Culture-free sequencing of *Mycobacterium tuberculosis* for diagnosis and genetic diversity identification

phD Thesis

Doctoral programme in Biotechnology

Carla Mariner Llicer

Supervisors:

Iñaki Comas Espadas
Mariana Gabriela López Rodríguez

UPV tutor:

Jorge García Hernández

October, 2024

INSTITUTO DE BIOMEDICINA DE VALENCIA CSIC

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Index

Abstract

Resum

La detecció i el diagnòstic de resistències en *Mycobacterium tuberculosis* (MTB) suposa un repte per al control de la tuberculosi (TB). La tècnica de referència que sempre s'ha utilitzat ha consistit a cultivar les mostres diagnòstiques. No obstant això, el cultiu depèn del creixement lent de MTB, retardant l'obtenció de resultats des de setmanes fins a mesos, i conseqüentement, la prescripció d'un tractament òptim per als pacients. A més, el cultiu sempre ha sigut el primer pas per a enriquir mostres diagnòstiques per a fins d'investigació, com per exemple per a seqüenciar el genoma de MTB. A pesar d'això, encara es desconeix si cultivar imposa un coll de botella de la diversitat genètica original de la mostra diagnòstica. Si aquesta hipòtesi fora certa implicaria un biaix en l'obtenció de resultats. Per ara, evitar el cultiu per al diagnòstic ha sigut possible amb la introducció de tècniques moleculars als sistemes de vigilància. Les tècniques moleculars, a banda de reduir el temps d'obtenció de resultats, permeten detectar el patogen i identificar resistència a antibiòtics. No obstant això, aquestes proves estan limitades a detectar les mutacions més comunes conegudes que confereixen resistència a la rifampicina i/o isoniazida, els fàrmacs de primera línia més usats al tractament de la TB. D'altra banda, respecte a la investigació genòmica, la composició complexa de les mostres diagnòstiques, com l'esput, encara suposa un repte per a la seqüenciació.

En aquest context, en aquesta tesi presentem diferents tècniques i estratègies de seqüenciació per a recuperar i seqüenciar MTB a partir de mostres d'esput, centrant-nos principalment a estudiar la diversitat present en esputs i cultius i explorar el potencial de les tècniques de seqüenciació per a determinar el perfil de resistències a antibiòtics.

Al Capítol 1 comparem la diversitat genètica entre parelles d'esputs i cultius en dos entorns diferents. Per dur-ho a terme, desenvolupem i implementem una tècnica de seqüenciació a partir de mostra directa d'esput des de l'extracció de l'ADN fins a l'anàlisi bioinformàtic de les seqüències. Depenent de la quantitat inicial de MTB, realitzem una seqüenciació directa o enriquida dels esputs. A més, desenvolupem un flux d'anàlisi per als esputs per a detectar i descartar tota la diversitat procedent d'artefactes del processament de les mostres. Per a generalitzar els resultats obtinguts, es descarreguen i es reanalitzen tres conjunts de dades prèviament publicades. Els resultats obtinguts mostren una concordança alta en la diversitat dels esputs i els seus respectius cultius, tant en termes de presència/absència com de freqüència de les variants. Aquests resultats, que contradiuen la hipòtesi inicial del coll de botella del cultiu, destaquen la importància d'incloure passos de filtrat depenent de la complexitat de les mostres analitzades.

Al Capítol 2 explorem les limitacions del Gene Xpert MTB/RIF Ultra (Xpert Ultra) en la identificació de mutacions que confereixen resistència a rifampicina i examinem les mutacions de resistència a antibiòtics més prevalents que circulen al sud de Moçambic. Per fer-ho, utilitzem la seqüenciació del genoma complet com a tècnica per a identificar totes les variants associades a resistència a antibiòtics que estan presents en les soques recol·lectades al districte de Manhiça (Moçambic) durant dos períodes de temps. Al nostre

conjunt de dades, identifiquem dues soques que alberguen mutacions de resistència a rifampicina que estan fora de l'abast del Xpert Ultra. A més, també detectem una alta prevalença de soques resistents a isoniazida que no presenten mutacions associades a la resistència a rifampicina (soques no resistents a múltiples fàrmacs, MDR-TB). Aquests resultats impliquen que l'Xpert Ultra, que és l'eina de diagnòstic de primera línia utilitzada al sud de Moçambic, estaria identificant com a sensibles alguns casos resistents a rifampicina o isoniazida. Per tant, aquest capítol destaca la importància d'incloure noves mutacions i nous gens de resistència a antibiòtics en les proves de diagnòstic moleculars.

Al Capítol 3 avaluem el potencial de la seqüenciació d'amplicons per a predir resistència a antibiòtics i identificar el llinatge de les soques. Per a això dissenyem un panell que consta de nou gens associats a resistència a set antibiòtics i dues regions que contenen marcadors de llinatge. A més, desenvolupem una PCR multiplex per a amplificar totes les regions en una sola reacció. A diferència d'altres panells de gens comercials, els nostres encebadors dissenyats cobreixen els gens complets per poder detectar totes les mutacions, no només les que es troben als punts calents. Per tant, per a seqüenciar usem la tecnologia de lectura llarga amb MinION (Nanopore). Es tracta d'una plataforma de seqüenciació portàtil que ofereix una anàlisi en temps real. Cal destacar que un dels objectius principals d'aquest capítol és calibrar els paràmetres per a cridar variants en seqüències de MinION. Per a això, comparem les variants cridades en Illumina amb les cridades en MinION i d'aquesta manera establir la freqüència mínima per a evitar cridar variants falses positives que es deguen a la taxa d'error de MinION. Els resultats obtinguts demostren una alta correlació entre la detecció de resistència i

identificació de llinatges entre MinION i Illumina, independentment del tipus de mostra (esput o cultiu). Aquests descobriments juntament amb l'anàlisi portàtil i en temps real de MinION representen un gran avantatge per a una futura implementació en sistemes d'atenció primària.

En conclusió, aquesta tesi contribueix als avanços al camp de la genòmica directa de mostra diagnòstica. Per una banda demostrant que el cultiu reflecteix la diversitat genètica present en l'esput, i per l'altra, revelant el potencial de la seqüenciació, tant del genoma complet com d'amplicons, per a la predicció d'antibiòtics, superant les limitacions que presenten les proves moleculars. Les tècniques i resultats que es presenten contribuiran a promoure la seqüenciació directa d'esput com a tècnica de diagnòstic ràpida i precisa, però també al seu ús per a la investigació.

Resumen

La detección y el diagnóstico de resistencias en *Mycobacterium tuberculosis* (MTB) supone un reto para el control de la tuberculosis (TB). La técnica de referencia que siempre se ha usado implica cultivar las muestras diagnósticas. Sin embargo, debido al lento crecimiento de las micobacterias, el cultivo demora la obtención de resultados desde semanas hasta meses, lo que conlleva a retrasar la prescripción de un tratamiento óptimo para los pacientes. Además, el cultivo siempre ha sido el primer paso para enriquecer muestras diagnósticas para fines de investigación, como por ejemplo para secuenciar el genoma de MTB. A pesar de ello, por el momento se desconoce si cultivar impone un cuello de botella de la diversidad genética original de la muestra diagnóstica. Si esto fuera cierto, implicaría un sesgo en la obtención de resultados. Hasta el momento, omitir el cultivo para el diagnóstico ha sido posible con la introducción de técnicas moleculares en los sistemas de vigilancia. Estas técnicas, además de reducir el tiempo de obtención de resultados a unas pocas horas, son capaces de detectar el patógeno y además identificar mutaciones de resistencia a algunos antibióticos. No obstante, estas pruebas están limitadas a detectar las mutaciones más comunes que confieren resistencia a los fármacos usados en el tratamiento de TB, principalmente isoniazida y/o rifampicina. Por otro lado, respecto a la investigación genómica, la composición compleja de las muestras diagnósticas, como el esputo, aún supone un reto para la secuenciación.

En este contexto, el objetivo general de esta tesis es evaluar diferentes técnicas y estrategias de secuenciación para recuperar y secuenciar MTB a partir de muestras clínicas, centrándonos principalmente en estudiar y

comparar la diversidad presente en esputos y cultivos, y explorar el potencial de las técnicas de secuenciación para determinar el perfil de resistencias a antibióticos a partir del esputo .

En el Capítulo 1 comparamos la diversidad genética entre parejas de esputos y cultivos en dos entornos de TB diferentes. Para ello desarrollamos e implementamos una técnica de secuenciación a partir de muestra directa de esputo, desde la extracción del ADN hasta el análisis bioinformático de las secuencias. Dependiendo de la cantidad inicial de MTB, realizamos una secuenciación directa o enriquecida de los esputos. Además, desarrollamos un flujo de análisis que permite detectar y descartar los artefactos procedentes del procesado de muestras clínicas complejas. Para generalizar los resultados obtenidos, se reanalizan tres conjuntos de datos previamente publicados. Los resultados obtenidos muestran una concordancia alta en la diversidad de los esputos y sus respectivos cultivos, tanto en término de presencia/ausencia como de frecuencia de las variantes. Estos resultados, que contradicen la hipótesis inicial del cuello de botella del cultivo, destacan la importancia de aplicar pasos de filtrado adecuados a la complejidad de las muestras analizadas.

En el Capítulo 2 exploramos las limitaciones del Gene Xpert MTB/RIF Ultra (Xpert Ultra) en la identificación de mutaciones que confieren resistencia a rifampicina y examinamos las mutaciones de resistencia a antibióticos más prevalentes que circulan en el sur de Mozambique. Para ello, utilizamos la secuenciación del genoma completo como técnica para identificar todas las variantes asociadas a resistencia a antibióticos que están presentes en las cepas recolectadas en el distrito de Manhiça (Mozambique) durante dos períodos de tiempo. En nuestro conjunto de datos, identificamos mutaciones

de resistencia a rifampicina que están fuera del alcance del Xpert Ultra y detectamos, también, una alta prevalencia de cepas resistentes a isoniazida que no presentan mutaciones asociadas a la resistencia a rifampicina (cepas no resistentes a múltiples fármacos, MDR-TB). Estos resultados implican que el Xpert Ultra, que es la herramienta de diagnóstico de primera línea utilizada en el sur de Mozambique, estaría identificando como sensibles algunos casos resistentes a rifampicina o isoniazida. Por tanto, este capítulo destaca la importancia de incluir nuevas mutaciones y nuevos genes de resistencia a antibióticos en los tests de diagnóstico molecular.

En el Capítulo 3 evaluamos el potencial de la secuenciación de amplicones, a partir de cultivo y esputo, para predecir resistencia a antibióticos e identificar el linaje de las cepas. Para ello diseñamos un panel que consta de nueve genes asociados a resistencia a siete antibióticos y dos regiones que contienen marcadores de linaje. Además, desarrollamos una PCR multiplex para amplificar todas las regiones en una sola reacción. A diferencia de otros paneles de genes comerciales, los cebadores que hemos diseñado permiten amplificar el gen completo y, por tanto, identificar todas las mutaciones presentes, no sólo las que se encuentran en los puntos calientes. Para secuenciar usamos la tecnología de lectura larga con MinION (Nanopore); se trata de una plataforma de secuenciación portátil que permite un análisis a tiempo real. Cabe destacar que uno de los objetivos principales de este capítulo es calibrar los parámetros para llamar variantes en secuencias de MinION. Para ello, comparamos las variantes llamadas en Illumina con las llamadas en MinION y de esta forma establecer el umbral de frecuencia mínima para evitar llamar variantes falsas positivas que se deban a la tasa de error de MinION. Los resultados obtenidos demuestran una alta correlación

entre la detección de resistencia e identificación de linajes entre MinION e Illumina, independientemente del tipo de muestra (esputo o cultivo). Estos hallazgos junto con el análisis portátil y en tiempo real de MinION representan una gran ventaja para una futura implementación en sistemas de atención primaria.

En conclusión, esta tesis contribuye a los avances en el campo de la genómica directa de muestra diagnóstica. Por un lado, demostrando que el cultivo refleja la diversidad genética presente en el esputo y, por otro, revelando el potencial de la secuenciación, tanto del genoma completo como de amplicones, para la predicción de resistencias a antibióticos, superando las limitaciones que presentan los tests moleculares. Las técnicas y resultados que se presentan contribuirán a promover la secuenciación directa de esputo como técnica de diagnóstico rápida y precisa, pero también a promover su uso para la investigación.

Summary

Mycobacterium tuberculosis (MTB) detection and diagnosis of drug resistance (DR) have been a challenge for tuberculosis (TB) control. Growing mycobacteria from diagnostic samples has, historically, been the most commonly used diagnostic technique. However, culturing depends on the slow growth rate of MTB which delays the obtention of results and, therefore, the prescription of optimal treatment from weeks to months. In addition, culture has always been the first step to enrich diagnostic samples also for research purposes such as MTB genomics. Although, there is still no evidence whether it represents a bottleneck of the original genetic diversity present in the diagnostic sample, which if true would represent a bias in the obtention of results. Skipping the culturing step and using the clinical sample has been possible for diagnosis since the development and introduction of molecular techniques into surveillance systems. Apart from reducing the turnaround time by a few hours, they allow a complete diagnosis in terms of both MTB and DR detection by performing only one assay. Nevertheless, these tests are limited to detect the most common known mutations that confer resistance to rifampicin and/or isoniazid, the first-line drugs used for TB treatment. However, regarding genomic research purposes, the complex composition of diagnostic samples, such as sputum, challenges the sequencing workflow.

In this thesis we use different sequencing approaches to recover and sequence MTB from sputum samples focusing mainly on overcoming the hypothesised bottleneck that culture poses and on exploring the potential of sequencing for MTB drug-resistance prediction.

In Chapter 1 we compare the genetic diversity within sputum-culture paired samples in two different settings. For this purpose, we implement a culture-free sequencing approach for sputum samples from MTB DNA extraction to downstream sequence analysis to deplete all the non-MTB DNA. Depending on the initial amount of MTB, we performed direct or enriched sequencing of sputum samples. We also develop a tailored analysis pipeline to detect and filter out all the artifactual variation present in culture-free approaches. To generalise the results obtained across settings, three additional published datasets are re-analysed. Overall, our results show a high concordance of genetic diversity between sputum-culture pairs, both in terms of the presence and frequency of the variants. These findings, which contradict the initial assumption of culture bottleneck, highlight the importance of applying important filtering steps depending on the complexity of the samples.

In Chapter 2 we explore the limitations of the Gene Xpert MTB/RIF Ultra (Xpert Ultra) in the identification of non-common mutations and screen for the most prevalent drug-resistance mutations circulating in southern Mozambique. To do so, we use whole genome sequencing to identify all the drug-resistance associated variants present in strains collected in the Manhiça district (Mozambique) during two time periods. In our dataset, we identify two strains harbouring rifampicin resistance mutations that are outside the scope of the Xpert Ultra. In addition, we also detect a high prevalence of isoniazid-resistant strains that do not present rifampicin resistance associated mutations (non multidrug-resistant strains, MDR-TB). This means that Xpert Ultra, that is the front line diagnostic tool used in southern Mozambique, sometimes misdiagnoses rifampicin and isoniazid resistant cases and, therefore, highlights

the importance of including new drugs and mutations in molecular diagnostic tests.

In Chapter 3 we evaluate the potential of targeted sequencing for drug resistance prediction and lineage identification by designing a gene panel consisting of nine genes associated with resistance to seven drugs, and two phylogenetic determining regions. For this end, we develop a multiplex PCR to amplify all the target regions in a single-tube reaction. Unlike other commercial gene panels, our primers designed cover the entire genes to identify all the mutations, not only the hot-spot regions. In that regard, we use the portable MinION (Nanopore) long-read sequencing platform that offers a real-time analysis. Importantly, one of our main aims is to focus on calibrating the parameters for MinION variant calling. Thus, we compare the variant calling of MinION and Illumina to establish the minimum frequency cut-off to avoid false positive variant calls due to the error rate of MinION. Our results demonstrate a high agreement between MinION and Illumina DR detection and lineage classification regardless of the sample type (sputum and culture). These findings together with the portable and real-time analysis of MinION represent a great advantage for a future implementation in point-of-care systems.

To sum up, this thesis contributes to advances in the field of culture-free genomics by validating that culture mirrors the genetic diversity present in sputum samples and by demonstrating the potential of whole genome and target sequencing to overcome the limitations of molecular diagnostic tests in the prediction of drug resistance. The techniques and findings presented will contribute to promote culture-free sequencing as a front-line approach for faster and more accurate diagnosis, and also for research purposes.

Introduction

Tuberculosis disease and control

TB Facts

Tuberculosis (TB) is an ancient infectious disease caused by *Mycobacterium tuberculosis* (MTB). It usually affects the lungs (pulmonary TB) and is transmitted through the air, but can also manifest in other parts of the body (extrapulmonary TB) [1]. Despite being preventable and curable, TB is still one of the top ten leading causes of death due to an infection worldwide. In 2022, the World Health Organisation (WHO) estimated 10.6 million new cases and 1.3 million deaths [2] (**Figure 1**).

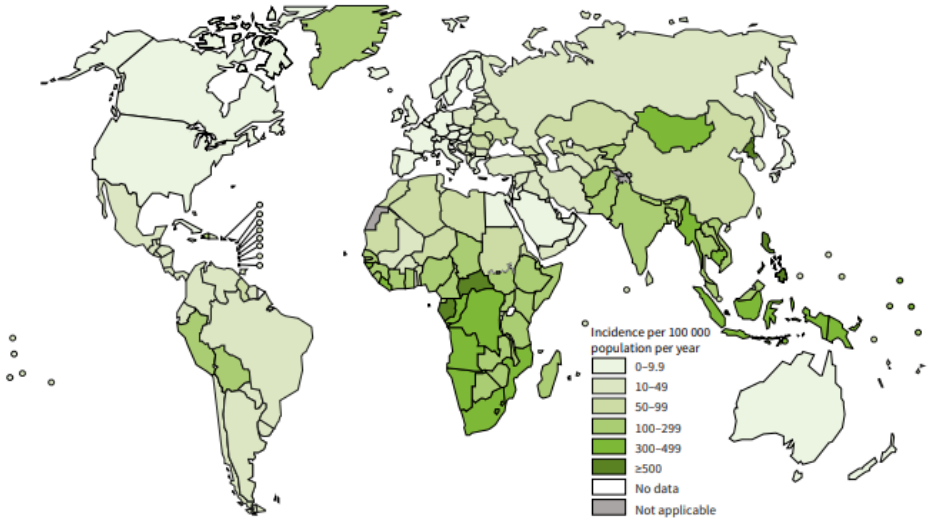


Figure 1: Estimated incidence rate of Tuberculosis in 2022. Figure extracted from WHO report 2023 [2].

Remarkably, more than 90% of noticed cases and deaths occur in low and middle-income countries, highlighting the substantial burden of this disease on vulnerable populations with limited access to public health resources [3] (**Figure 2**).

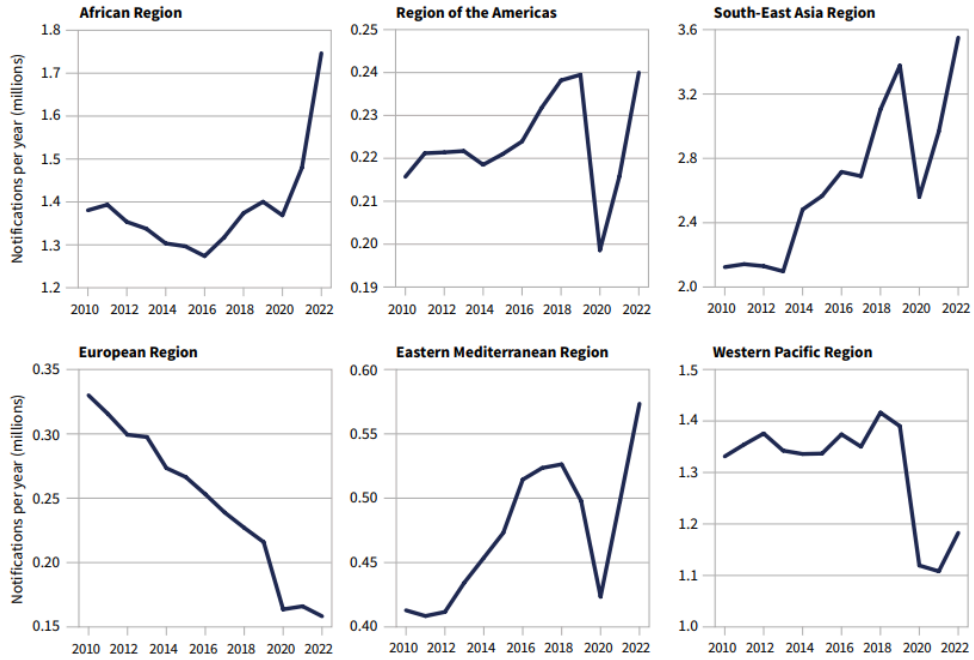
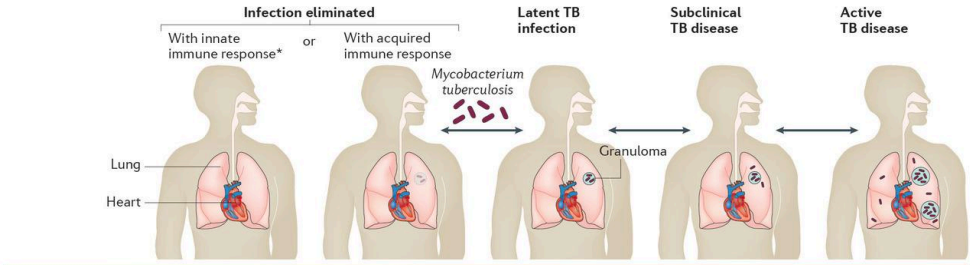


Figure 2: Trends in case notifications of people newly diagnosed with TB by WHO region (2010-2022). Figure from the WHO Tuberculosis report from 2023 [2].

