos económicamente limitados. Adicionalmente, demostramos que nuestros ensayos moleculares funcionan incluso en muestras biológicas con baja concentración de ADN, lo que es un problema común en muestras clínicas de rutina (por ejemplo, ADN obtenido de muestras inactivadas por calor y de esputo). Una vez puesto a punto, se genotipó una colección clínica de alrededor de 500 aislados de países con diferente incidencia de TB. Gracias a la baja diversidad de MTBC y sus características genómicas clonales, esta metodología no se limitan solo a la detección de SNP filogenéticos, de hecho, se pueden adaptar a otras aplicaciones específicas. Por ejemplo, los paneles para la detección de SNPs relacionados con resistencias se usan comúnmente en diagnósticos clínicos de rutina. Igualmente, marcadores con interés epidemiológico (para identificar brotes locales o diseminación transcontinental de TB) se pueden desenmascarar con estas técnicas moleculares. Sin embargo, siempre se requiere una noción previa de lo que se quiere amplificar para proceder a hacer WGS de casos representativos e identificar posiciones específicas de SNP.

Consideraciones adicionales

A partir del análisis y los resultados reportados aquí, podemos afirmar que la implementación de WGS en los sistemas de vigilancia de salud pública mejorará la predicción resistencias y el manejo individual de los casos. Además, ayudará a caracterizar con precisión los patrones de transmisión de TB a nivel poblacional. Sin embargo, su implementación requiere la integración de otras disciplinas relacionadas, como la microbiología, la epidemiología y la bioinformática. Por lo tanto, se necesita un equipo multidisciplinario que involucre personal con conocimientos sobre genómica y habilidades para analizar grandes bases de datos. Actualmente, solo algunos países desarrollados están utilizando WGS como herramienta de diagnóstico rutinario, pero se estima que más países lo adoptarán en los próximos años. En la Comunidad Valenciana, su integración al sistema de vigilancia local no parece muy lejana. De hecho, muchos profesionales clínicos y epidemiólogos son conscientes del uso y las aplicaciones de WGS en el campo de la tuberculosis. Un total de 18 unidades de microbiología de la región han contribuido generosamente al estudio en lo que representa, uno de los mejores ejemplos de estudios multicéntricos en enfermedades infecciosas en el país. Gracias a este esfuerzo coordinado, describimos y probamos cómo esta tecnología complementa los enfoques actuales utilizados en la Comunidad Valenciana para manejar la enfermedad. La implementación de WGS como herramienta complementaria para el diagnóstico y la epidemiología de la TB en la Comunidad Valenciana ya están ayudando al control local de la TB y, creemos que puede servir como plantilla para usar WGS en otras enfermedades infecciosas en el sistema local de vigilancia de salud pública.

Conclusiones

- La incidencia de transmisión de tuberculosis es alta según WGS en comparación con otros países como el Reino Unido. Los factores de riesgo asociados con brotes de transmisión en la Comunidad Valenciana son casos autóctonos (asociados), jóvenes (asociados) y mayores (no asociados).

- La mayoría de la incidencia de TB y su transmisión se asocia con personas nacidas en el lugar.

- Aunque las investigaciones epidemiológicas actuales de TB (rastreo de contactos) mantienen una baja incidencia de TB, subestiman la carga real de transmisión de TB.

- Los análisis de sensibilidad y especificidad demostraron que WGS es una herramienta precisa y confiable para detectar eventos de transmisión de TB recientes y anteriores y predecir la resistencia a los medicamentos. Podría convertirse en una metodología transformadora para el sistema de vigilancia de salud pública.

- El modelo filogenético basado en WGS combinado con datos epidemiológicos permite inferir transmisores de alta probabilidad e índices de casos dentro de un grupo de transmisión. Sin embargo, su aplicación está limitada por la diversidad nula observada muchas veces entre casos agrupados.

- En algunos individuos, la transmisión de TB puede ocurrir durante la enfermedad subclínica y puede poner en peligro el progreso de la búsqueda pasiva de casos y el estudio de contactos. Junto con los datos emergentes de diferentes campos que controlan la enfermedad subclínica a través de enfoques activos de búsqueda de casos, serán relevantes en el futuro cercano.

- El uso de WGS en tiempo real se puede utilizar para detectar variantes de resistencia a los medicamentos durante el tratamiento de la tuberculosis. Más importante aún, ayuda a identificar mutaciones poco comunes que surgen a través del tiempo.

- Los enfoques moleculares de tipificación SNP basados en PCR brindan una herramienta rápida alternativa para clasificar las cepas de MTBC y es un método que podría traducirse en múltiples aplicaciones de TB.