Pathogenesis

From a clinical perspective, pathogenesis of (pulmonary) TB has been classified into a latent asymptomatic state characterised by a reduced metabolic activity of the bacteria; and an active state in which the patient is symptomatic and infectious. Adult patients with active TB can manifest several symptoms such as fever, fatigue, weight loss, lack of appetite, chest pain, persistent cough and haemoptysis (coughing up blood) [1]. However, in some patients TB has been classified as subclinical as patients are culture-positive but asymptomatic [1,4]. Over time, it has been thought that cough was mandatory for TB transmission, although recent studies have evidenced that the presence of symptoms is not essential for MTB spread [5]. Despite being easily transmissible, when tubercle bacilli arrive to the lungs after inhalation, they can be eliminated by the immune system or persist inside granulomas in a dormant, or non-replicative, state. These latency periods can vary from months to years. Remarkably, only 5-15% of infected patients develop the active TB disease [6] and nearly a quarter of the culture-positive people have no obvious symptoms, posing a challenge for diagnostics [7,8] (**Figure 3**).



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 3: Stages of Tuberculosis pathogenesis. From MTB infection to active disease. Figure adapted from Pai *et al.* 2016 [1].

MTB detection and diagnosis

TB is firstly suspected by clinicians based on patients symptoms. Confirmation of active TB requires a microbiological diagnosis, typically achieved through sputum smear microscopy, the conventional approach. Smear microscopy is based on the identification and quantification of MTB bacilli. The most commonly used technique is based on the Ziehl-Neelsen acid-alcohol stain due to the acid-fast nature of MTB bacilli and its thick cell wall [9]. It requires a good quality sputum to ensure the sensitivity of the technique and personnel trained to reliably identify and distinguish mycobacteria from debris that can be non-specifically stained [10]. However, this technique has low sensitivity and specificity and is not able to differentiate between MTBC species [10]. In that regard, culture is still the most widely used technique for both MTB detection

and drug-susceptibility testing (DST) and plays an important role in the outcome prediction [11,12]. MTB can be cultured either in liquid (Bactec Mycobacteria Growth Indicator Tube or MGIT, Middlebrook 7H9) or solid media (Löwenstein–Jensen or Middlebrook 7H11). However, its main drawback is the extended time required to obtain the results, as MTB is a slow-growing organism [12].

Given the limitations of microscopy and culturing, the WHO has endorsed molecular testing (e.g. GeneXpert Ultra) in low- and middle-income countries for faster and more accurate diagnosis [13]. Molecular testing not only identifies the presence of MTB but also detects some resistance mutations. This is why WHO recommends substituting smear microscopy with WHO-approved molecular tests to enhance the sensitivity of primary TB diagnostics and provide a more comprehensive result [13,14]. Furthermore, other molecular techniques that are currently used consist of sequencing the whole genome (WGS) or targeted regions (tNGS) by using next-generation technology (NGS). These approaches are the ones used for achieving the objectives of this thesis, thus are explained in more detail in the following sections.

Diagnosis of MTB latent infection is more challenging than active TB. It can be primarily detected by tests based on the immune response identification. These tests are mainly conducted to screen actively for new TB suspicious cases, for example in people that have been in contact with patients diagnosed as TB positive; before giving immunosuppressive therapy to prevent activation of MTB or in other risk groups as people living with HIV [15].Tuberculin skin test (TST or Mantoux) measures the hypersensitivity reaction when exposing a patient to the MTB specific purified protein derivative (PPD). Thus, it can give false positive results in BCG vaccinated patients. The other technique,

Interferon-Gamma Release Assay (IGRA), detects the secretion of interferon-gamma by T-cells when exposing them to MTB-specific antigens [16]. Imaging techniques, as well as chest radiography, are also used to screen for lung lesions [1]. Nevertheless, it is important to note that these tests have a low predictive value and do not distinguish between latent and active TB [1].

Treatment

Making a prompt, complete and accurate diagnosis is crucial to provide an optimum personalised treatment to the patients and prevent TB transmission. The selection of an accurate treatment depends on the drug susceptibility profile of the pathogenic strains. The recommended treatment for susceptible TB (DS-TB) consists of a 6-month regimen combining isoniazid (H), rifampicin (R), pyrazinamide (Z) and ethambutol (E) during the first 2 months, followed by four months with H and R [17]. In 2022, the WHO updated the treatment guidelines for drug-resistant TB (DR-TB) by shortening the regiments, which were previously longer than those for DS-TB and more toxic [2,17]. Therefore, for those multidrug resistance (MDR) cases with isoniazid and rifampicin resistance (or rifampicin monoresistance), commonly known as MDR/RR-TB, the WHO recommends a 6-month all-oral regimen treatment (which may be extended 3 months if needed) consisting of bedaquiline (B), pretomanid (Pa), linezolid (L) and moxifloxacin (M) known as BPaLM [18]. For those cases with additional resistance to fluoroquinolones, also known as pre-XDR-TB (pre-extensively DR-TB), moxifloxacin has to be removed and the regiment continues as BPaL [18]. A BPaL extension up to 9-months all-oral is recommended when culture remains positive between months 4 and 6 [18]. Moreover, longer regiments of 18-20 months including injectable drugs

(amikacin) are recommended when shorter regimens can not be implemented due to drug intolerances, drug-drug interactions, treatment failure or XDR-TB (pre-XDR-TB with additional resistance to at least bedaquiline) [2,18]. It must be noted that this update on DR-TB treatment guidelines is a great achievement for patient care. Nevertheless, challenges persist in providing widespread access to new drugs such as bedaquiline and also DST tests for all countries. Thus, a current major concern is to address these challenges as soon as possible, facilitating a smooth and global implementation of the updated guidelines.

Prevention and disease control

Global efforts are crucial for tuberculosis control to decrease the incidence and mortality rates by ensuring the access to high-quality diagnostic and treatment in all countries [19,20]. A key point in the disease control lies on stop transmission, emphasising the importance of early diagnosis and treatment [21]. Hence, prevention plays an important role for TB control. The current and unique approved attenuated vaccine is still BCG (*Mycobacterium bovis* bacillus Calmette–Guérin) since 1921 [22]. It derives from *M. bovis* and it is used in more than 80% of neonates [23], especially in high-burden TB countries [24]. It offers moderate efficacy in preventing severe forms of TB in infants and young children [25]. However, over the years, it has been proven to be ineffective in preventing primary MTB infection and, more importantly, reactivation of latent infection in adolescents and adults, who are responsible for 90% of TB transmission [26]. Advances in molecular tools and design of viral vectors and adjuvants have facilitated the development of new vaccines and there are, currently, more than a dozen vaccines under clinical trials [22,27].

Beyond vaccines, preventive treatment is a crucial measure to stop progression from MTB infection to clinical disease. The World Health Organisation (WHO) recommends preventive treatment to high-risk groups like immunocompromised patients as people living with HIV or household contacts [15]. This involves an isoniazid (may be combined with rifampicin or rifapentine) prophylaxis regimen [2]. Therefore, screening for MTB infection is of vital importance to identify individuals that require preventive treatment.

Mycobacterium tuberculosis biology and evolution

Mycobacterium genus comprises close to 190 distinct species. It was defined as a genus after the identification of the *Mycobacterium tuberculosis* complex (MTBC), relying on phenotypic characteristics within the complex which include the mycolic acid-enriched cell wall, anaerobic growth and the bacillary shape [28]. This genus encompasses major human pathogens such as *Mycobacterium leprae* and *Mycobacterium tuberculosis* (MTB), as well as nontuberculous (NTM) opportunistic mycobacteria such as *Mycobacterium abscessus* and *Mycobacterium avium [29].* Within the members of the MTBC, *Mycobacterium tuberculosis* (MTB) and *M. africanum* are the responsible agents for human infection. Other MTBC ecotypes are mainly adapted to wild and domestic animals: *M. caprae, M. bovis, M. pinipedii*, *M. microti, M. mungi, M. orygis* and *M. suricattae [30]*.

Mycobacterium tuberculosis was identified in 1882 by Robert Koch as the causative agent of tuberculosis (Koch 1882). MTB is a slow growing, non-motile, aerobic microorganism. In spite of being classified as gram-positive, its cell wall

functionality closely resembles that of gram-negative bacteria. Its main characteristics include the presence of two outer membranes, the first one composed by peptidoglycan and arabinogalactan; and the second one riched in mycolic acids with long chain branched fatty acids, glycolipids, polysaccharides [31–33]. Indeed, its waxy outer membrane together with the lipid synthesis capabilities contribute to the MTB virulence, antibiotics impermeability and pathogenicity [34,35].

Human adapted MTBC ecotypes are classified in 9 major lineages (L) [36] (**Figure 4**), with lineages L8 [37] and L9 [38] being the most recently discovered. Within this classification, lineages are grouped in "ancestral" or "modern" depending on the presence or absence of the TbD1 genomic region. The "modern" lineages, include L2, L3 and L4, also known as Beijing (L2), CAS/Delhi (L3) and Haarlem/Ural/X/LAM (L4), and are responsible for the vast majority of cases in high burden TB countries [39]. They present a deletion of TbD1, a marker of evolution [40] that provides a better adaptation to oxidative stress and hypoxic conditions generating advantage to host-pathogen interactions during the infection. This explains the global predominance of L2, L3 and L4 nowadays [39]. However, "ancient" lineages, such as L1, L5 and L6 are more restricted to concrete regions. L1 or Indo-Oceanic, is prevalent in East Asia, The Philippines and the Rim of Indian Ocean [41] corresponding to 30% of cases. L5 and L6, belonging to *M. africanum*, are mainly located in West Africa [38]; L7 is located in the Horn of Africa and Ethiopia [42]; and L8 and L9 in East Africa. L8 was identified in the African Great Lakes region [37] and L9 in Somalia [38]. Further beyond these major classifications, the modern lineages can also be divided into sublineages that also correlate with the geographical distribution of human populations [43,44].

Importantly, lineage plays an important role in TB outcome, virulence and drug-resistance emergence [44]. Lineages 2 and 4 are associated with more virulent phenotypes compared to ancient lineages such as L1 and *M. africanum*. This virulence variability within patients may be due to variations in specific genomic regions (RD1-15) over time, which include several virulence associated genes [45]. Mutations in RD regions are related to patient relapse and occur more frequently in L2 strains [45].

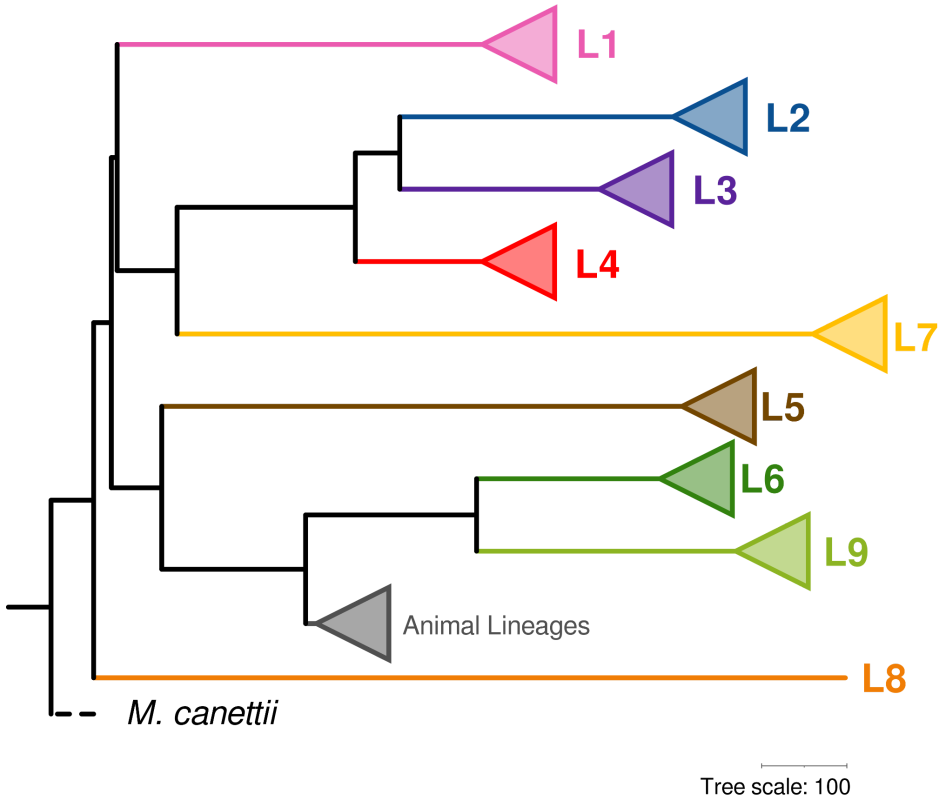


Figure 4: Neighbour joining tree of *M. tuberculosis* strains belonging to all 9 human and animal lineages. Figure adapted from Coscollá *et al.* 2021 [38]. The topology is rooted with *Mycobacterium canettii*.

Mycobacterium tuberculosis genomics

MTB genome characteristics

The complete sequence of the *M. tuberculosis* genome was published for the first time in 1998 [35]. The MTB genome is characterised by a 4.4 Mb size, 65% of GC content and a high amount of repetitive regions such as multigene families and duplicated housekeeping genes [35]. MTB is a clonal mycobacteria and the average rate of acquired mutations per genome is 0.5 single nucleotide polymorphisms (SNPs) per year [46]. Variations on the MTB genome rely on SNPs, insertions or deletions; there are no horizontal gene transfer nor recombination mechanisms [47].

Sequencing technologies

Next-generation sequencing approaches are now affordable, scalable and high-resolution to provide high-quality sequencing data not only in research but also in clinics. The standard workflow consists of growing MTBC strains in liquid or solid media, extracting DNA, preparing libraries and sequencing. Regarding the analysis pipeline, it involves several filtering steps to discard non-MTBC reads and low-confidence variants during the mapping and variant calling steps. The most widely used technology is based on short-read sequencing (Illumina platform), but there is a growing trend towards long-read sequencing (Nanopore and PacBio) technologies [48]. Each sequencing technology offers different advantages and limitations for MTB research. While short read sequencing has high accuracy and low error rate to detect single nucleotide polymorphisms (SNPs) and small insertions or deletions (indels), it is

not able to fully resolve the challenging regions of MTB genome including repetitive gene families or duplications thus limiting the comprehensive analysis [49]. In contrast, long-read sequencing allows a better characterization of these large structural variations which may contain relevant epidemiological information or may be involved in virulence and drug-resistance mechanisms offering a more complete picture of the genome [49,50].

WGS applications

With the identification of genetic variants, whole genome sequencing (WGS) of MTB has several applications including TB surveillance, drug-resistance profiling, phylogenetic classification and epidemiology (**Figure 5**) [48].

Surveillance and DR prediction

Sequencing of MTB is currently used for drug-resistance TB surveillance in some countries such as the United Kingdom [51], specially targeted NGS that has been recently endorsed by the WHO [52]. The development of catalogues containing drug-resistant associated mutations standardises the interpretation of the detection of variants conferring resistance to several anti-TB drugs. Application of WGS for drug-resistance prediction is detailed in a separate section as it has been used to achieve the objectives of this dissertation.

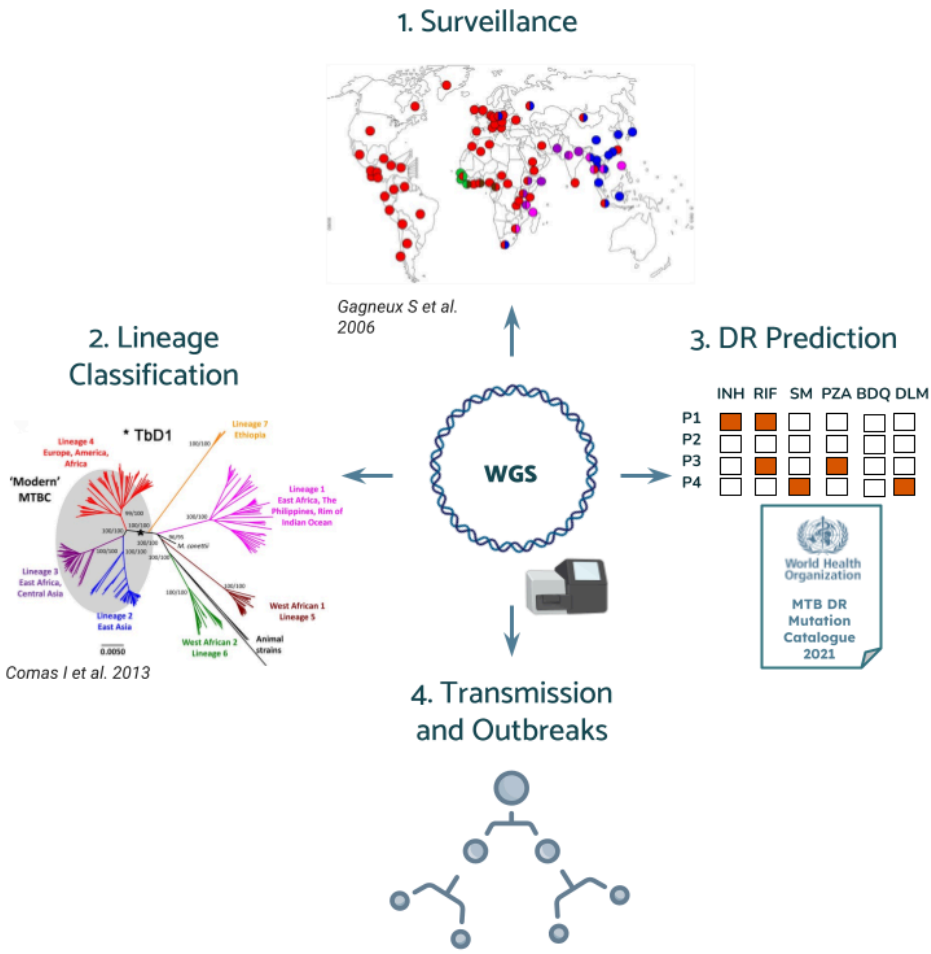


Figure 5: Applications of *M. tuberculosis* WGS adapted from Meehan C. *et al. [48]*.

Genotyping and transmission

Similar to DR-mutations catalogues, genotyping MTB strains by using WGS also requires lists containing phylogenetic determining variants which are

lineage-specific SNPs [43,44]. WGS provides higher resolution than other molecular techniques used for MTBC genotyping [53]. In the past decade, the RFLP (restriction length polymorphism) method, based on the IS6110 region, was used to track MTBC strains [54]. More recent molecular techniques include spoligotyping and VNTRs. Spoligotyping is a reverse hybridization PCR based approach that identifies polymorphisms in the direct repeat (DR) locus [55], while VNTR is based on the identification of variable number tandem repeats of the mycobacterial interspersed repetitive units also known as MIRUs [56]. Although these techniques are cost-effective, they fail on identifying some lineages [57,58]. Therefore, using SNPs for phylogenetic classification is ideal as they represent unique events and show no homoplasy [58].

Regarding transmission, genomic epidemiology is based on comparative genomics of MTBC strains to obtain pairwise-genetic SNP distances within strains [59]. WGS offers higher performance than contact tracing [60] and better resolution to identify transmission clusters than traditional approaches such as the previously mentioned VNTR [59]. Additionally, WGS also can identify outbreaks, super-spreaders and infer transmission directionality [59]. The low variability of MTB allows the standardisation of SNP cut-off values to establish transmission groups. Thus, a threshold of 5 or 12 SNP distance is the most common used to identify a recent transmission group [48,60,61].

Sequencing the MTB genome from diagnostic samples

Obtaining sufficient bacteria is a key point for achieving high-quality MTBC sequences. Therefore, the initial step of the workflow involves culturing diagnostic samples in solid or liquid media, which can significantly increase the turnaround-time of the technique by up to months due to the slow growth rate of *M. tuberculosis* strains, especially when performing phenotypic drug-susceptibility testing. To overcome this time-consuming limitation, WGS directly MTB from diagnostic samples is a promising alternative both for research and clinical applications.

Culture-free WGS (cfWGS) has been demonstrated to be possible in the last decade. Several research groups have performed cfWGS mainly for drug-resistance prediction but also to study epidemiology, obtaining accurate results [62–66]. The primary challenge of cfWGS is the composition of the diagnostic samples, mainly sputa. Sputum samples contain low amounts of MTBC DNA present in a complex matrix with plenty of human DNA and DNA from other microorganisms [67]. Considering this complex composition, sputum samples require additional processing steps to deplete non-MTBC DNA. First, a homogenization step needs to be performed to decontaminate the sputa without affecting the viability of mycobacteria [68,69]. Second, regarding the DNA extraction protocol, methods based on a differential cell lysis are preferred to eliminate non-MTBC DNA from the sputum sample [62,65]. Nonetheless, these approaches do not yield MTBC DNA enough to achieve the minimum depth for sequencing in most sputa [70]. Additionally, an enrichment step is required to capture MTBC genomes. Brown *et al.* [66]

designed an enrichment approach that allowed the recovery of 90% of the genome at more than 20X depth. It was based on an RNA-bait hybridization performed after preparing the sequencing libraries. Then, a PCR step amplifies the fragments captured by the RNA-baits [66]. Other enrichment approaches using DNA-based capture protocols have also been developed [71]. Despite the increased costs associated with these approaches, some studies have explored alternative strategies to make them more affordable [65,72].

Once the MTBC sequences are obtained, the bioinformatics pipeline to analyse them is also of great importance to obtain high confidence results. Even though strategies to minimise non-MTBC DNA are employed, remnants from other organisms may persist after sequencing, compromising the accuracy of downstream analysis [73]. Therefore, an initial step to classify the reads taxonomically and filter those belonging to MTBC, is required [74]. Furthermore, applying additional filters in the downstream analysis of variants is also crucial to remove amplification and mapping errors. However, there is currently a lack of standardised bioinformatic pipelines that comprehensively evaluate this process, which complicates the comparison of results within research groups [75].

It is commonly assumed that, when culturing from diagnostic samples, the original genetic diversity present in the patient's lungs is not well represented [76]. Published results based on post-mortem sampling and from lung surgeries have demonstrated that there are different genotypes depending on the body location of samples that, often in sputum samples, are not present [57,76]. Specially, in those cases where one of the genotypes has higher fitness cost due, for example, to drug-resistance. Despite showing the near real MTBC diversity within individuals, for obvious reasons, surgeries are risky and are not

the best way for diagnosis sample collection. In addition, in these studies, biopsies were cultured prior to sequencing to obtain enough bacteria, as performed in the vast majority of research and clinics [57,76].

Nevertheless, little is known about the bottleneck effect that culture may impose on genetic diversity. It is not strange to consider that bacteria more suited for *in vitro* growth may predominate and alter the original diversity during culture. A good example is lineage 6, which has been demonstrated that its growth rate is significantly slow compared to other lineages [77]. In addition, when studying other pathogens, such as *Helicobacter pylori,* an alteration on genetic diversity producing changes in virulence have been observed [78]. Few studies have achieved enough sequencing quality in sputum samples to compare the diversity between sputum and cultures, getting conflicting results. On the one hand, Votinseva *et al.*, Brown *et al.* [66] and Goig *et al.* [65] showed that the resistance profile obtained was the same in each sputum and culture pair [62,65,66,79]. On the other hand, Shockey *et al.* and Nimmo *et al.* results reported a higher genetic diversity in sputum samples compared with cultures [64,80,81]. Possible reasons that may explain these contradictory results include not only the usage of different approaches in sample collection and processing but also the analysis pipelines applied. Furthermore, composition of sputum samples, including mixed or polyclonal infections together with the presence of different DR-TB genotypes may also contribute to these differences [76,81]. Overall, sequencing MTBC directly from sputum samples is now a reality and offers significant advantages for both clinical and research settings, especially in terms of time efficiency. Although there is still a lack of standardisation of sample processing workflows and bioinformatic pipelines to achieve high confidence results.

Antimicrobial resistance and DR prediction tools

Since the release of the first antimicrobial drugs, resistance has become one of the most important challenges for global public health. In 2022, the estimated number of people developing drug-resistant TB (DR-TB) was 410,000 [2]. In addition, the proportion of MDR/RR-TB new cases was 3.3% and 17% in previously treated cases. Despite there has been a downward trend in the DR-TB incidence between 2015-2019, the number of new cases is still high. According to the WHO estimates, 73% of all the pulmonary TB confirmed cases (2.9/4.0 million) were tested for rifampicin resistance and, among them, only the 4.4% (176,586/4.0 million) were classified as MDR/RR-TB, pre-XDR or XDR [2]. While the number of people tested was higher than the absolute number reported in 2019 (2.2/3.6 million), the total number of MDR/RR-TB cases detected was lower in 2022 than 2019. These numbers are in line with the estimated decline in the prevalence of MDR/RR-TB cases [2]. Furthermore, globally the amount of MDR/RR-TB patients enrolled in treatment in 2022 was higher than in 2020, but still lower than pre-pandemic figures in 2019.

MTBC acquires resistance through single nucleotide polymorphisms (SNPs), indels and chromosomal rearrangements in genes that include drug target sites, enzymes that activate the prodrug or drug efflux systems [82]. These mechanisms are the basis for the survival of the bacteria even in the presence of the drug. Resistant populations persist when they are not accurately detected or when treatment is not correctly followed, leading to suboptimal outcomes and a continued transmission of the strains [1]. Therefore, it is crucial to make a prompt and accurate diagnosis.

Phenotypic tests

Drug-resistance detection requires both bacteriological confirmation of MTBC and DR testing by culturing or performing molecular assays. Drug susceptibility testing (DST) assays are not always accessible and depend on the setting [83]. Several options exist. The gold standard technique is culture-based and relies on phenotypic drug susceptibility testing (pDST). This involves a confirmed TB sample being grown in solid or liquid media containing an antibiotic at its critical concentration. In this regard, MTB growth indicates that the strain is resistant, while the absence of growth is associated with susceptibility [84]. Solid cultures are slower than liquid ones requiring 30-45 and 10-42 days, respectively [85]. One of the most common pDST assay for first line drugs is performed by using BD BACTEC MGIT system that test SIRE (streptomycin, isoniazid, rifampicin, ethambutol), plus pyrazinamide in a different tube [86]. pDST assays also allow measuring the minimum inhibitory concentration (MIC) of the antibiotic, which is the lowest concentration repressing the strains growth. Commercial microdilution plates contain standardised lyophilized drug concentrations. These include Sensititre MYCOTB MIC plates that, a part from the most common first-line drugs, they also test several second line drugs such as moxifloxacin, ofloxacin, amikacin, cycloserine and para-aminosalicylic acid (PAS), in total 12 antimicrobials [87]. These microdilution plates are expensive, but cheaper non-commercial alternatives exist. The resazurin microtiter assay (REMA) is also used for MIC testing and is based on a colorimetric method that uses the reduction of resazurin as a mycobacterial growth indicator [88]. Although pDST assays are the most commonly used, they are slow, technically demanding in terms of laboratory equipment and trained technicians, and

there is a lack of standardisation that compromises their reproducibility between laboratories [83].

Molecular tests

Molecular DST assays are based on the detection of drug-resistant conferring mutations (**Figure 6**). They are faster than pDST as they can be performed directly on the diagnostic samples [89]. These techniques are mainly based on an amplification of targeted (NAAT) regions of the MTBC genome. Several commercial tests have been developed, but currently, the most globally used and recommended by the WHO is the GeneXpert MTB/RIF and its updated version XpertUltra (Cepheid, Sunnyvale, CA, USA) [90]. XpertUltra is an easy-to-use assay based on a real-time PCR (qPCR) occurring inside a cartridge. It detects MTBC by targeting two multicopy genes (IS6110 and IS1810) as well as rifampicin resistance by targeting a specific region of the *rpoB* gene containing the most common mutations (RRDR, rifampicin resistance determining region) [91]. Despite its high sensitivity and specificity for RIF-R detection, XpertUltra has not been designed to target a broad spectrum of mutations. Therefore, RIF-R strains harbouring uncommon mutations (e.g. *rpoB*_I491F) are misdiagnosed [92] potentially contributing to the spread of resistant strains; as happened in countries such as Eswatini, South Africa and Botswana [93–95]. Furthermore, to identify pre-XDR and XDR cases, a new cartridge has been developed, the Xpert MTB/XDR test that detects DR associated mutations to isoniazid and second-line drugs (fluoroquinolones, kanamycin, amikacin and capreomycin) [96,97]. Its major limitation is that it is performed separately from Xpert MTB/RIF increasing the overall diagnostic costs. In addition, its implementation requires new instrument modules with a 10-colour multiplex

technology which represent shipping and cost challenges particularly for middle and low-income countries [98].

Another widely used molecular approach is line probe assays (LPAs), such as the Genotype MTBDRplus and MTBDRsl 2.0 assays (Hain LifeScience GmbH, Nehren, Germany) [99,100]. These tests target resistance conferring mutations to both first and second line anti-TB drugs. However, their low sensitivity limits their usage to cultures or smear-positive sputum. This limitation, together with their complex laboratory requirements, makes them unsuitable for point-of-care diagnostics [101].

Figure 6: WHO recommended molecular tests in 2023 report for MTB diagnostic. Asterisk highlights the test that only detects MTB but not drug resistance.

Whole Genome Sequencing (WGS) overcomes molecular NAATs-based testing's limited detection range as it provides a full report of the whole genomic variation. Moreover, targeted sequencing (tNGS) is also an attractive

alternative for drug susceptibility prediction as it requires fewer bacteria than WGS and can be easily applied to diagnostic samples [102]. The combination of tNGS with portable sequencing technologies like MinION (Nanopore) increases its potential to be used in the clinics and point-of-care due to the advantages of MinION in terms of costs, maintenance, size and the real-time analysis [48,102]. Some gene panels have been designed, commercialised and even recommended by WHO for DR diagnosis [52], like Deeplex Myc-TB (Genoscreen, Lille, France). The Deeplex Myc-TB targets the full sequences or relevant regions of 18 genes associated with DR, in combination with lineage-determining regions [103]. Hence, it provides drug resistance profile and the genotype of the strains. In addition, it also includes a web-based user-friendly platform for the analysis [103].

The usage of mutation lists allows to standardise the interpretation of DST prediction from sequencing results [104]. The WHO has released in 2023 the second version of a catalogue containing more than 30000 mutations associated with resistance and susceptibility to 13 anti-TB drugs including first and second line [104,105]. This catalogue has been created by using phenotypic-genotypic associated data obtained from several research groups, consortia, public health organisations and published results. It provides valuable information for reference laboratories, surveillance programs and developers of molecular tests [104]. In addition, it presents more than 80% sensitivity for the prediction of mutations associated with the main first line drugs, and more than 95% specificity for all drugs except ethionamide, ethambutol and moxifloxacin [104,106]. Furthermore, it includes variants associated with resistance bedaquiline, delamanid and linezolid [105].

The advent of molecular techniques has been a major step forward for MTB diagnostics. Although, sometimes, discrepancies in the phenotype and genotype prediction of the resistance profile have been reported. These conflicting results can be false positives or false negatives, and may cause confusion and lead to delays or suboptimal treatments. Some scenarios contributing to phenotypic-genotypic discrepancies are related to the presence of DR mutations undetected by the molecular test used or the presence of a low-level resistance conferring mutation, silence mutations, heteroresistance or an unknown DR mechanism. When comparing with WGS, the presence of compensatory mutations may also explain these differences. Technical problems in the pDST also contribute, as well as errors associated with the amplification or probe interaction in the molecular test performed [89]. In particular, regarding pDST, when the MIC of the strain is close to the critical concentration, minor variations in the media used in the pDST or in the initial inoculum concentration, may lead to misinterpretation of results [107,108]. For this reason, this technique should only be applied in laboratories that maintain high standards of performance [89,107].

Molecular techniques, including WGS and tNGS, provide a rapid diagnostic that can guide treatment for the majority of patients, avoiding unnecessary exposure to toxic drugs when resistance is detected. Nowadays, the traditional concept of gold standard technique for diagnosis has been challenged by the appearance of discrepancies between molecular and phenotypic drug-resistance profiling in some cases. Therefore, it has to be assumed that each drug may have a unique standard and it is important to note that, sometimes, the genotypic results will take priority over phenotypic results [89,109].

Objectives

Mycobacterium tuberculosis research and clinics have commonly relied on growing the bacilli in vitro. However, the real consequences of the bottleneck effect imposed by the culture on the genetic diversity of the bacterial population are still unknown. Contradictory results have been obtained by different research groups highlighting the need for further investigation. The aim of this dissertation is to sequence sputum samples and compare the genetic diversity within their paired cultures. Additionally, the potential of whole genome and target sequencing for drug resistance prediction are also explored. These objectives would help to consider sputum sequencing as a frontline tool to enhance the rapidity and accuracy of tuberculosis diagnosis as well as for research purposes. They are described in more detailed below:

Objective 1: Comparison of *Mycobacterium tuberculosis* genomic diversity within sputum and culture paired samples. This objective involves culture-free sequencing techniques and the design of an analysis pipeline tailored to sputum samples. In addition, sputum-culture pairs from different settings are compared, including already published datasets. This objective is addressed in Chapter 1.

Objective 2: Monitoring drug-resistance mutations in Southern Mozambique by using Whole Genome Sequencing. This aim is developed in Chapter 2. It explores the usage of whole genome sequencing to predict the resistance profile in Southern Mozambique and compares it with the molecular frontline technique currently used there.

Objective 3: Designing a gene panel to predict drug resistance profile and lineage of *Mycobacterium tuberculosis* strains from both sputum and culture samples. Chapter 3 includes the development of a set of amplicons tailored for resistance profiling and genotyping MTB strains by using the MinION. As a portable sequencer, this platform can be used in point-of-care centres for a prompt and accurate diagnosis of tuberculosis including drug-resistance and lineage prediction. This objective evaluates this technology to be used in both sputum and culture samples.

Chapter 1

Genetic diversity within diagnostic sputum samples is mirrored in the culture of *Mycobacterium tuberculosis* across different settings

Carla Mariner-Llicer, Galo A. Goig, Manuela Torres-Puente, Sergo Vashakidze, Luis M. Villamayor, Belén Saavedra-Cervera, Edson Mambuque, Iza Khurtsilava, Zaza Avaliani, Alex Rosenthal, Andrei Gabrielian, Marika Shurgaia, Natalia Shubladze, Alberto L. García-Basteiro, Mariana G. López, Iñaki Comas

Published in Nature Communications

DOI: https://doi.org/10.1038/s41467-024-51266-0

Publication Cite:

Mariner-Llicer, C., Goig, G.A., Torres-Puente, M. *et al*. Genetic diversity within diagnostic sputum samples is mirrored in the culture of *Mycobacterium tuberculosis* across different settings. Nat Commun 15, 7114 (2024). https://doi.org/10.1038/s41467-024-51266-0

Abstract

Culturing and genomic sequencing of *Mycobacterium tuberculosis* (MTB) from tuberculosis cases is the basis for many research and clinical applications. The alternative, culture-free sequencing from diagnostic samples, is promising but poses challenges to obtain and analyse the MTB genome. Paradoxically, culture is assumed to impose a diversity bottleneck, which, if true, would entail unexplored consequences. To unravel this paradox we generate high-quality genomes of sputum-culture pairs from two different settings after developing a workflow for sequencing from sputum and a tailored bioinformatics pipeline. Our approach reveals that 88% of variants called in culture-free sequencing analysis are false positives due to supplementary alignments, mostly in enriched-sputa samples. Overall, contrary to the bottleneck dogma, we identify a 97% variant agreement within sputum-culture pairs, with a high correlation also in the variants' frequency (0.98). The combined analysis from five different settings and more than 100 available samples shows that our results can be extrapolated to different TB epidemic scenarios, demonstrating that for most cases culture accurately mirrors clinical samples.

Introduction

Mycobacterium tuberculosis (MTB) research from clinical samples usually involves a culturing step to obtain sufficient bacteria for downstream applications, including MTB whole-genome sequencing (WGS) studies. As a consequence, our current knowledge of MTB characteristics, including its biology during infection, evolution, epidemiology, and diagnostics, is largely based on cultured samples [48,110]. It has been hypothesised that the

cultivation procedure may constrain the genetic diversity of MTB, either by selecting for specific variants more suited to *in vitro* growth as happens for some drug resistance mutations or lineages [111–113], or simply due to the bottleneck imposed by the culture inoculum [114]. If true, this could distort our understanding of bacterial diversity, particularly at the within host level and even affect epidemiological and drug resistance inferences that rely on the presence or absence of a few single nucleotide polymorphisms (SNPs). Direct sequencing from clinical samples is an alternative as it could bypass the potential disadvantages of culturing [115]. However, the implementation of culture-free sequencing techniques is challenging due to the complexity of the sample matrix with low amounts of mycobacterial DNA and a mix of contaminants including host genetic material [73,116].

Previous efforts on culture-free genome sequencing have focused on developing new and affordable protocols for MTB culture-free WGS and assessing their reliability for clinical applications, mainly for AMR diagnosis and transmission inference [62,63,65,66,80,117]. Those studies show contradictory results regarding overall genetic diversity comparison between culture-based and culture-free WGS. Some publications have identified no significant differences [62,66], while others have reported a reduction in genetic diversity when culturing [80,118]. These contradictory results probably reflect limitations of the studies, often focused in one single setting and the limited quality of culture-free sequences prevent a proper comparison of genetic diversity between settings. This is specially true to identify low frequency variants which are more likely to suffer from any technical (ie. sample processing) and analytical limitations [119]. Therefore, the question of whether culture does actually impose such a bottleneck remains largely unsolved despite its

importance. The main goal of this work is to determine, in different TB settings, whether culture reflects the original MTB variability present in the diagnostic samples, typically sputum, making the culture a suitable sample for clinical research.

Here we have put together our own comprehensive dataset including samples from two settings differing in tuberculosis (TB) incidence, burden of AMR, and HIV co-infection; and achieving sufficient sequencing depth to properly compare genetic diversity, especially regarding low-frequency variants. We successfully sequence 61 high-quality sputum-culture pairs from Georgia and Mozambique, which are middle- and high- burden TB settings. For the sputa, we implemented a culture-free WGS approach based either in direct (dWGS) or bait-enrichment (eWGS) sequencing, depending on the amount of MTB DNA. In addition we carry out experimental benchmarking to detect major sources of artifactual genetic variation in culture-free approaches. We also reanalyze available datasets [65,80,118] to generalise our results across settings and sequencing approaches. Importantly, we develop a tailored analysis workflow to address the absence of standardised laboratory protocols and bioinformatic pipelines, carefully considering artefacts beyond the role of potential contaminants in the sample. Our customised workflow is capable of demonstrating that culture accurately reflects sputum diversity in all the evaluated settings, albeit with some individual exceptions. Our results indicate that the current knowledge of *M. tuberculosis* diversity based on culturing methods is robust, in the scenarios tested, and therefore the genetic diversity observed in culture generally mirrors that present in the diagnostic sample. In addition, we provide in-vitro and in-silico tools to test the correlation in scenarios not contemplated in this work.

Methods

Ethics

The samples from Mozambique came from a study that was approved by the National Bioethics Committee for Health of Mozambique (CNBS, Ref:369/CNBS/17) and the Internal Bioethics Committee of CISM. Regarding samples coming from Georgia, the ethical approval was obtained from the Institutional Review board (IRB) of the National Center for Tuberculosis and Lung Diseases within the framework of observational clinical study NCT02715271. All methods were performed in accordance with the relevant guidelines and regulations. An informed consent was signed by all participants after providing a verbal explanation and written information about the study. Regarding participants under 18 years of age, the informed consent was obtained from their relatives (parents or guardians). All data were de-identified before the analysis.

Dataset

We received sputum-culture paired samples from the Centro de Investigação em Saúde de Manhiça (Mozambique) and the National centre for Tuberculosis and Lung Diseases located in Tbilisi (Georgia). Mozambique and Georgia are considered high- and medium- burden tuberculosis countries. No statistical method was used to predetermine sample size. To ensure that the bacillary load present in sputum samples was enough for sequencing, we selected 50 sputum samples from Mozambique graded High/Medium according XpertUltra result and 3+/2+ according to smear microscopy result. Regarding Georgia, we

processed all 45 sputum samples received due to the lack of bacillary load information. In summary, we processed 95 sputum-culture paired samples.

Sputum samples, collected at diagnosis, were homogenised and decontaminated in origin countries. In Mozambique, they performed the N-acetyl-L-Cysteine-sodium hydroxide (NALC-NaOH) or Kubica method [120], while in Georgia they followed the modified Petroff protocol, that uses 4% NaOH, which was validated under supervision of the WHO Supranational TB Reference Laboratory in Antwerp [121].

The paired cultures were grown in 7H11 solid media, supplemented with OADC and glycerol to ensure a high amount of bacteria for sequencing.

All samples, cultures and sputa, were received and processed at FISABIO biosafety level 3 (BSL-3) facilities in Valencia, Spain.

Samples' Processing

DNA extraction from sputum samples

DNA extraction from sputum leftovers was based on a differential cell lysis procedure to remove non-MTB contaminant DNA using MolYsis basic5 kit (Molzym, Germany). We followed a modified version of the original protocol that entailed an initial lysis of non-mycobacterial cells to remove contaminant DNA, followed by an inactivation of MTB cells at 95°C for 15 minutes and a mechanical cell disruption using FastPrep. DNA precipitation steps were performed using ethanol, sodium acetate and Glycoblue. DNA extraction steps are detailed as follows:

1. Saline wash:

1.1. Centrifuge 1 mL of sample at 13,000rpm during 15 min

1.2. Remove supernatant by pipetting without disturbing the pellet (leave 200uL of liquid)

1.3. Add 1 mL of PBS and resuspend the pellet by pipetting.

2. Lysis of human cells with MolYsis Basic5 kit:

2.1. Add 250µL of CM buffer, mix by slow pipetting

2.2. Incubate 5 min at room temperature (Note: CM is a chaotropic buffer)

2.3. Add 250µl buffer DB1 and 10µl MolDNase B (do not premix) to each sample and immediately vortex for 15s.

2.4. Incubate at room temperature for 15 min.

2.5. Centrifuge the tubes at 13,000rpm for 10 min.

2.6. Discard the supernatant taking care to not disturb the pellet.

2.7. Add 1 mL of RS buffer and pipette up and down until the pellet is resuspended.

2.8. Centrifuge the tubes at 13,000 rpm for 10 min (During this step, set a thermal block to 95°C for MTB inactivation.

2.9. Discard the supernatant taking care to not disturb the pellet.

2.10. Add 700uL of sterile mqH2O and resuspend the pellet by pipetting up and down.

3. MTB Cells Inactivation:

3.1. Spin down tubes at maximum speed 30 seconds and incubate at 95°C for 15 min.

4. Mechanical cell disruption

4.1. Transfer all the volume 700uL to a 2mL FastPrep Lysing Matrix B tube.

4.2. Break the cells using the FastPrep Mycobacterium tuberculosis program using 1 pulse instead of two (45s at 6.5m/s). Microscopic glass beads break the mycobacterial cell.

4.3. Spin the tubes in the centrifuge and transfer 450 uL to a new microcentrifuge tube.

5. DNA ethanol precipitation procedure:

5.1. Add 1/10 vols (45ul) of sodium acetate 3M to the supernatant.

5.2. Add 1.5ul of GlycoBlue.

5.3. Add 2 vol (1000ul) of cold EtOH 96% and vortex 10s.

5.4. Incubate at -20°C for 30-60 min.

5.5. Centrifuge at 13,000 rpm or maximum speed for 15 minutes and remove the supernatant leaving a small volume of liquid to not disturb the pellet.

5.6. Add 1 ml of EtOH 70% centrifuge at maximum speed for 5 minutes and remove the ethanol without disturbing the pellet.

5.7. Let the ethanol dry (but not overdry), preferably in a vacuum centrifuge machine.

5.8. Resuspend the pellet in 50ul of Tris-Hcl 10mM and dissolve the DNA by pipetting up and down.

DNA extraction from culture samples

Cultures were heat inactivated at 90°C during 30 minutes and then MTB DNA was extracted following the standardised CTAB protocol [122] based on an overnight cell wall digestion with lysozyme, followed by incubation steps with proteinaseK, SDS, CTAB and NaCl and a final DNA precipitation step with

isopropanol. All bacterial cultures and DNA extraction steps were performed in a BSL-3 laboratory.

qPCR Conditions

We assess the concentration of MTB in sputum samples to decide the sequencing approach. The qPCR was performed in a total volume of 20 uL including 10 uL of Kapa Fast Probe Master Mix 2X, 2 uL of forward and reverse primers mix 2.5 uM, 0.6 uL of probe 10 uM; and 1 ng of DNA. We used DNA normalised to 0.5ng/uL. The qPCR assay consisted on the amplification of a 65 bp region within the *Rv2341* gene using the following primers: Forward-GCCGCTCATGCTCCTTGGAT, Reverse-AGGTCGGTTCGCTGGTCTTG, Probe-TGAGTGCCTGCGGCCGCAGCGC [123].

Sequencing selection

We prepared NexteraXT (Illumina) libraries for all samples. Libraries from cultures were sequenced directly. Regarding sputum samples, we classified libraries for dWGS (direct) or eWGS (enriched) depending on Cq and %TB obtained in pre-sequencing runs as described in [65]. Therefore, we performed a qPCR targeting *Rv2341* gene [65,123] to quantify the amount of MTB DNA. Sputa that obtained a Cq value above 35 (equivalent to less than 2 genome copies) were considered negative (not suitable for sequencing) (See ***qPCR conditions*** section). Libraries were prepared for all MTB positive samples (Cq<35) and a pre-sequencing run was performed to estimate the percentage of MTB. According to qPCR and pre-sequencing %MTB result we performed dWGS in sputa obtaining a Cq<25 and more than 20% of MTB and the rest sputum samples underwent an enrichment step before sequencing. The

enrichment step consisted of a MTB DNA capture using RNA biotinylated baits (See ***Enrichment step*** section). See the flowchart in **Figure 7.**

WGS was performed in Illumina Miseq (2x300bp) or NextSeq (2x150bp) platforms. Samples were sequenced until a minimum of 30X median depth and 95% of genome covered was reached.

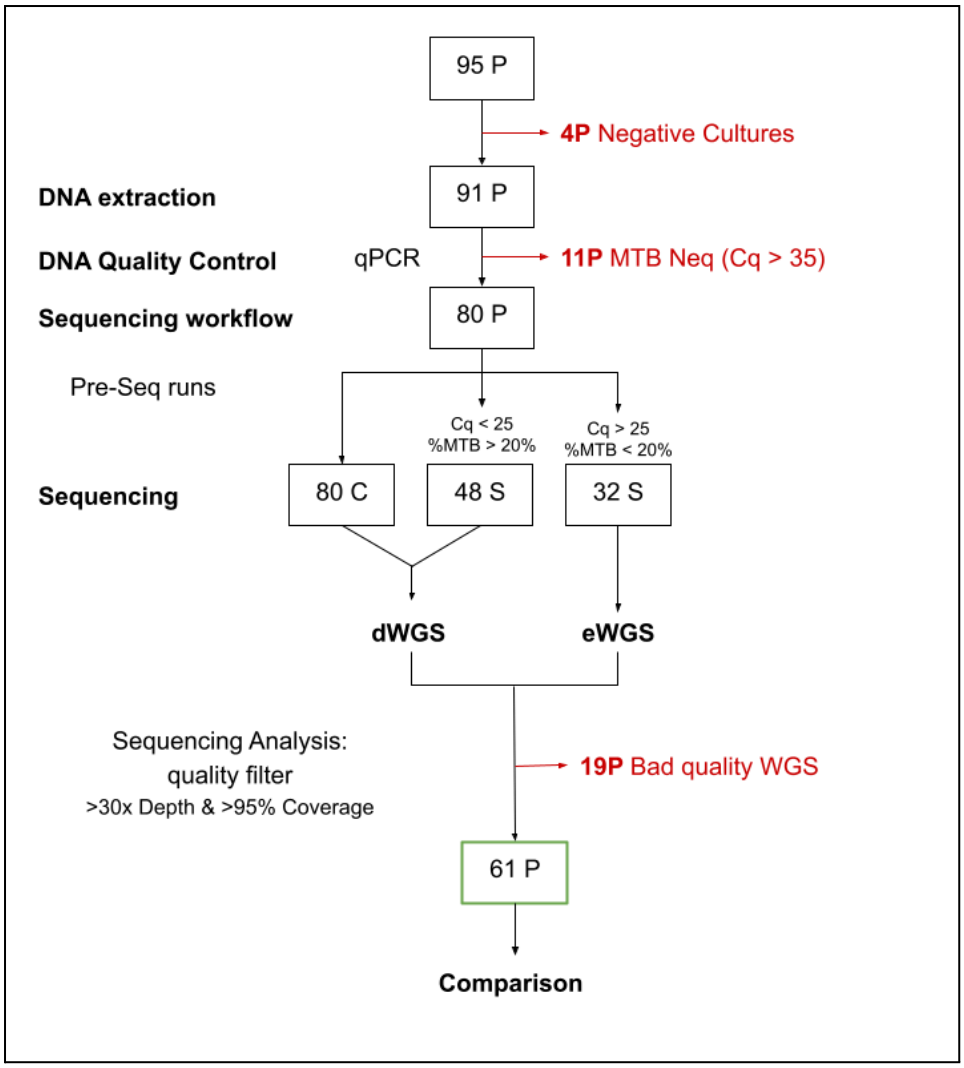


Figure 7: Sequencing workflow. Diagram summarising the sequencing steps and the amount of samples that have passed or have been discarded (in red). Abbreviations: C-Culture, S-Sputum, P-Pair, dWGS-sputum samples not enriched, eWGS-sputum samples enriched.

Enrichment Step

We used myBaits kit (Arbor Biosciences) to perform the hybridization MTB DNA capture step following the myBaits protocol Version 4.01. This protocol is based on a hybridization of already prepared whole genome sequencing libraries (prepared using NexteraXT kit, Illumina) with RNA biotinylated baits at 65°C overnight (at least 24h). During this step, baits hybridised to denatured MTB DNA. Then, most non-captured DNA is discarded by different cleaning steps with streptavidin-coated magnetic beads. Finally, the MTB library is amplified by a post-capture PCR in a 50uL reaction containing 25uL of KAPA HiFi HotStart ReadyMix PCR Kit (Roche), 200nM of Illumina sequencing primers P5 and P7; and 15uL of captured library. Post-capture PRC conditions are the following: 15 cycles of 20 seconds to 98°C, 30 seconds to 65°C and 1 minute to 72°C.

The RNA-biotinylated baits panel was designed and developed by Arbor Biosciences using as reference the inferred ancestor genome of the MTB (NC_000962.3), genetically equidistant to all the MTB lineages. This panel is available upon request.

We evaluated the enrichment step by obtaining the fold-change of the percentage of MTB reads between the 32 sputum samples before and after the enrichment. The median fold-change was 10.96 (**Figure 8**). In other words, sputa containing an initial MTB% within 0.5-10% got a 64.8% (average) while those with an initial MTB% of 10-25% reached 94.3% after undergoing the enrichment step (**Figure 8**).

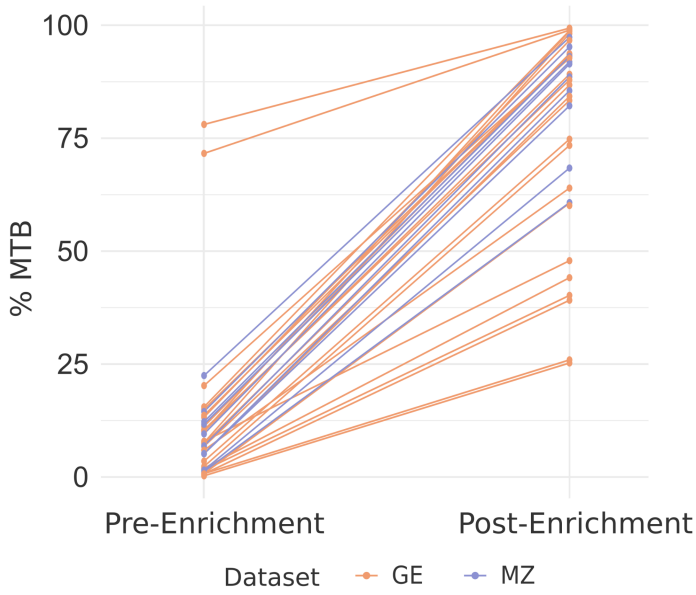


Figure 8: Comparison of percentage of MTB. Comparison of %MTB DNA before and after the enrichment step of the 32 sputum samples that have been enriched. Each point represents a sputum sample. Colours represent samples' origin orange for samples from Georgia and purple for samples for Mozambique. Lines link each sputum sample before and after the enrichment. Source data are provided as a Source Data file.

Bioinformatics Analysis

Core analysis

Analysis was performed using our routine pipeline available at https://gitlab.com/tbgenomicsunit/ThePipeline for both culture and sputum samples. First, reads were trimmed by quality with FastP [124]; non-MTB reads were discarded by using Kraken (v0.10.5) [73,74]. MTB reads were mapped to the reference ancestor genome (NC_000962.3) [125], which is genetically equidistant to all lineages, using BWA (version 0.7.10-r789) [126].

Samples with a median depth below 30X and less than 95% of the genome covered were discarded for downstream variant comparison analysis. For the remaining good quality samples, variants were called with VarScan (version 2.3.7) [127] and Samtools v1.15 [128] by applying stricter coverage and frequency filters for culture samples than for sputum samples (variants called in at least 3 reads, in both strands for sputum; variants appearing in 6 reads, in at 10X depth and in both strands, for culture) (see parameters for VarScan in **Figure 9**). For the genetic diversity comparison analysis, we discarded SNPs appearing in high density regions with GATK (version 4.0.2.1) [129] or genomic regions that are known to be challenging for short-read mapping such as repetitive genes or mobile elements (PE/PPE gene families) [49] by using a customised Python2 (version 2.7.5) script. Highly conserved genes such as *rrs* and *rrl* were also discarded for the comparisons to avoid false positive variability coming from non-MTB reads not discarded by Kraken. We also obtained Pearson's correlation of the variant frequency obtained in each sputum-culture pair (the mean, median and the range).

Sputum & Culture	Sputum Filtering	Culture Filtering
Raw Variants	**All SNPs:**	**All SNPs:**
--min-coverage 3	--p-value 0.01	--p-value 0.01
--min-reads2 3	--min-reads2 3	--min-reads2 6
--min-freq-for-hom 0.90	--min-coverage 3	--min-coverage 10
--min-var-freq 0.05	--min-avg-qual 20	--min-avg-qual 25
	--min-strands2 2	--min-strands2 2
	--min-var-freq 0.10	--min-var-freq 0.10
	fSNPs	**fSNPs**
	--p-value 0.01	--p-value 0.01
	--min-coverage 20	--min-coverage 20
	--min-reads2 20	--min-reads2 20
	--min-avg-qual 25	--min-avg-qual 25
	--min-strands2 2	--min-strands2 2
	--min-var-freq 0.90	--min-var-freq 0.90

Figure 9: Parameters used for variant calling in VarScan. Fixed SNPs (fSNPs) are the ones called at a frequency above 90%. After obtaining SNP files additional filters were applied as described in the *Core analysis* section.

Capture technical validation

To investigate if the capture step introduced bias in variant calling, we analysed sputum samples sequenced by direct (dWGS) and enrichment (eWGS) methods and compared to the respective culture (hereafter, trios-analysis). All discrepant positions within trios, but particularly between eWGS and dGWS samples, were manually checked on the reads alignment to the reference genome in Tablet viewer (v1.17.08.17, see **Supplementary Figure 2**) to identify the causes of the inconsistency. All discrepant variants were identified in supplementary alignments, which are part of a read that is split and aligned to a different part of the genome than the primary alignment [126,130], were identified as the main source of false positive variants. Therefore, we discarded them from bam

files running samtools v1.15 (command: `samtools view -bh -f0 -F256 -F2048 $BAM_FILE`), before variant calling step. Afterwards, we evaluated the impact of discarding supplementary alignments in diversity comparison in all the 61 pairs (sputum-culture) analysed.

After applying all filters described above, a pairwise comparison of variants was conducted within sputum samples and their corresponding cultures. At this point we included a recovery step based on searching for the discrepant SNPs, in the files of the paired sample obtained from a less restrictive variant calling. The reason was that in samples with suboptimal depth the missing SNP may still exist but may have been lost during the filtering steps. Parameters used in VarScan pileup2snp for generating rescued SNP files for comparison analysis were: --min-coverage 2, --min-reads2 1, --min-freq-for-hom 0.9, --min-var-freq 0.01.

The average of rescued SNP shared within sputum and culture pairs was 24 (5-69) representing the 2.5% (0.5-7.9%) of the total variants, highlighting the importance of the step when dealing with samples of heterogeneous coverage.

To verify if rescue was biassing the comparison analysis, we conducted a concordance analysis to corroborate that our rescue approach was not artificially removing true discrepant positions. We analysed 35 sputum-culture pairs with depth of coverage >50x and >95% of the genome covered by applying the same calling pipeline (see the variant calling parameters for cultures in **Figure 9**) and skipping the rescue step. Results showed a median of 886 (range: 828-1059) of common variants, 2 (range: 0-6) sputum-exclusive variants and 2 (range: 0-16) culture-exclusive variants. This corresponded to a median Cohen's Kappa coefficient [131] of 0.998 (0.992-1) which represented an

almost perfect agreement. In this case, the median Pearson's correlation of variants frequencies was 0.939 (0.454-0.991).

Finally, variants appearing in both samples were classified as "common", variants rescued were classified as "common rescued", sputum-exclusive variants were called "only sputum" and those culture-exclusive were classified as "only culture".

Validation using available datasets

We applied our pipeline described to analyse genetic variability on available sputum-culture paired data sets. We downloaded the sequences under the following project accession numbers from ENA: PRJEB9206 [66], PRJNA486713 [118] and PRJEB37609 [65]. Sequences were analysed and only pairs passing the quality filters were compared as explained above (see Bioinformatic Analysis from Methods).

Resistance profiling and phylogenetic classification

Resistance profile was obtained by looking for DR-conferring mutations (>10% frequency) listed in the WHO catalogue (version from 03/09/2021) [104]. Lineage and sublineage classification of the strains was determined by looking for phylogenetic variants reported in bibliography [43,44].

Overall, even though the sequencing quality for some sputum was not enough for the comparison analysis, we were able to classify phylogenetically 74/80 (92.5%) sputum samples sequenced and 72/80 (90.0%) at a sublineage level according Coll *et al.* classification [44].

Pairwise distance between each sputum-culture pair was obtained using the R package ape [132] based on a multiple alignment of fixed SNPs (fSNPs, >90% frequency). In addition, we also obtained the pairwise distance to see whether paired samples clustered together. Neighbour joining trees were constructed using MEGA version X [133].

Results

Selection of sputum-culture pairs

Out of the initial 95 sputum-culture pairs available, 80/95 (84.2%) were suitable for WGS since 4/95 (4.2%) cultures did not grow and 11/95 (11.6%) sputa were MTB negative according to qPCR results and %MTB (Cq>35 and <1%MTB). Given their quality, in terms of Cq and %MTB DNA, 48/80 (60%) sputa underwent dWGS and 32/80 (40%) were enriched before sequencing (**Figure 10a**; see Methods). Percentage of MTB was obtained before and after the enrichment step. We observed that sputa with an initial MTB% within 0.5-10% reached 25.5-98.4% after the enrichment, while those containing initially 10-25% of MTB reached more than 90% of MTB (see Methods).

A total of 19 sputa sequences did not meet the minimum quality criteria (30X depth and 95% coverage) (**Figure 10b**), thus these pairs were excluded from further analysis. The sequencing performance of the remaining 61 pairs was: median depth 48X for eWGS (30-157X), 74X (33-302X) for dWGS and 114X (46-347X) for cultures. Median genome coverage, at a minimum of 20X depth, was above 96% in all sputa and cultures (**Table 1**). See below the samples'

sequencing workflow in the Methods section. All samples' information is available at **Supplementary Table 1**.

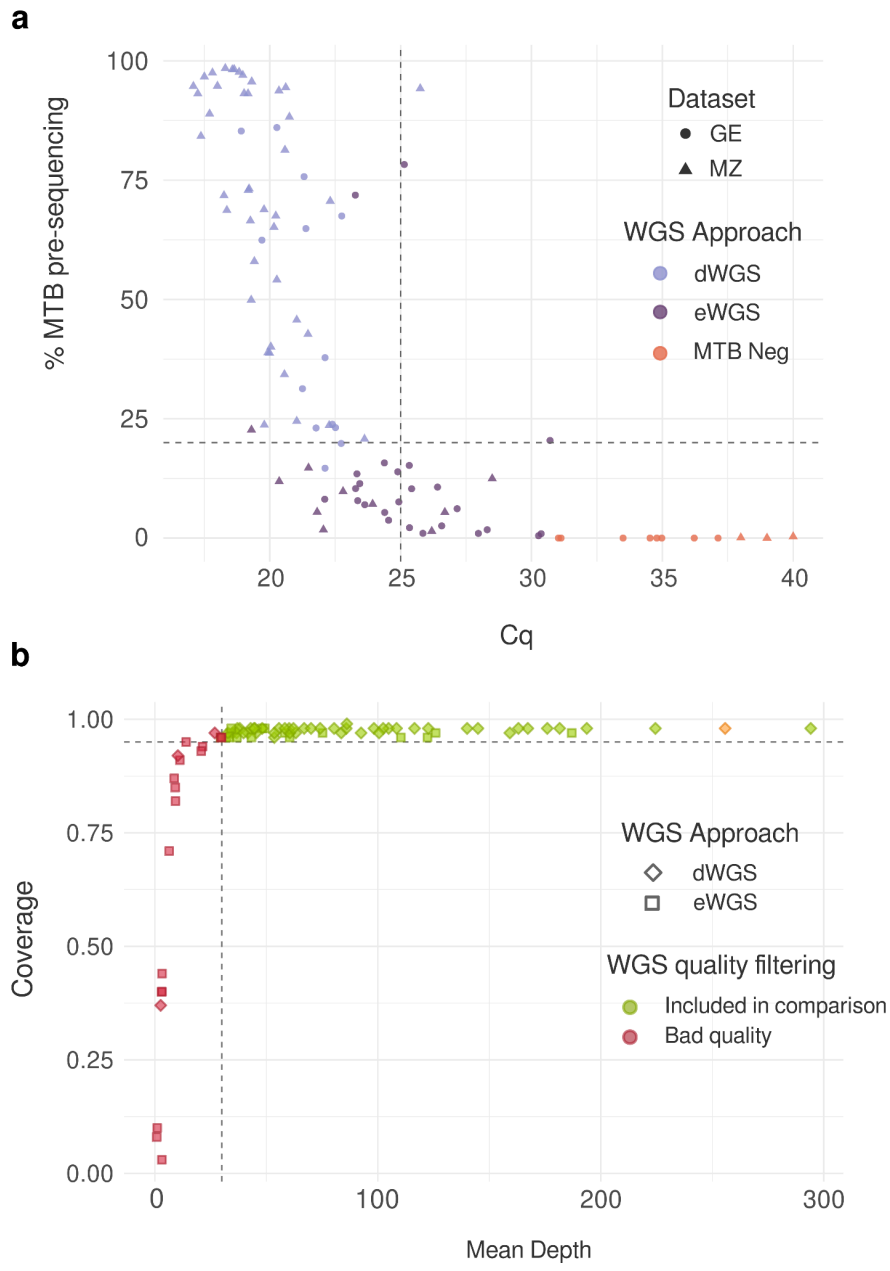


Figure 10: Evaluation of sputum samples for sequencing. **a**, qPCR Cq vs %MTB. 95 sputa are each one represented by a point. %MTB was obtained by performing a pre-sequencing run in order to determine if each sputum

contained enough MTB DNA to be sequenced directly (light purple) or if it required a previous enrichment step (dark purple). Sputum samples in orange were considered negative (Cq>35 and %MTB <1%). Shape indicates sample origin: triangles for Mozambique, dots for Georgia. Dashed lines represent thresholds to decide which WGS approach to follow, the horizontal line highlights MTB% = 20%, the vertical line highlights Cq = 25. **b,** Mean depth vs coverage. 86 sputa sequenced with enrichment (eWGS, square) or directly (dWGS, diamond). Colour represents the sequencing quality, good quality samples (the ones used for comparison analysis) are in green and bad quality samples in red. The dashed lines represent the coverage and depth cut-off values to consider a good quality sample, the horizontal line highlights a coverage = 0.95 and the vertical line highlights depth = 30X. Source data is provided as a Source Data file.

Table 1: Summary of sequencing results of the 61 paired samples.

	Total pairs = 61		
	eWGS	**dWGS**	**Culture WGS**
Total WGS	16	45	61
Median Depth	48.00	74.00	114.00
Q1	33.75	47.00	94.00
Q3	76.25	120.00	127.00
IQR	42.50	73.00	33.00
min depth	30.00	33.00	49.00
max depth	157.00	302.00	347.00
Median Coverage	0.96	0.98	0.98
Q1	0.96	0.97	0.98
Q3	0.97	0.98	0.98
IQR	0.01	0.01	0.01
min coverage	0.96	0.96	0.97
max coverage	0.98	0.99	0.98

Abbreviations: Q - quartile, IQR - Interquartile range, min - minimum, max - maximum

Tailored SNP calling pipeline for culture-free WGS analysis

To determine if culture-free sequencing approaches introduced a bias during the variant calling, we compared fixed and low-frequency variants detected in trios of eWGS, dWGS and culture-WGS from the same TB case in three suitable examples. After variant calling, we implemented a "recovery" step in which all

variants not passing filters in one sequencing approach, but accurately called in another approach, were re-included or "rescued" in the analysis. Our rationale was that culture-free WGS is performed in samples with low amounts of MTB DNA, hence increasing the chance of variants not passing the filters due to lower sequencing depths. Before applying our customised calling filters, we identified up to 22 variants only detected in eWGS, all of them at low frequencies (median=22%, range=10-54%). Exclusive variants were also detected in dWGS (up to 8) and in culture-WGS (up to 3) but the magnitude of the differences were minor compared to eWGS (**Figure 11a** Venn diagrams in blue). We screened manually the discrepant variants in the eWGS approach, and identified that all of them appeared in supplementary alignments (**Supplementary Figure 2**). A supplementary alignment is a read segment split from the primary read and aligned to a different region of the genome that can produce false positive calls in the variant calling step [126,130]. After discarding supplementary alignments, most discrepant SNPs disappeared giving a concordance of 99-100% in the variant calling between the three sequencing approaches (**Figure 11a**).

The extended analysis to all the 61 sputum-culture pairs demonstrated that overall the 88.5% (307/347) of the discrepant variants between sputa (either eWGS or dWGS) and paired cultures were false positives introduced by supplementary alignments, and these false positive variation was accumulated in 11/61 sputa (taking into account the sputum samples with more than 5 SNPs discarded due to supplementary alignments). Particularly, eWGS sputa accumulated significantly higher false positive variants than dWGS sputa and cultures (Wilcox test, p-value < 0.001. **Figure 11b**) likely due to a higher amount of chimeric reads causing supplementary alignments. After discarding them

from the bam files, we identified a mean of 18.3 false positive SNPs (range 0-106 SNPs corresponding to 1-7.4%) per eWGS- culture pair (**Figure 11c**, **Supplementary Figure 3**) whereas, dWGS-culture pairs showed a lower mean value of 0.3 (1-5 SNP corresponding to 0.3-5%) false positive SNPs (**Figure 11c**, **Supplementary Figure 3**). Notably, a high, positive and significant correlation (corr=0.95 [confidence interval: 0.85-0.98], p-value < 0.001) was obtained between supplementary alignments and false positive SNPs for eWGS, whereas, in the case of dWGS, the correlation was lower and with limited impact on variation (**Figure 11d**). Supplementary alignments accounted for only 1.2% of the reads on average (range: 0.2-7.5; dWGS: 1.0%, eWGS: 2.3%), therefore not impacting the depth and genome coverage (**Supplementary Table 1**).

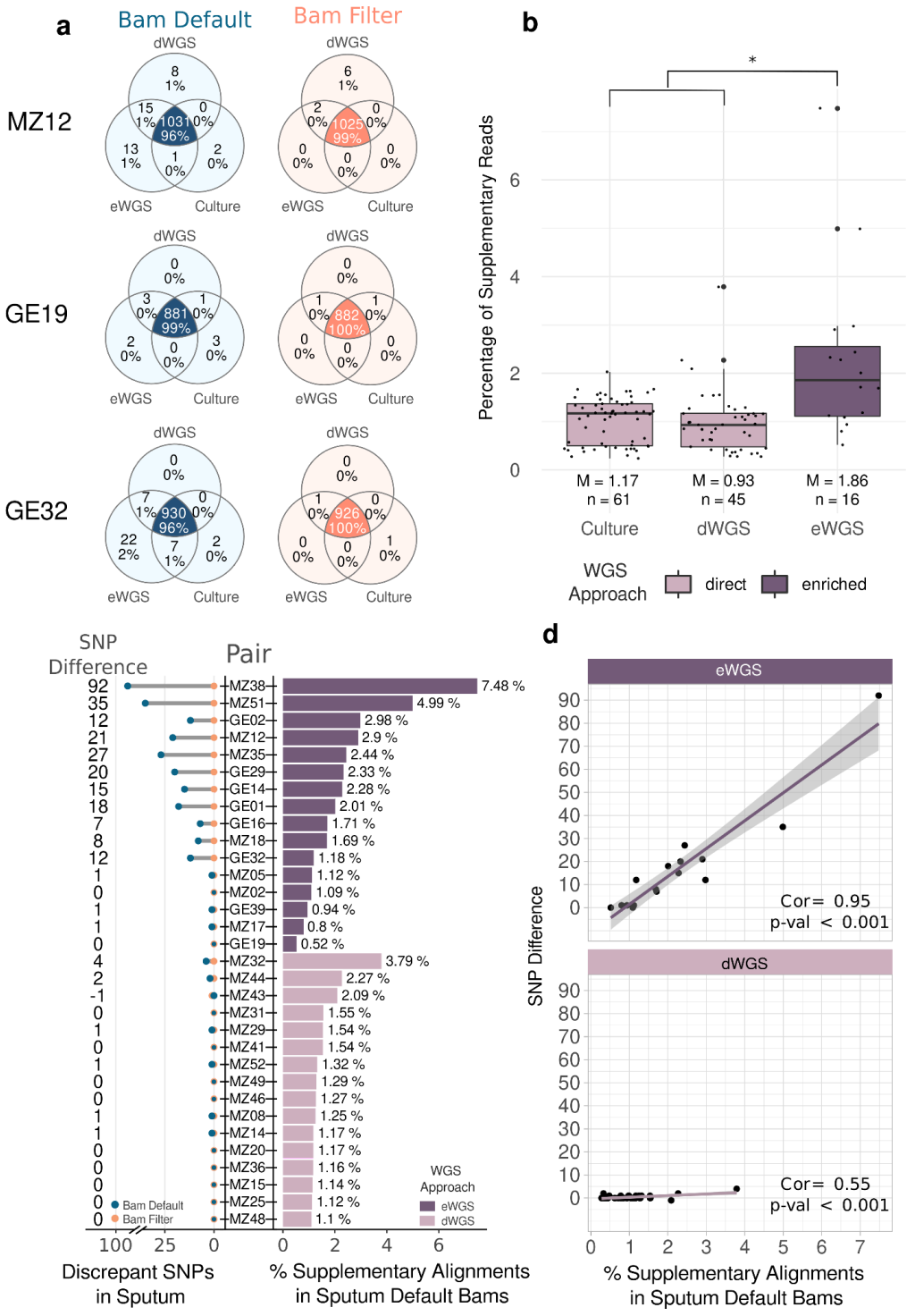

a
Bam Default
Bam Filter
MZ12
GE19
GE32
dWGS
eWGS
Culture
8 1% 15 1% 1031 96% 0 0% 13 1% 1 0% 2 0%
6 1% 2 0% 1025 99% 0 0% 0 0% 0 0% 0 0%
0 0% 3 0% 881 99% 1 0% 2 0% 0 0% 3 0%
0 0% 1 0% 882 100% 1 0% 0 0% 0 0% 0 0%
0 0% 7 1% 930 96% 0 0% 22 2% 7 1% 2 0%
0 0% 1 0% 926 100% 0 0% 0 0% 0 0% 1 0%
b
Percentage of Supplementary Reads
0
2
4
6
*
M = 1.17
n = 61
M = 0.93
n = 45
M = 1.86
n = 16
Culture
dWGS
eWGS
WGS Approach
direct
enriched
c
SNP Difference
Pair
92 MZ38 7.48 %
35 MZ51 4.99 %
12 GE02 2.98 %
21 MZ12 2.9 %
27 MZ35 2.44 %
20 GE29 2.33 %
15 GE14 2.28 %
18 GE01 2.01 %
7 GE16 1.71 %
8 MZ18 1.69 %
12 GE32 1.18 %
1 MZ05 1.12 %
0 MZ02 1.09 %
1 GE39 0.94 %
1 MZ17 0.8 %
0 GE19 0.52 %
4 MZ32 3.79 %
2 MZ44 2.27 %
-1 MZ43 2.09 %
0 MZ31 1.55 %
1 MZ29 1.54 %
0 MZ41 1.54 %
1 MZ52 1.32 %
0 MZ49 1.29 %
0 MZ46 1.27 %
1 MZ08 1.25 %
1 MZ14 1.17 %
0 MZ20 1.17 %
0 MZ36 1.16 %
0 MZ15 1.14 %
0 MZ25 1.12 %
0 MZ48 1.1 %
Bam Default
Bam Filter
WGS Approach
eWGS
dWGS
100
25
0
0
2
4
6
Discrepant SNPs in Sputum
% Supplementary Alignments in Sputum Default Bams
d
eWGS
dWGS
SNP Difference
Cor= 0.95
p-val < 0.001
Cor= 0.55
p-val < 0.001
0 1 2 3 4 5 6 7
% Supplementary Alignments in Sputum Default Bams

Figure 11: Analysis of supplementary alignments. **a,** Venn diagrams of the comparisons between trios of direct sputum (dWGS), enriched sputum (eWGS), culture WGS. Amount and percentage of exclusive and common variants are denoted. Blue and orange Venn diagrams represent comparisons of variant calls from default unfiltered bams (including supplementary alignments and filtered bams, respectively. **b,** Comparison of the amount of supplementary alignments between direct (sputum dWGS and culture WGS, light purple) and enriched sputum samples (eWGS, dark purple) in all 61 paired-samples. Median (M) and the total amount of samples (n) are shown. Asterisk (*) highlights a significant p-value (Wilcox test, p-value = 0.0002049). Data are presented as box-plots: centre line represents the median, upper bound located at 75th percentile, lower bound at 25th percentile, whiskers at minimum and maximum values and the outliers. Each dot represents one sample. **c,** On the left there is the comparison of the discrepant SNPs exclusive in sputum, either dWGS and eWGS, before and after filtering supplementary alignments. Colours stand for variant calls from bams before discarding supplementary alignments (blue) and after discarding them (orange). The x-axis is discontinued. The right part shows the percentage of supplementary alignments in sputum files, either dWGS (light purple) and eWGS (dark purple). Plot c contains 32/61 pairs, the 16 ones containing a higher percentage of supplementary alignments in each eWGS and dWGS. Samples are ordered from the highest to the lowest amount of supplementary alignments. The complete version containing the 61 pairs can be seen in **Supplementary Figure 3**. **d,** Correlation between the percentage of supplementary alignments and the amount of SNPs removed when discarding supplementary alignments from sputum bam files (represented as SNP difference and calculated as follows: discrepant SNPs exclusive in sputum in Default Bams - Filtered Bams). Colours represent whether the sputum samples

have been sequenced directly (light purple) or previously enriched (dark purple). Regression lines, Pearson correlation coefficients (one-side) (Corr) and p-values are shown in the plot. Source data are provided as a Source Data file.

Sputum-culture genetic diversity comparison

After applying our tailored SNP calling pipeline we evaluated the genetic diversity between sputum and culture. Overall, a mean of 913 (range 770-1063), representing 97% of the total SNPs detected, were common in sputum and culture. The SNPs rescued represent approx 3% of total SNPs, with a mean of 24 (range 5-69) (**Table 2**; **Figure 12a**; **Supplementary Figure 4**). We noticed that rescued SNPs can have an impact, added to that of/in addition to supplementary alignments, when evaluating the correlation between sputum-culture pairs and can affect both sputum and culture estimated diversity. The effect of both supplementary alignments and the rescue step in the total of discrepancies per sample is detailed in **Supplementary Figure 6**. Regarding exclusive variants, we also observed a small proportion of sputum-exclusive SNPs (mean 1, range 0-10) and culture-exclusive (mean 1, range 0-11). In other words, most of the sputum-culture pairs (49/61, 80.3%) did not display any discrepancies in sputum; 9/61 (14.8%) presented 1-5; and only 3/61 (4.9%) had between 5 and 10 sputum-exclusive SNPs (**Figure 12b**).

It could be argued that culture-free WGS will be more relevant for determining the frequency of SNPs rather than just their presence/absence. The correlation between the frequencies of SNPs observed in both sputum, either dWGS and eWGS, and culture (common SNPs) was 0.98 (0.93-0.99, p-value<0.001) (**Figure 12a**; **Supplementary Figure 5**). By going deeper, we observed that most of the

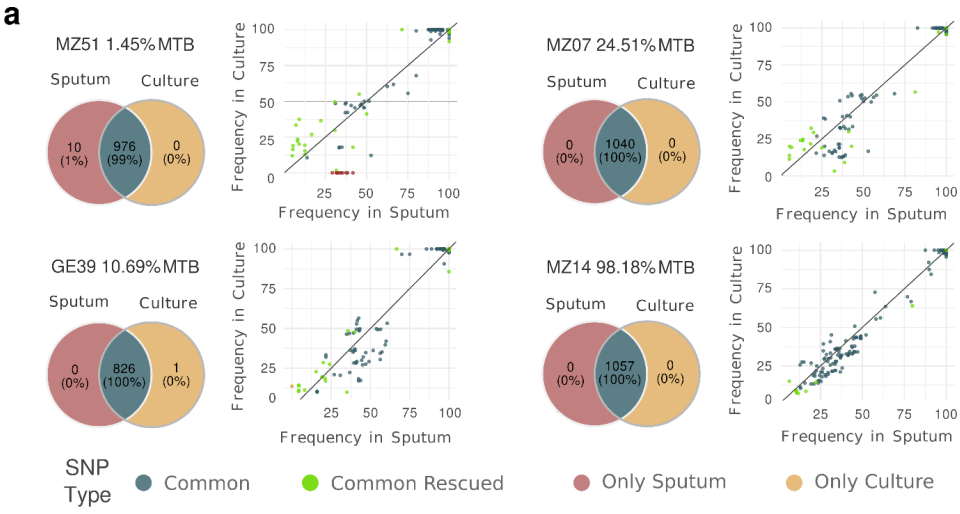
common SNPs (98%) had a small frequency difference within both paired samples (median = 0, ranging 0-10%) with higher frequencies in culture than in sputum (**Figure 12c**). To highlight, 83.4% (46445/55700) presented the same frequency in sputum and culture pairs. As expected, these differences were slightly higher for rescued SNPs with 70% of them ranging 0-10% frequency and 28% ranging 11-30% frequency (**Figure 12c**).

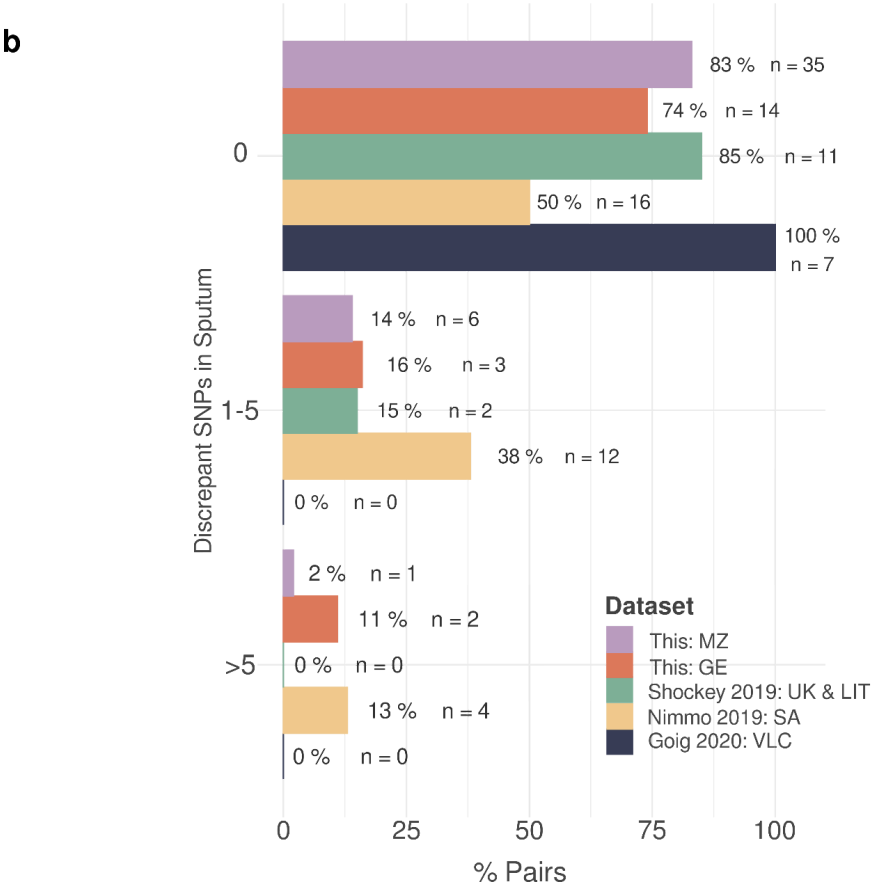
Regarding the exclusive variants, we observed a total of 40 sputum-exclusive SNPs distributed among 12 sputa, with 17 out of 40 (42%) falling within a frequency interval of 30-40%. For those culture-exclusives, which were present in 12 cultures, 73.8% (31/42) were found within a frequency range of 10-20% (**Figure 12c**).

The results for discrepant variants suggest that, in some cases, variants with intermediate frequencies (between 30-80%) in sputum can be missed in culture. In any case, the number of exclusive variants present only in sputum or in cultures were similar and low compared to those commonly called. To highlight, no fixed-SNP (>90% frequency) was detected within them.

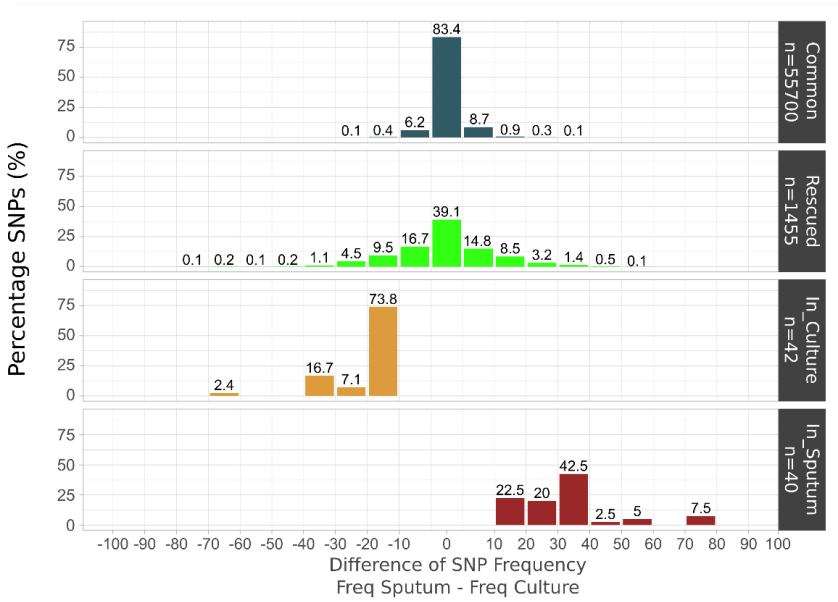
Table 2: Counts of common and discrepant variants in 61 pairs. Exclusive SNPs shown were obtained after the recovery step.

	Common	Rescued	Only Sputum	Only Culture
Mean	913	24	1	1
Median	889	21	0	0
Range	770-1063	5-69	0-10	0-11


a
MZ51 1.45% MTB
Sputum
Culture
10 (1%)
976 (99%)
0 (0%)
Frequency in Culture
Frequency in Sputum
MZ07 24.51% MTB
0 (0%)
1040 (100%)
0 (0%)
GE39 10.69% MTB
0 (0%)
826 (100%)
1 (0%)
MZ14 98.18% MTB
0 (0%)
1057 (100%)
0 (0%)
SNP Type
Common
Common Rescued
Only Sputum
Only Culture


b
Discrepant SNPs in Sputum
0
1-5
>5
83 % n = 35
74 % n = 14
85 % n = 11
50 % n = 16
100 % n = 7
14 % n = 6
16 % n = 3
15 % n = 2
38 % n = 12
0 % n = 0
2 % n = 1
11 % n = 2
0 % n = 0
13 % n = 4
0 % n = 0
Dataset
This: MZ
This: GE
Shockey 2019: UK & LIT
Nimmo 2019: SA
Goig 2020: VLC
% Pairs

c


Percentage SNPs (%)
Common n=55700
Rescued n=1455
In_Culture n=42
In_Sputum n=40
0.1 0.4 6.2 83.4 8.7 0.9 0.3 0.1
0.1 0.2 0.1 0.2 1.1 4.5 9.5 16.7 39.1 14.8 8.5 3.2 1.4 0.5 0.1
2.4 16.7 7.1 73.8
22.5 20 42.5 2.5 5 7.5
-100 -90 -80 -70 -60 -50 -40 -30 -20 -10 0 10 20 30 40 50 60 70 80 90 100
Difference of SNP Frequency
Freq Sputum - Freq Culture

d

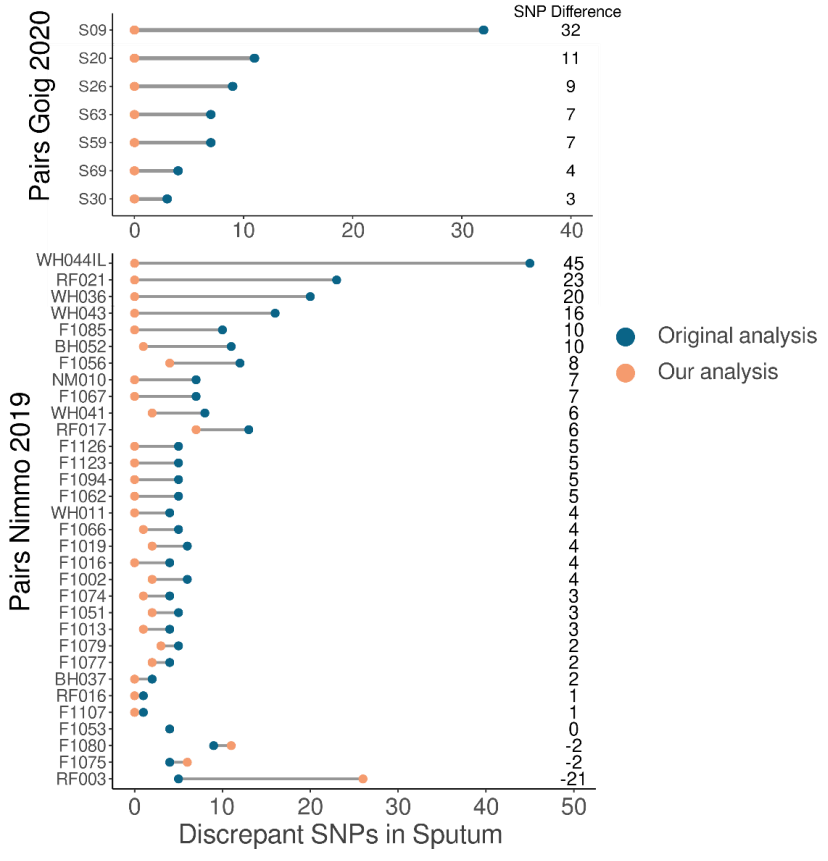

Pairs Goig 2020
SNP Difference
S09 32
S20 11
S26 9
S63 7
S59 7
S69 4
S30 3
0 10 20 30 40
Pairs Nimmo 2019
WH044IL 45
RF021 23
WH036 20
WH043 16
F1085 10
BH052 10
F1056 8
NM010 7
F1067 7
WH041 6
RF017 6
F1126 5
F1123 5
F1094 5
F1062 5
WH011 4
F1066 4
F1019 4
F1016 4
F1002 4
F1074 3
F1051 3
F1013 3
F1079 2
F1077 2
BH037 2
RF016 1
F1107 1
F1053 0
F1080 -2
F1075 -2
RF003 -21
0 10 20 30 40 50
Discrepant SNPs in Sputum
Original analysis
Our analysis

Figure 12: Comparison of variants between sputum-culture pairs. **a,** Comparison of the amount of common and exclusive variants (Venn diagrams on the left) and comparison of frequency of variants in sputum (dWGS or eWGS) and culture (on the right). Colours represent common or exclusive variants. Percentage of MTB reads in the dWGS is shown above each plot. The complete figures containing the 61 pairs can be seen in **Supplementary Figure 4** and **Supplementary Figure 5. b,** Analysis of sputum-exclusive variants in this dataset and published ones. Percentage of pairs in each dataset with 0, 1-5 or more than 5 sputum-exclusive SNPs (all were not fixed variants). Colours represent the dataset. Abbreviations of countries/regions stand for: MZ-Mozambique, GE-Georgia, UK-United Kingdom, LIT-Lituania, SA-South Africa, VLC-Valencia (Spain). c, Histogram of the difference of frequency (freq) between variants obtained in sputum versus culture (frequency in sputum - frequency in culture). Colours represent whether the variants are common or exclusive. Percentages of SNPs are shown on the top of the bars. The plot includes the percentage of SNPs that have a frequency difference equal to 0.Total number of variants is shown in grey boxes (n). d, Differences of sputum-exclusive SNPs published in the original paper (in blue) versus the ones found by running our pipeline (in orange) for Goig *et al.* 2020 [65] (upper panel) and Nimmo *et al.* 2019 [118] (bottom panel) datasets. Source data are provided as a Source Data file.

Generalisation of results in different datasets

Finally, we extended our analysis to publicly available sputum-culture sequencing data from three different clinical settings [65,80,118]. In those, authors described differences in genetic diversity between sputum and culture

pairs. We only analysed those sequences meeting the quality criteria (see Methods) for comparisons. After applying our pipeline, we observed that 88% of pairs differed in less than 5 SNPs (**Figure 12b**), all of them below 90% frequency. These results are very similar to the ones obtained for the dataset generated in this study (samples from Georgia and Mozambique; **Figure 12b**). For those studies with the number of discrepant SNPs available in the publication (Goig *et al.* 2020 [65] and Nimmo *et al.* 2019 [118]), our pipeline reduces the number of discrepant SNPs from an average of 10.4 SNPs (median 7, range 4-32) to 0 SNPs in all 7/7 sputum-culture pairs of Goig *et al.* [65] dataset; and from an average of 8.7 SNPs (median 5, range 1-45) to 1 SNPs (median: 0, range 0-7) in 28/32 pairs from Nimmo *et al.* [118] dataset (**Figure 12d**). Contrary to the Goig *et al.* [65] dataset, in three samples of Nimmo *et al.* [118] dataset, the number of SNPs did not decrease before and after applying our tailored pipeline, and there was even an increase in one of the pairs. For that sample (RF0003, **Figure 12d**) the number of discrepant SNPs in sputum is higher than for the rest of comparisons.

Overall, the results from our dataset suggest that almost all the genetic variability present in the original sputum is represented by culture both in terms of presence/absence of SNPs but also in terms of SNP frequency correlation within the samples. Re-analysis of available datasets also points to a high sputum-culture concordance after discarding false-positive variants, mostly due to the bioinformatic analysis implemented.

Phylogenetic classification and drug-resistance profile

We obtained a 100% concordance between sputum and culture lineage (L) prediction. We identified 31/61 L4 strains (20 from Mozambique and 11 from

Georgia) and 16/61 L2 (8 strains from each dataset). The remaining strains from Mozambique belonged to L1 (13/61 samples) and L3 (1/61 sample). Regarding the re-analysed datasets the lineage distribution was 1/52 L1, 12/52 L2, 2/52 L3 and 37/52 L4. Concordance at lower taxonomic levels according to Coll *et al.* classification [44] was also 100%. Sublineage classification is shown at **Supplementary Information 2**.

All the 61 sputum samples matched with their paired cultures when constructing a Maximum likelihood tree (**Supplementary Figure 1**) obtaining a 0 SNP in all pairs (fixed-SNP >90% frequency) pairwise distance. Two samples from Mozambique were in the same transmission cluster (measured at a 5 SNP distance cut-off) which was equally detected when culture and sputum genomes were analysed separately.

Regarding AMR-conferring variants, the agreement between the resistance profile predicted in sputum and culture was 100%. We obtained 10/61 pairs with at least one AMR-conferring mutation (6 from Georgia, 4 from Mozambique): 7/10 mono or poly-resistant, 2 MDR and 1 pre-XDR. All these were fixed-SNPs in both sputum and culture. No low-frequency AMR-conferring variants were found. See **Supplementary Information 3** for resistance SNPs information. Regarding the re-analysed datasets, we identified 31/52 strains with at least one AMR-conferring mutation in the sputum and the culture. All of them were fSNPs except one from Shockey's [65,80,118] dataset (P10) which was a heteroresistant variant present at 70% in both sputum and culture.

Discussion

In this study we explore the role of culturing from sputum diagnostic samples as a bottleneck for MTB genetic diversity. In terms of SNPs' presence/absence, both fixed and minority variants, after comparing the genomic variability, most culture-sputum pairs (91.8%, 56/61) showed a difference of less than 5 (0-3) SNPs. Strikingly, the correlation of the SNPs' frequencies was very high (0.98, p-value < 0.001). Our findings show that culture mirrors the genetic diversity present in the sputa analysed, considering both the presence/absence of SNPs and their frequencies. We corroborate our results by reanalyzing previously published datasets, suggesting that any apparent contradictions were likely due to differences in bioinformatic pipelines. Importantly, our findings were consistent across the datasets included, encompassing samples from different clinical settings and laboratories.

One key finding of our study was the detection of false-positive variation driven by chimeric reads most likely produced during sample enrichment and library preparation steps. While this is a common issue, we observed that they were significantly higher in the enriched sputa as compared to direct sputa sequenced. The presence of chimeric reads was potentially due to the low concentration of DNA in sputa. Considering that not all DNA was from MTB, the low quantity of target DNA could lead to non-specific amplifications in both library preparation and PCR reactions carried out during the enrichment. Our hypothesis is that such amplifications likely generate chimeric reads, resulting in supplementary alignments during the mapping and, therefore, false variant calls. Furthermore, a high correlation between supplementary alignments and false-positive SNPs was obtained for eWGS sputa, with special impact on those

samples with very low MTB DNA to be captured. Importantly, while removing supplementary alignments reduces by more than 80% the discrepant variants in some cases, it has no impact on depth and coverage. However, a more dedicated study of the factors affecting supplementary alignments is warranted as we identify some non-enriched samples with a high number of alignments.

Another source of discrepancies is driven by differences in depth throughout the genome, which is more relevant in culture-free WGS approaches. To solve this, here we implemented a recovery step to rescue the false negative variants present in the sputum or culture. It was observed that the rescued SNPs only make up less than 3% of the total SNPs analysed in the joint dataset. However, because differences between culture and sputum are subtle, they can make a great impact in the interpretation of sputum-culture differences in specific pairs. This phenomenon is more obvious when looking at specific examples like those presented in **Figure 12a**. It is worth noting that the absence of culture could have resulted in the oversight of these variants in the sputum, highlighting the critical role of achieving minimum sequencing depths for diversity comparisons. Overall, these findings unequivocally indicate that achieving an accurate representation of the original genetic diversity depends on sequencing depth, appropriate bioinformatic filters and careful post-processing of results.

The main aim of this study was to assess whether culturing distorts the genetic diversity of the bacterial population present in sputum. By performing an enrichment step, we succeed in increasing the MTB DNA by 26.5-fold on average, allowing high-quality sequences of sputa with as low as 1% of MTB DNA, similar to the values obtained by Mann *et al.* [134]. Since we have not

explored the potential of culture-free WGS as a diagnostic tool for TB, we performed as many runs as necessary to achieve sufficient coverage and depth to reach our goal, regardless of the sequencing cost per sample. With this effort we were able to perform culture-free sequencing, with high quality, in 76% (61/80) of diagnostic samples, suggesting that further improvements of the technique will be needed as shown by others [117]. In fact, while previous studies showed that smear-negative and scanty sputa could be enriched and sequenced occasionally [62,65,66,80], a formal testing of the limit of detection is still needed. Meanwhile, intermediate alternatives like targeted next-generation sequencing tools are finding a room in the TB diagnostics pipelines [135].

Understanding that the diversity of *M. tuberculosis* is reflected in both sputum and culture has an impact in TB research, since the overwhelming majority of studies today rely on culture sequencing to assess diversity, even in those cases where surgery [76] or post-mortem [57] samples are interrogated. Our analysis suggests that diversity of the diagnostic sample is well represented in those studies and that substantial variation is not missed in the matching culture. However, as a limitation, in this study we have not been able to analyse the detection of genotypes carrying AMR mutations (or any other mutation) with a fitness cost in culture. Most of our samples are drug susceptible and our analysis has identified only one case of heteroresistance for quinolones which is equally detected in sputum and culture. Thus, the high correlation observed in our sputum-culture pairs may not apply in situations where a subpopulation has a mutation with associated fitness cost. Mutations such as these are known to cause growth delay and affect culture-based diagnostics, particularly in the case of rifampicin-resistance conferring mutations occurring in *rpoB* [112,136].

Similarly, this limitation also applies to cases where lineages with different growth rates coexist, potentially affecting the identification of polyclonal infections, as it has been previously suggested for Lineage 6 and animal strains [113]. Identification of those scenarios requires specific analysis [76] out of the scope of this work. For some cases, particularly in the datasets from Georgia and South Africa [118], we identified a higher disagreement between sputum and culture. This may be due to the presence of unknown fitness-cost associated mutations or to polyclonal infections with genetically close strains, which are more likely to happen in those settings, particularly if there are high transmission hotspots [137]. Finally, a proportion of non-cultivable bacterial populations in the sputum of patients has been described in some studies, which may be relevant at the clinical level [138,139]. In the case that those populations are genotypically different (currently unknown) it may be reflected in the few SNPs and pairs with sputum exclusive SNPs or it may well be a phenomenon occurring below our limit of detection of 10% frequency. Those specific situations remain to be tested and here we provide tools for an accurate assessment of the sputum-culture diversity. Nevertheless, our dataset reflects the reality of many TB cases, which are from single infections and carry drug resistance mutations that do not cause growth problems in culture.

In conclusion, our results highlight the importance of evaluating and applying appropriate filtering steps when sequencing complex samples, such as sputum, in order to detect and discard sources of false variation. After developing and applying a tailored bioinformatics analysis, we show that culture accurately captures the genetic diversity present in diagnostic samples and this is true across settings and laboratories tested. More specific scenarios remain to be analysed and here we provide laboratory and bioinformatics tools to do so. On

one hand, from a diagnostic point of view our results reflect the long road ahead towards whole genome-based diagnostics, as highlighted in recent publications [117]. On the other hand, from a research point of view our results support the large body of work based on culture sequencing.

Data availability

Sequences generated in this paper have been deposited in the European Nucleotide Archive (ENA) under project accession number: PRJEB64897. Detailed information about the samples can be found at **Supplementary Information 1**.

We also downloaded the sequences under the following project accession numbers from ENA: PRJEB9206 (doi: https://doi.org/10.1128/jcm.00486-15) [66], PRJNA486713 (doi: https://doi.org/10.1186/s12864-019-5782-2) [118] and PRJEB37609 (doi: https://doi.org/10.1016/S2666-5247(20)30060-4) [65].

Source data are provided with this paper.

Code availability

The pipeline used for analysing sputum and culture sequences has been developed and validated in our laboratory and it is available at: https://gitlab.com/tbgenomicsunit/ThePipeline.

Authors' Contributions

A.G.B., B.S.C. and E.M. were responsible for data and samples collection in Mozambique. S.V., I.K., Z.A., A.R., A.G., M.S. and N.S. were responsible for data and samples collection in Georgia. C.M.L., M.T.P., L.V. processed and sequenced the samples. M.T.P., M.G. and I.C. conceived and designed the experiments. C. M.L., M. G. and I. C. were involved in data interpretation. C.M.L., G.G., M.G. and I.C. wrote the paper. All the authors reviewed and approved the manuscript for submission.

Acknowledgment

This work has been supported by the following: European Research Council (ERC): H2020-ERC-COG/0800; Ministerio Español de Ciencia e Innovación: PID2022-137607OB-I00; Fundació La Caixa: HR21-00415; Stop TB partnership (TB REACH): STBP/TBREACH/GSA/W5-30; and International Science and Technology Center (ISTC): Project #G-2143; National Institute of Allergy and Infectious Diseases (NIH).

We would like to thank Katharine Walter for reviewing the manuscript and providing her feedback.

Conflicts of interests

The authors declare no competing interests.

Chapter 2

Monitoring of first-line drug resistance mutations outside the scope of Xpert MTB/RIF Ultra is needed for successful control of DR-TB in Southern Mozambique

Carla Mariner-Llicer, Belén Saavedra-Cervera, Edson Mambuque, Neide Gomes, Shilzia Munguambe, Luis Villamayor, Irving Cancino-Muñoz, Manuela Torres-Puente, Dinis Nguenha, Durval Respeito, Gustavo Tembe, Mariana G. López, Iñaki Comas, Alberto L. García-Basteiro

Published in Clinical Infectious Diseases

DOI: 10.1093/cid/ciad684

Publication cite:

Mariner-Llicer C, Saavedra Cervera B, Mambuque E, Gomes N, Munguambe S, Villamayor L, Cancino-Muñoz I, Torres-Puente M, Nguenha D, Respeito D, Tembe G, López MG, Comas I, García-Basteiro AL. Monitoring of First-line Drug Resistance Mutations Outside the Scope of Xpert MTB/RIF Ultra is Needed for Successful Control of DR-TB in Southern Mozambique. Clin Infect Dis. 2023 Dec 4:ciad684. doi: 10.1093/cid/ciad684. Epub ahead of print. PMID: 38048599.

Abstract

Multidrug-resistant tuberculosis in Southern Africa is of great concern, exacerbated by the spread of a clone harbouring a mutation missed by Xpert Ultra. In Southern Mozambique, the presence of such mutation and rising cases of non-MDR isoniazid resistance highlights the need to ensure accurate detection of antimicrobial-resistance in the country. In this study we analyse 610 Mycobacterium tuberculosis genomes from two different drug resistance surveys conducted in Manhiça during 2014 and 2018 in order to determine which resistance mutations are circulating in this region and analyse if they are being detected by the frontline diagnostic techniques used there.

Introduction

Mozambique is among the 30 countries with the highest burden of multidrug-resistant tuberculosis (MDR-TB) according to the World Health Organisation (WHO). In 2021, 3.5% and 11% MDR-TB cases were reported among new and previously treated patients respectively, matching the last national estimates of 2008 [140]. Recently, they have implemented Xpert MTB/RIF Ultra (XpertUltra), a WHO-endorsed rapid method that identifies *Mycobacterium tuberculosis* complex (MTBC) and rifampicin-resistance (RIF-R) (WHO treatment guidelines, 2020). This assay detects well-characterised mutations in the RIF-R determining region (RRDR) of *rpoB*, a 81-bp "core region" where the most common mutations, responsible for 95-97% of RIF-R cases, are located [141]. Currently, the monitoring of RIF-R using XpertUltra has been challenged in southern Africa due to the spread of a MDR-TB clone harbouring a low-level resistance mutation, *rpoB*_I4941F, which is missed by

XpertUltra but also by phenotypic DST (pDST) [93,94] because it hampers the growth of MTB in culture. This mutation was reported in Eswatini for the first time in 2015, where it had been silently transmitted since 2009 causing a MDR-TB outbreak [93]. Later, two clones bearing the mentioned variant were detected in South Africa (2013-2016), one introduced from Eswatini and the other likely emerging independently [94]. Additionally, the MDR prevalence in both countries is high with 5% and 14% among new and retreated cases in Eswatini; and 4.1% and 28% in South Africa (WHO).

The latest antimicrobial resistance (AMR) prevalence estimates for southern Mozambique, obtained through pDST in 2014, showed a MDR-TB profile comparable to the latest national estimates (3.8% new and 13.2% retreated cases) [142]. However, these MDR prevalence rates may be likely underestimated, as some RIF-R mutations that can be missed by XpertUltra and undetected by pDST, as they cause bacterial growth retardation [112]. In this regard, a comprehensive screening for AMR-conferring mutations in this setting is essential to develop targeted public health strategies to improve diagnosis and treatment of drug-resistant TB (DR-TB).

Given the concerning MDR figures in Mozambique's bordering countries together with the high level of migration, particularly to South Africa, it is imperative to assess the presence and spread of *rpoB*_I491F and additional mutations missed by XpertUltra, including those associated to isoniazid, especially in the southern border region. The aim of this study is to screen the AMR-conferring mutations, focusing mainly on those located outside the RRDR, and to reassess the current AMR situation in Southern Mozambique.

Methods

Dataset and sequencing

This study was conducted in the district of Manhiça, Maputo province, a rural area with high TB/HIV burden and TB notification rate in southern Mozambique [143]. Manhiça is close to the South Africa and Eswatini borders and reports a high mobility rate with both countries [144]. The samples included in this study are from two different surveys conducted in the Centro de Investigação de Saúde de Manhiça (CISM): (i) 337 TB culture-positive from an index-contacts study (January-December, 2018). (ii) 275 publicly available sequences obtained from a prospective surveillance-based study (August 2013 to 2014) [61]. Furthermore, the CISM receives samples from several point-of-care centres throughout the province.

All isolates were whole-genome sequenced (WGS) by short-read sequencing approach (Illumina MiSeq) and analysed running our published and validated routine pipeline (https://gitlab.com/tbgenomicsunit/ThePipeline) for mapping and variant calling. Quality thresholds used for excluding sequences for downstream analysis were: median depth < 50X and genome coverage < 90%.

Drug resistance

We screen for all variants above 5% frequency to identify all the AMR-associated SNPs and indels listed in the WHO catalogue (2021 version) [104]. We took a closer look by performing an expert-knowledge analysis to find polymorphisms in candidate genes conferring resistance to bedaquiline (BDQ) (*atpE*, *Rv0678* and *pepQ*) and delamanid (DLM) (*ddn, fgd1, fbiA, fbiB, fbiC* and *fbiD*). To do that, we looked for all the mutations listed in Kadura *et al.*

2020 present in the strains [145]. We also searched for frameshift and insertion elements in the *Rv0678* gene as they have been associated to increase the value of the minimum inhibitory concentration (MIC) for BDQ as detailed in Roberts *et al.* 2022 [146]. Of note, the version of the WHO catalogue used (2021) lacked mutations conferring resistance to BDQ and clofazimine (CLZ).

Transmission analysis

Transmission events within Mozambican strains were defined as the number of cases in cluster minus one. They were measured by taking into account those "unmixed" clusters containing only resistant or only susceptible strains, as detailed by Walker *et al.* 2014 [147]. We calculated clustering by obtaining the pairwise distance within strains to identify recent (5 SNPs) local transmission [61] and quantify transmission events. First, we constructed an alignment of high confidence fixed-SNPs (fSNPs, >90% frequency), and second we obtained the transmission clusters.

The analysis of tuberculosis transmission within Mozambique and other 19 African countries included sequences from the neighbouring countries Eswatini, South Africa, Tanzania, Malawi, Zimbabwe and Botswana. See **Supplementary Table 4** for detailed information of the datasets. We downloaded all 10,000 available WGS African sequences from the ENA repository and analysed them by applying the same pipeline mentioned in the section *Dataset and sequencing*. Only those sequences meeting the quality criteria mentioned above were included in the downstream transmission analysis. We identify any potential epidemiological link (10 SNPs) (**Supplementary Table 4)** by following the same pipeline described above.

Data availability

Sequences generated in this study (study from 2018) were uploaded to the European Nucleotide Archive (ENA) under the accession number PRJEB61426. Sequences from 2014-study are available under the accession numbers: PRJEB27421, PRJEB50708 and PRJEB49677.

Accession study codes from the African datasets are detailed in **S2. Supplementary Material Table1**.

Results

Drug resistance mutations

We identified 78/612 (12.7%) strains containing at least one AMR-conferring mutation: 38 (6.2%) mono-resistant, 21 (3.4%) poly-resistant, 17 (2.8%) MDR-TB and 2 (0.3%) pre-extensively resistant (pre-XDR, according to 2021 WHO definition) (**Supplementary Table 4**). Regarding the two most common first-line drugs, we found 59 (9.6%) isoniazid resistant (INH-R) and 25 (4.1%) RIF-R strains. The most prevalent mutations were: *katG*_S315T in 31/59 (53%) INH-R strains; *inhA*_c-777t (*fabG1*_c-15t), conferring cross-resistance to INH and ethionamide (ETH), in 20/59 (34%) INH-R strains; *rpoB*_S450L in 15/25 (63%) RIF-R strains. We identified two additional RIF-R conferring mutations located outside the RRDR region and undetected by XpertUltra: *rpoB*_V170F (fSNP) in two strains from 2014; and *rpoB*_I491F (6.5% frequency) in one strain from 2018. These strains had no additional AMR mutations. Furthermore, we detected two

mutations in the RRDR region, usually missed by pDST [112,136]: *rpoB*_D435F (81% frequency) in one MDR-TB strain harbouring another variant (*rpoB*_S450L, fSNP); and *rpoB*_H445N at 6.6% frequency in one strain with no additional variants but fixed in another MDR-TB strain. Regarding second-line drugs, we detected two mutations (*gyrA*_D94G, *gyrA*_D94A) conferring resistance to levofloxacin and moxifloxacin in 2 pre-XDR strains. Importantly, we did not find variants in known genes associated with BDQ and DLM resistance, or truncations in *Rv0678* (**Table 3**).

Table 3: AMR-conferring mutations found in Mozambican strains (N=612).

AB	Strains with resistance mutations (%)	Gene	Mutation	2018	2014	Total
INH	38 (6.2)	*katG*	S315T	22	9	31
		katG	katG_1285_del_1_gc_g	0	3	3
		katG	katG_1433_del_1_gc_g	2	0	2
		katG	katG_1866_del_1_ag_a	0	1	1
INH, ETH	25 (4.1)	*IG_Rv1482c_Rv1483*	inhA_c-777t (fabG1_c-15t)	11	9	20
		fabG1-inhA	L203L	3	0	3
		IG_Rv1482c_Rv1483	inhA_t-770a (fabG1_t-8a)	0	2	2
RIF	25 (4.1)	*rpoB*	S450L	9	6	15
		rpoB	H445D	2	0	2
		rpoB	H445N	1	1	2
		rpoB	S450Y	2	0	2
		rpoB	V170F*	2	0	2
		rpoB	D435F**	1	0	1
		rpoB	H445Y	0	1	1
		rpoB	I491F*	0	1	1
EMB	14 (2.3)	*embB*	M306I	4	4	8
		embB	G406D	2	0	2
		embB	M306L	2	0	2
		embB	D354A	1	0	1
		embB	M306V	1	0	1
ETH	5 (0.8)	*ethA*	ethA_33_del_1_gc_g	2	0	2
		ethA	ethA_861_del_1_gt_g	2	0	2
		inhA	S94A	0	1	1
PZA	8 (1.3)	*pncA*	pncA_390_del_1_ca_c	0	2	2
		pncA	T135P	2	0	2

AB	Strains with resistance mutations (%)	Gene	Mutation	2018	2014	Total
		pncA	C14R	1	0	1
		pncA	L27P	0	1	1
		pncA	V130G***	0	1	1
		pncA	V139G	1	0	1
		pncA	Y34D	0	1	1
SM	25 (4.1)****	*rpsL*	K43R	8	2	10
		IG_Rv1315_Rv1316c	rrs_a1401g	0	5	5
		gid	gid_116_del_1_cg_c	4	0	4
		gid	gid_103_del_1_gc_g	0	3	3
		IG_Rv1315_Rv1316c	rrs_a514c	1	1	2
		gid	gid_352_del_1_gc_g	1	0	1
		gid	gid_472_del_1_gc_g	1	0	1
LEV, MXF	2 (0.3)	*gyrA*	D94A	1	0	1
		gyrA	D94G	1	0	1
AMI, CAP, KAN	1 (0.2)	*IG_Rv1315_Rv1316c*	rrs_a1401g	1	0	1
CAP	1 (0.2)	*tlyA*	Q22!	0	1	1

Abbreviations: AB: Antibiotic, AMI: amikacin, CAP: capreomycin, INH: isoniazid, RIF: rifampicin, ETH: ethionamide, SM: streptomycin, PZA: pyrazinamide, EMB: ethambutol, LEV: levofloxacin, MXF: moxifloxacin

***** represents mutations outside the RRDR region and, therefore, undetected by XpertUltra

****** rpoB_D435F appears in one strain with rpoB_S450L

******* pncA_V130G appears in the same strain with pncA_390_del_1_ca_c

******** one strain SM-R harbours two mutations: rrs_a1401g and rpsL_K43R

Local transmission

Regarding local transmission, we overall identified 190 putative transmission events among Mozambican cases, 28 of them occurring between AMR strains (14.7%, 28/190); this proportion was lower than transmission events occurring between susceptible strains (85.3%, 162/190) (odds-ratio=1.42; 95%CI:0.83-2.44; p-value=0.2). We also characterised transmission of the most prevalent mutations to first-line drugs: 19/31 strains with *katG*_S315T were within 6 different clusters; and 4/15 strains with *rpoB*_S450L were within 2 transmission clusters.

Cross-border transmission

Through the transmission analysis between Mozambique and other African countries, we identified 30 intercountry clusters involving at least one Mozambican strain and one, or more, from others. Clusters involved 52 Mozambican samples and 150 from surrounding countries, mainly South Africa (73) and Eswatini (40), but also from Botswana, Malawi and Zimbabwe; with a cluster size range of 2-35 samples. Importantly, 5/30 (16.7%) clusters presented a mean distance below 5 SNPs (range 1-5) suggesting very recent cross-border transmission.

Discussion

Here we presented a WGS screening of drug-resistance conferring mutations in Mozambique. The *rpoB*_I491F mutation was detected (not fixed) in one case from 2014, which was not in cluster with other samples. The strain did not present a polyclonal infection suggesting that the mutation arose *de novo.*

Furthermore, we detected two strains from 2018 harbouring *rpoB*_V170F mutation, previously reported in the country [148]. As we detected borderline resistance mutations in the RRDR, which are easily missed by pDST, it is likely that a proportion of RIF resistant patients are being misdiagnosed and receiving ineffective treatment. Therefore increasing the risk of transmission of AMR strains and negative outcomes as have been shown [136].

We have not detected a spread of *rpoB*_I491F from neighbouring countries to Southern Mozambique although it has been circulating widely in South Africa and Eswatini since 2009 [93,94]. However, we show evidence of transmission between countries for other strains even though our transmission analysis is limited in time and not nationwide. This, together with the high rate of cross-border mobility, alerts the need for increasing surveillance of AMR strains.

We found a prevalence of DR-TB of 2.8% (2018), slightly lower than the obtained in the last pDST-based survey (3.8%) conducted in Manhiça (2014) [142], suggesting a broadly stable situation compared to neighbouring countries [93,94]. We detected an increasing trend in new INH-R cases, 10.9%, compared to 2014 (9.2%) [142] driven by a higher percentage of INH-R non-MDR strains (7.6% in 2018 vs 5.5% in 2014). The results suggest that almost one in ten cases by 2018 were INH-R before starting treatment and undetected by XpertUltra. High-burden TB countries have limited tools to detect INH-R cases as they tend to be centralised in reference laboratories, diagnosing and treating them as pansusceptible [149]. Close monitoring of AMR is therefore essential to guarantee that the currently recommended standardised regimens for treating DR-TB continue to be effective.

A limitation of this analysis is that the estimates cannot be extrapolated to the whole country, as the strains included to identify AMR-conferring mutations were collected in Manhiça district. In addition, we cannot assess the correlation with pDST data. However the aim of the study was to identify circulation and prevalence of mutations rather than concordance with pDST. Opportunely, the accuracy of WGS for first-line AMR prediction shows a high sensitivity and specificity, (>90%), for RIF and INH making both techniques comparable [104].

Our results show the most recent AMR figures in southern Mozambique. While much attention must be paid to RIF-R, especially by expanding the AMR-targets beyond those implemented in XpertUltra, INH-R is now recognised as another major challenge for TB control and successful treatment outcomes [149]. The genomic evidence provided here, together with the MDR situation in the region, call for a thoughtful choice of surveillance tools and diagnostic algorithms. It is thus needed to assess the implementation of new molecular tests, such as Xpert MTB/XDR, low complexity automated nucleic acid amplification tests (NAAT) or targeted sequencing as rapid approaches for the detection of both INH and RIF-resistance, thereby simultaneously, interrogating a larger number of well-characterised mutations [104], and providing a more reliable drug-resistance profile.

Ethics

The study was approved by the National Bioethics Committee for Health of Mozambique (CNBS, Ref:369/CNBS/17) and the Internal Bioethics Committee of CISM. All methods were performed in accordance with the relevant guidelines and regulations. All study participants signed written informed consent after a

verbal explanation and written information about the study was provided. For participants under 18 years old, informed consent was obtained from their relatives, parents or guardians.

Authors' contribution

IC, AGB and MGL have contributed to the conception and the design of the work. EM, SM, NG, LV, ICM, MTP, DN, DDR and GT have processed the samples and acquired the epidemiological data that have been analysed by BSC and CML. Interpretation of results has been made by CML, MGL, IC, AGB and BSC. CML and MGL have drafted the manuscript and IC, BSC and AGB have contributed to its revision. All authors have agreed to final approval for submission.

Acknowledgment

This work was supported by STOP TB partnership [STBP/TBREACH/GSA/W5-30: CA-3-D000920001 to AGB], the TB Portals programme of the National Institutes of Health, the European Research Council [TB RECONNECT: H2020-ERC-COG/0800 to IC], Generalitat Valenciana [AICO/2018/113 to IC] and the Spanish Ministry of Science and Innovation [PID2019-104477RB-I00 to IC]. The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

We would like to express our gratitude to Stefan Niemann (Research Center Borstel Leibniz Lung Center, Borstel, Germany) and to Alex Rosenthal (Office of Cyber Infrastructure and Computational Biology, National Institutes of Health,

Rockville, Maryland, USA) from TB PORTALS for performing and funding the sequencing of the samples included in this study.

Conflicts of interest

IC received consultancy fees from Foundation for innovative new diagnostics for the development of the WHO mutation catalogue v1. The author has no other conflicts of interests to declare. The other authors declare no conflict of interest.

Chapter 3

Accuracy of an amplicon-sequencing nanopore approach to identify variants in tuberculosis drug-resistance-associated genes

Carla Mariner-Llicer, Galo A Goig, Laura Zaragoza-Infante, Manuela Torres-Puente, Luis Villamayor, David Navarro, Rafa Borràs, Álvaro Chiner-Oms, Iñaki Comas

Published in Microbial Genomics

DOI: 10.1099/mgen.0.000740

Publication cite:

Mariner-Llicer C, Goig GA, Zaragoza-Infante L, Torres-Puente M, Villamayor L, Navarro D, Borras R, Chiner-Oms Á, Comas I. Accuracy of an amplicon-sequencing nanopore approach to identify variants in tuberculosis drug-resistance-associated genes. Microb Genom. 2021 Dec;7(12):000740. doi: 10.1099/mgen.0.000740. PMID: 34919513; PMCID: PMC8767343.

Abstract

A rapid and accurate diagnostic assay represents an important means to detect *M. tuberculosis*, identify drug-resistant strains, and ensure treatment success. Currently employed techniques to diagnose drug-resistant tuberculosis include slow phenotypic tests or more rapid molecular assays that evaluate a limited range of drugs. Whole-genome sequencing-based approaches can detect known drug resistance-conferring mutations and novel variations; however, the dependence on growing samples in culture, and the associated delays in achieving results, represents a significant limitation. As an alternative, targeted sequencing strategies can be directly performed on clinical samples at high throughput. This study proposes a targeted sequencing assay to rapidly detect drug-resistant strains of *M. tuberculosis* using the Nanopore MinION sequencing platform. We designed a single-tube assay that targets nine genes associated with drug resistance to seven drugs and two phylogenetic-determining regions to determine strain lineage and tested it in nine clinical isolates and six sputa. The study's main aim is to calibrate MinNION variant calling to detect drug resistance-associated mutations with different frequencies to match the accuracy of Illumina (the current gold standard sequencing technology) from both culture and sputum samples. After calibrating Nanopore MinION variant calling we demonstrated 100% agreement between Illumina WGS and our MinION set up to detect known drug resistance and phylogenetic variants in our dataset. Importantly, other variants in the amplicons are also detected, decreasing the recall. We identify minority variants and insertions/deletions as crucial bioinformatics challenges to fully reproduce Illumina WGS results. Nanopore long-amplicon sequencing represents an under-used technology for diagnosing tuberculosis despite its

potential to predict drug resistance and genotype at almost real-time. This work describes a proof-of-concept study in which a single-tube PCR assay to identify drug resistance and phylogenetic determining mutations has been designed and evaluated with clinical cultures and diagnostic sputa. In addition, this study demonstrates the viability of the long-amplicon sequencing strategy regardless of the sample type. Our bioinformatics analyses establish a high degree of congruence to Illumina short-reads (as the gold-standard technique); however, we note the challenges associated with interpreting heteroresistance from Nanopore low-frequency variants.

Introduction

Tuberculosis (TB) remains a significant health problem worldwide and represents the principal cause of death by a single infectious agent. Approximately 10 million new cases and 1.5 million deaths were reported in 2019 [150]. TB has been successfully treated since the late 1940s thanks to the development of the first anti-tubercular drugs. However, the number of drug-resistant cases has increased 13 fold in the last decade, complicating the global control of TB and delaying its eradication [150–152].

Standard TB treatment consists of a combination of four different drugs administered daily for up to four months. Treatment of drug-resistant TB (DR-TB) represents a more complex task, comprising a combination of second-line antibiotics, which requires more extended treatment times and is more expensive [153,154]. Multidrug-resistant TB (MDR-TB) strains are resistant to both rifampicin (RIF) and isoniazid (INH), two key first-line drugs. MDR-TB is treatable by second-line drugs including oral and injectable drugs. Since 2019,

injectable drugs are being replaced by all-oral regimens following WHO guidelines [155]. Consequently MDR- and XDR-TB definitions are being revisited [156]. However, injectable drugs are still being used in many regimes and countries including those that have not yet approved bedaquiline, an oral drug, to treat MDR-TB cases [157].

The current reference technique for DR-TB diagnosis is based on phenotypic tests that evaluate the susceptibility of MTBC strains to different antibiotics (Drug Susceptibility Tests, DSTs). DSTs assess the ability of mycobacteria to grow in a solid or liquid medium containing antimicrobial drugs [84]. Despite being a commonly used technique, phenotypic diagnosis methods are laborious, time-consuming due to the slow growth rate of mycobacteria and require dedicated facilities [158]; furthermore, the results are not entirely reproducible in some cases [92]. Such limitations can be overcome with rapid molecular tests; however, these approaches target only a reduced number of genomic positions located in hot-spot regions for the most frequent drug resistance-associated mutations [100,159–162]. This limitation increases false-negative results due to silent mutations and other mutations located outside the hot-spot regions [163,164]. As an alternative, whole-genome sequencing (WGS) provides a maximum level of detail regarding the complete genomic sequence. Published catalogues of mutations robustly associated with drug-resistant phenotypes can be used to design targeted approaches against a wide number of regions involved in drug resistance. However, obtaining the whole genome sequence from sputum samples is challenging and expensive (see for example Galo *et al.* 2020 [65]).

Compared to WGS, targeted sequencing of drug resistance-associated genes represents a cheaper approach that can be routinely applied to sputum

samples. Two assays based on amplicon sequencing have been developed by Next Generation Sequencing (NGS) technologies – the Deeplex® Myc-TB assay (GenoScreen, Lille, France) [103] and the Next Gen-RDST assay (Translational Genomics Research Institute, Phoenix, Arizona, USA). Both assays target specific genomic regions known to be hot-spots for drug-resistance conferring variants. Due to its flexibility, custom panels can incorporate other target regions, such as those that allow *M. tuberculosis* genotyping. Importantly, strain type plays an important role in disease outcome, transmission, drug resistance emergence, host responses to the pathogen, disease epidemiology, and virulence [165,166]. Amplicon sequencing can be directly performed on patient samples, forgoing the need for long in-vitro culture times, and obtaining high depth of coverage [167]. In addition, an increasing body of evidence has suggested the role of heteroresistance to some drugs, in which drug resistance subpopulation undetected during in-vitro culture may impact treatment outcomes [168].

Nanopore (Oxford Technologies) released MinION, a portable device that allows real-time sequencing and analysis. MinION long-read sequencing is based on nanopores that act as biosensors, measuring the changes in voltage produced when a nucleotide passes through them [169]. This sequencing technology has been recently applied to the point-of-care diagnosis of viral and bacterial infections [170,171]. In MTBC research, MinION technology has been employed to assemble *M. tuberculosis* Beijing XDR strains and detect DR-TB in Madagascar [172,173]. The Nanopore MinION platform presents certain limitations, including the generation of noisy reads and a high error rate [174]. However, these limitations are balanced by the higher depth of coverage obtained compared to Illumina sequencing technology, which makes

free-culture sequencing and shorter diagnostic turnaround times possible [62].

This study aims to develop a single-tube PCR amplification protocol to sequence a panel of genes of interest using Nanopore technology and calibrate MinION variant calling parameters for prospective sequencing by comparing them to paired Illumina samples. We analyse recall and true negative rate to identify known drug resistance-associated variants in culture and sputum samples. We also analyse accuracy in identifying new variants and the source of errors associated with Nanopore MinION amplicon sequencing by comparing results with Illumina MiSeq WGS technology, the current gold standard technique.

Methods

Samples and DNA extraction

Fifteen samples were barcoded and sequenced in two MinION runs. In the first run, nine samples extracted from solid pure cultures were sequenced, including lineage control strains (L1-L6) and one sample extracted from smear-positive sputum (no smear grading was provided for sample 182320_M20). In the second run, five smear-positive sputa extracted directly from respiratory specimens of TB culture-positive patients were sequenced. Acid-fast bacilli (AFB) smear results were four 3+ sputa (G2267_W20, G2103_W23, G2284_W26, G1335_W27), and one 1+ sputum (G1646_W19). Cultures from all samples were previously whole-genome sequenced in our laboratory using the Illumina

MiSeq platform (**Table 5**).

DNA from culture samples was extracted following the standard CTAB protocol [122], while DNA from sputum samples was extracted following a differential cell lysis-based protocol to remove, as much as possible, all the contaminant non-MTBC DNA, according to Goig *et al.* [65]. All steps performed before library preparation were carried out in a BSL-3 laboratory.

Culture and sputum samples included in this study were derived from an ongoing surveillance program of communicable diseases by the General Directorate of Public Health of the Comunitat Valenciana (Spain); therefore, the samples fall outside the mandate of the corresponding Ethics Committee for Biomedical Research. All personal information was anonymized, and no data allowing individual identification were retained.

All sequences were uploaded to the ENA repository and are available under the accession numbers detailed in **Table 4**.

Table 4: Accession numbers of samples included in this study.

	Illumina		MinION	
Sample	**Study accession**	**RUN accession**	**Study accession**	**RUN accession**
G841	PRJEB22237	ERR2099789	PRJEB44496	ERS6293816
N0067	PRJEB3223	ERR233356	PRJEB44496	ERS6293819
G1335_W27	PRJEB37609	ERR3994128	PRJEB44496	ERS6293807
G1646_W19	PRJEB37609	ERR3994105	PRJEB44496	ERS6293808
G2103_W23	PRJEB37609	ERR3994108	PRJEB44496	ERS6293812
G2267_W20	PRJEB37609	ERR3994106	PRJEB44496	ERS6293813
G2284_W26	PRJEB37609	ERR3994111	PRJEB44496	ERS6293814
G107	PRJEB38719	ERR4195665	PRJEB44496	ERS6293806
G1800	PRJEB38719	ERR4195305	PRJEB44496	ERS6293809
G1961	PRJEB38719	ERR4195372	PRJEB44496	ERS6293811
G770	PRJEB38719	ERR4195357	PRJEB44496	ERS6293815
G870	PRJEB38719	ERR4195619	PRJEB44496	ERS6293817
G981	PRJEB38719	ERR4195376	PRJEB44496	ERS6293818
182320_M20	PRJEB44496	ERS6337728	PRJEB44496	ERS6293805
G1952	PRJEB44496	ERS6337729	PRJEB44496	ERS6293810

Table 5: Description of the dataset included in this study.

Sample	Sample Type	Run MinION	Resistance Profile	Resistance Agreement	Lineage	Lineage Agreement
182320_M20	Sputum	Run1	Susceptible	100	Lineage 4	100
G107	Culture	Run1	Susceptible	100	Lineage 3	100
G1800	Culture	Run1	FQ, RMP, INH, KAN, EMB	100	Lineage 2	100
G1952	Culture	Run1	Susceptible	100	Lineage 6	100
G1961	Culture	Run1	RMP, EMB	100	Lineage 5	100
G770	Culture	Run1	Susceptible	100	Lineage 4	100
G841	Culture	Run1	RMP	100	Lineage 4	100
G870	Culture	Run1	FQ, RMP, INH, KAN, EMB	100	Lineage 2	100
G981	Culture	Run1	Susceptible	100	Lineage 2	100
N0067	Culture	Run1	Susceptible	100	Lineage 1	100
G1335_W27	Sputum	Run2	Susceptible	100	Lineage 2	100
G1646_W19	Sputum	Run2	Susceptible	100	Lineage 4	100
G2103_W23	Sputum	Run2	Susceptible	100	Lineage 4	100
G2267_W20	Sputum	Run2	Susceptible	100	Lineage 4	100
G2284_W26	Sputum	Run2	Susceptible	100	Lineage 4	100

Abbreviations: FQ, fluoroquinolones; RMP, rifampicin; SM, streptomycin; INH, isoniazid; PZA, pyrazinamide; KAN, kanamycin; EMB, ethambutol

Primers' Design

The gene panel consisted of nine resistance genes and two phylogenetic determining regions. We selected nine genes known to be associated with first and second-line drug resistance in MTBC (*rpoB, inhA, katG, pncA, embB, rrs, gyrA, gyrB,* and *eis*) and two regions containing phylogenetic-determining SNPs (one region for lineages 1, 2, and 5 and another region for lineages 3, 4, and 6)

(**Table 6**; **Figure 13**) [175]. Primers were designed to amplify the entire genomic regions, obtaining products between 2,500 - 3,600 bp. Their specificity for MTBC was verified with Primer-BLAST (NCBI). The primer concentration for PCR reactions was optimised for an equimolar amplification of all genes. **Supplementary Table 6** contains all primer concentrations and sequences.

Table 6: Genes and phylogenetic regions included in the gene panel, their corresponding positions in the MTBC genome and the antibiotic they confer resistance to.

Amplicon MTBC genomic Positions		Gene	Prediction
Start	End		
4,193	7,279	*gyrB*	Fluoroquinolones
7,300	9,882	*gyrA*	Fluoroquinolones
759,779	763,387	*rpoB*	Rifampicin
1,471,177	1,474,319	*rrs*	Streptomycin
1,673,088	1,675,890	*inhA*	Isoniazid
2,153,843	2,156,356	*katG*	Isoniazid
2,287,571	2,290,686	*pncA*	Pyrazinamide
2,713,324	2,716,363	*eis*	Kanamycin
4,246,440	4,249,901	*embB*	Ethambutol
4,356,331	4,359,445	Rv3878 / Rv3879c	L1, L2, L5
1,279,773	1,282,976	Rv1155 / intergenic region	L3, L4, L6

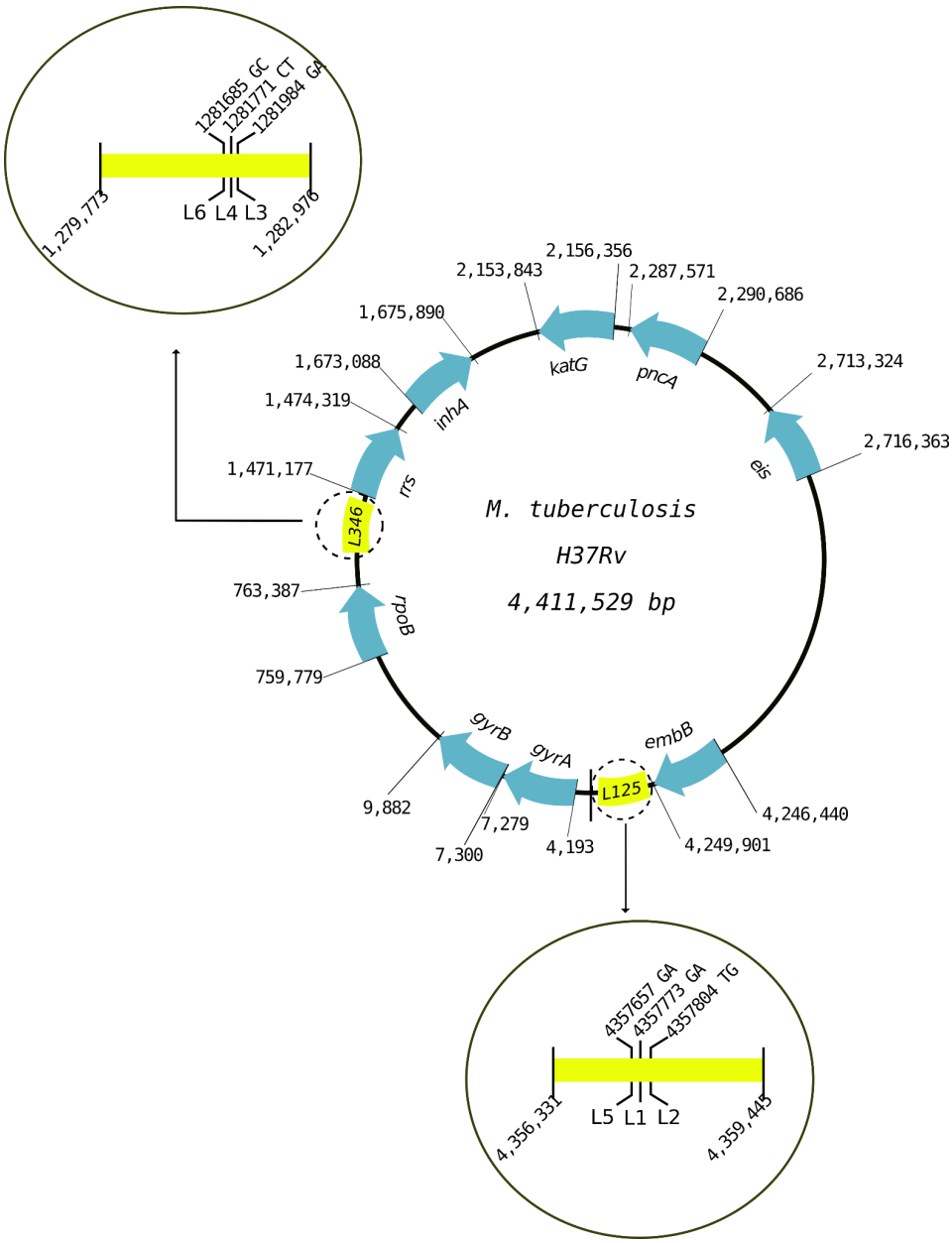


Figure 13: Diagram of the target regions of the gene panel. The lengths of regions are not scaled to genome size.

Single-tube multiplex PCR

A single-tube multiplex PCR reaction was performed for each sample using the Kapa HiFi HS polymerase kit (Kapa Biosystems®). The final reaction volume was adjusted depending on the DNA concentration of the samples to optimise yield (50 µl for samples above 10^{-3} ng μl^{-1}, and 100 µl for samples below 10^{-3} ng μl^{-1}). The optimum concentration of primers was previously set-up. PCR conditions were 3 minutes of DNA denaturation at 95°C followed by 30 cycles of amplification as follows: 20 seconds at 98°C for denaturing, 15 seconds at 65°C for primer annealing, 2 minutes at 72°C for extension, followed by 5 minutes at 72°C for a final extension. The PCR product was purified with AMPure XP magnetic beads at 0.6X volume (following the Agencourt AMPure XP purification protocol). The pure PCR product was eluted with 10 mM Tris (pH=8.5) and quantified with Qubit® to evaluate reaction yield. **Figure 14** summarises the steps of the workflow. See **Supplementary Methods 1** for reaction details.

MinION Library preparation and sequencing

Samples were barcoded by ligation using customised barcodes (See **Supplementary Methods 2**) and pooled in an equimolar manner. An additional PCR enrichment step after barcoding was performed in samples that failed to reach 500 ng of DNA after the first PCR (See **Supplementary Methods 3**).

The library was prepared using the 1D SQK-LSK108 Ligation Sequencing Kit (Oxford Nanopore technology) following the recommended protocol with modifications to increase process efficiency (See **Supplementary Methods 4**).

Analysis of nanopore reads

After sequencing, Fast5 files were base-called with Guppy version 2.3.5 (Oxford Nanopore Technologies) using the 'flip flop' model described by Wick *et al.* [176], which is slow but provides high read accuracy (See **Supplementary Methods 5** for the code). The resulting fastq file was demultiplexed with Porechop version 0.2.3 (https://github.com/rrwick/Porechop), generating a fastq file for each sample containing its corresponding reads (**Supplementary Methods 6**). The source code of Porechop was modified by substituting the default sequences of Oxford Nanopore barcodes with the sequences of our custom barcodes (**Supplementary Table 7**). Reads were mapped to the MTBC reference sequences of target genes included in the gene panel using minimap2 version 2.5-r607-dirty [177] with default parameters (**Supplementary Methods 7**) (Reference sequences were obtained from an inferred MTBC ancestral genome (NC_000962.3) [125]). Aligned reads with a map quality under 60 (MAPQ60) were discarded. The remaining BAM file was sorted using samtools [178] (See **Supplementary Methods 7** for the mapping code). Variant calling and indel (insertion/deletion) calling were performed by running VarScan v2.3.7 [179] (https://github.com/dkoboldt/varscan) (See **Supplementary Methods 8** for the variant calling parameters).

Illumina sequencing and analysis

DNA libraries for WGS were prepared according to the standardised NexteraXT® protocol and loaded into the Illumina MiSeq sequencer. Reads were analysed following our tested and publicly available analysis pipeline (https://gitlab.com/tbgenomicsunit/ThePipeline) for mapping and variant

calling (http://tgu.ibv.csic.es/?page_id=1794). This pipeline includes a first step to filter out low-quality and non-MTBC reads with fastp (https://github.com/OpenGene/fastp) and Kraken [74]. Filtered MTBC reads were aligned to an inferred MTBC ancestral genome [125] with BWA-MEM version 0.7.10 [180] and applied the MAPQ60 filter. Next, SNP calling and indel calling were performed with VarScan v2.3.7 [179] considering only positions covered at least by six reads, a minimum depth of coverage of 10, a minimum base quality of 25, and that present a minimum frequency of the variant allele above 0.1 (for variant positions, vSNPs) and 0.9 (for fixed variants, fSNPs). Additional information and details of the pipeline can be found at http://tgu.ibv.csic.es/wp-content/uploads/2020/03/ThePipeline_flowchart.pdf.

Resistance profiling and phylogenetic classification

The resistance profile of each strain was obtained using a customised script (Python 2.7.5) that searched for drug-conferring variants in the target genes of the panel using a catalogue of well-known variants obtained from two databases (PhyResSe (http://www.phyresse.org/) and ReSeqTB (https://platform.reseqtb.org/), 09/07/2019). In addition, lineage-defining variants (shown in **Supplementary Table 8** and described in a previously published study [175]) were searched to determine the major MTBC lineage of each strain. The customised script for resistance prediction and genotyping is available at GitLab https://gitlab.com/tbgenomicsunit/paneltb_scripts/.

Calibration of variant calling in MinION

A ROC (receiver operating characteristic curve) analysis was performed to determine the optimum variant allele frequency to reduce false-positive calls

by calling variants at different frequencies (from 0.1 to 0.9 using increments of 0.1) taking Illumina MiSeq variants (Illumina variants from now to the end) at 0.1 frequency as reference. These analyses focused on positions that pass all variant-calling thresholds, as described above in the case of Illumina. For Nanopore MinION variants (MinION variants from now), stricter filtering was applied to reduce false-positive calls that increase background noise and decrease the recall. Positions appearing in all fifteen samples sequenced by MinION were considered systematic errors and discarded. Triallelic positions with more than one nucleotide change were also discounted from downstream analysis. In addition, only positions that obtained a good depth value in both sequencing platforms were compared (positions above 10X depth in Illumina and 50X depth in MinION). Illumina was selected as the gold standard. Thus, true positive variants (TP) were defined as variants detected in both sequencing technologies, false-positive variants (FP) as variants detected in MinION but not in Illumina, false negative (FN) as missing variants in MinION, and true negatives (TN) as sites that contained the wild type allele in both technologies. According to these definitions, true positive rate (TPR) or recall, true negative rate (TNR), precision, agreement, error rate and F1-score were obtained (See **Supplementary Methods 9**)

Insertions/Deletions analysis

We performed an indel calling in MinION at different frequencies (from 0.1 to 1 with 0.1 increment) and validated with indels called by Illumina at 0.1 frequency. Two different comparisons were performed, one considering all indels regardless of the length and the other one filtering out indels longer than 50bp.

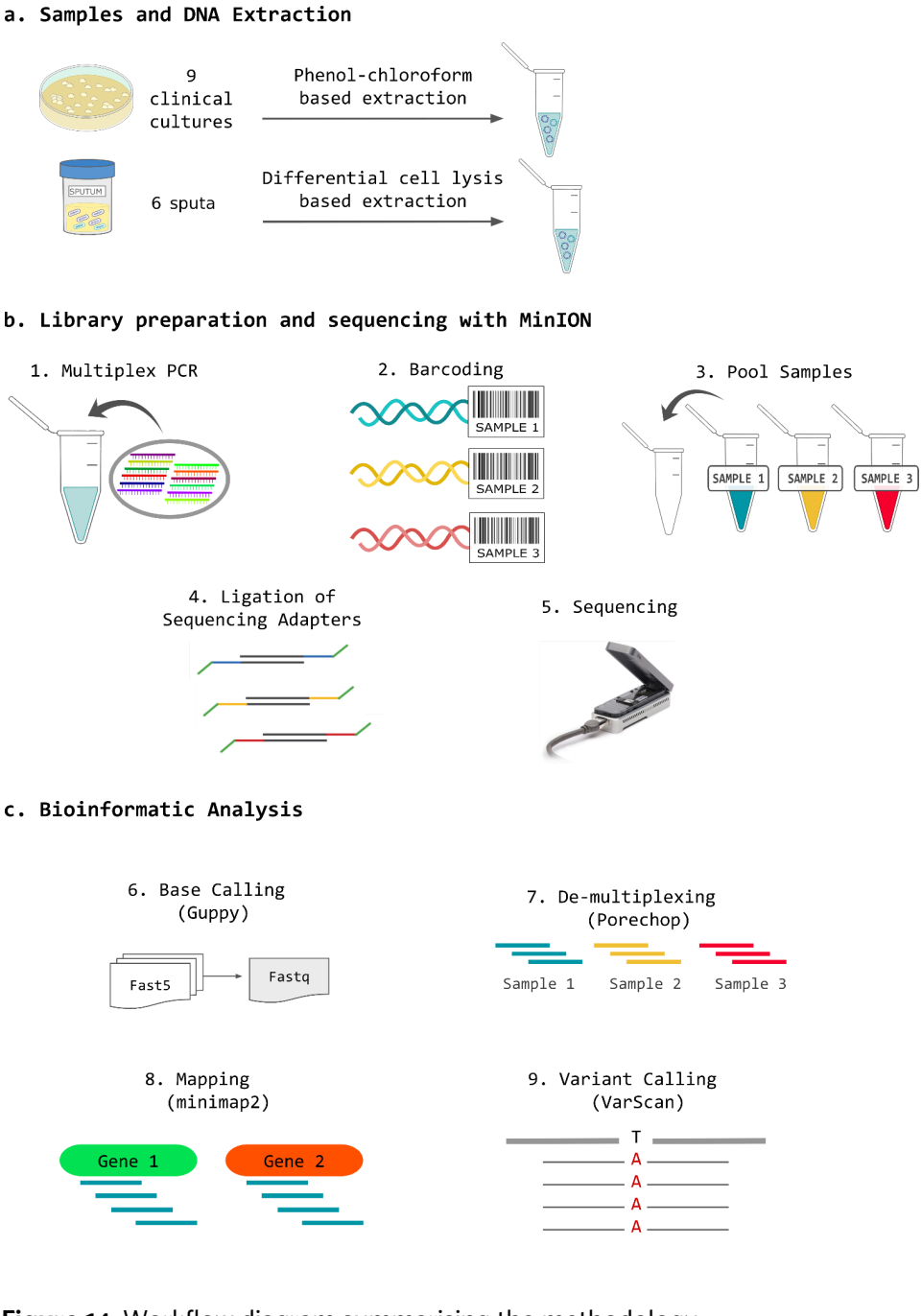


Figure 14: Workflow diagram summarising the methodology.

Results

Selection of samples, genes and qPCR set-up

We selected nine culture samples from an ongoing TB population study of the Comunitat Valenciana (Spain), while the Hospital Clínico Universitario de Valencia (Spain) provided six sputum samples. Samples were selected to maximise the number of MTBC lineages and antibiotic resistance profiles for this proof-of-concept study. All the fifteen samples were barcoded and then sequenced in two MinION runs. In the first run, we sequenced nine samples extracted from solid pure cultures, including lineage control strains (L1-L6) and one sample extracted from smear-positive sputum. In the second run, we sequenced five samples extracted directly from respiratory specimens of TB culture-positive patients. All fifteen samples were previously WGS by Illumina MiSeq.

The gene panel consisted of nine genes (*gyrA, gyrB, rpoB, rrs, inhA, embB, eis, pncA*, and *katG*) associated with resistance to first- and second-line drugs such as isoniazid, rifampicin, pyrazinamide, ethambutol, streptomycin, kanamycin, and fluoroquinolones; and two phylogenetic regions containing lineage determining SNPs (L1, L2, and L5 in the first region and L3, L4 and L6 in the second region) (**Figure 13**). We designed this panel to cover the whole region of interest and as a one-tube test. We carried out qPCR of the multiplex PCR product as a quality control step to assay the proportion of each amplicon before sequencing. Observed Cq values from 9.81 (L346 region) to 11.42 (*rpoB*) suggested that all target regions were amplified at the same ratio (**Figure 15**).

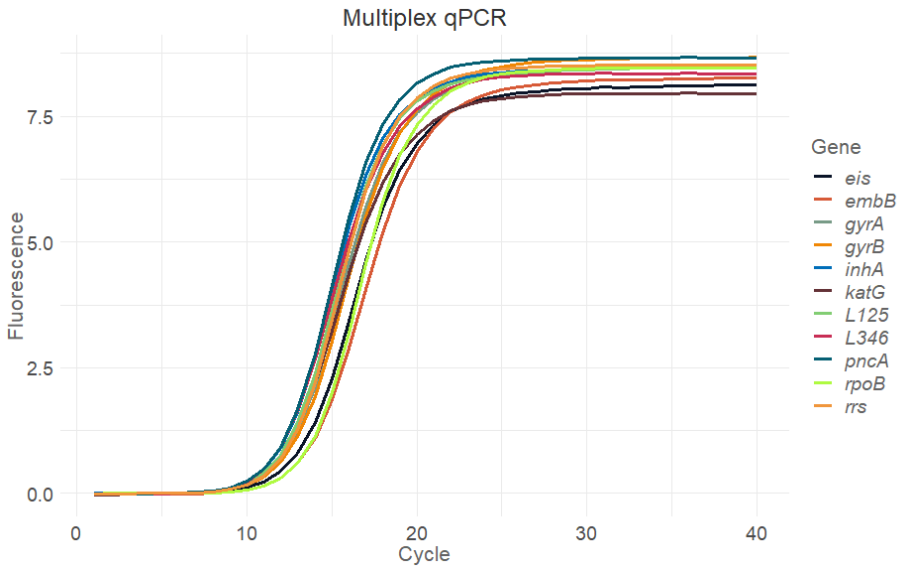


Figure 15: qPCR of the multiplex PCR product. Each colour represents a different gene. Two replicates per gene and negative controls were included.

Nanopore MinION sequencing

Samples were loaded in two Nanopore MinION R9 flow cells with 764 (first run) and 1375 (second run) active pores. In the first run, 848,583 reads passed base calling with an N50 read length of 3123 bp, median read length of 2941 bp, median read quality of 11.4, and median depth of 4194.5X per gene. In the second run, a total number of 623,989 reads passed base calling with an N50 read length of 3153 bp, median read length of 3102 bp, median read quality of 10.6, and median depth 1846X per gene. We hypothesise that the quick depletion of the library prompted the lower yield in the second run.

Porechop demultiplexed 65.9% and 61.6% of reads in the first and second runs, respectively; however, we failed to detect barcodes for around 22% of reads in

both runs, maybe due to problems in the barcode ligation step. In addition, 11.4% and 16.2% of reads in the first and second runs, respectively, could not be demultiplexed as they did not pass the default threshold used in Porechop (See demultiplexing parameters in **Supplementary Methods 6**).

Regarding the depth of coverage, it was higher in the first run, perhaps due to the sample type, the library, or flow cell differences. The first run included DNA extracted from cultures (only one sputum sample), while the second run included only sputum samples. The greater heterogeneity of sputum samples generally associates with lower DNA quality and DNA fragmentation. **Figure 16** depicts median depth coverage values for each gene per sample; we found that the *rpoB* and *rrs* genes had the lowest depth of coverage in both runs (**Table 7**).

Table 7: Median depth of coverage for each gene per MinION run.

Region	Run 1			Run 2		
	Median Depth	IQR	Min-Max	Median Depth	IQR	Min-Max
gyrA	6677	737.25	4213-7493	1719	373	231-1856
katG	6505	1561	4453-8510	2383	702	2056-3129
inhA	5802	1059.25	3916-7695	1479	410	863-1580
gyrB	6222	1590	1310-7489	3011	695	686-3058
embB	4564	801	2962-6318	3592	976.5	2833-4370
pncA	4358	1754.75	2086-6780	3888	960	587-4763
eis	2156	1135.75	1299-3985	1211	1067	152-1967
rpoB	2016	2011.75	1605-2643	84	78	3-118
rrs	2022	1088	852-4278	166	241	40-495
L125	2124	1597.5	994-3109	1520	1684	37-1896
L346	5222	1618	3844-7428	3593	3930.6	1860-3943

Abbreviations: IQR: interquartile range

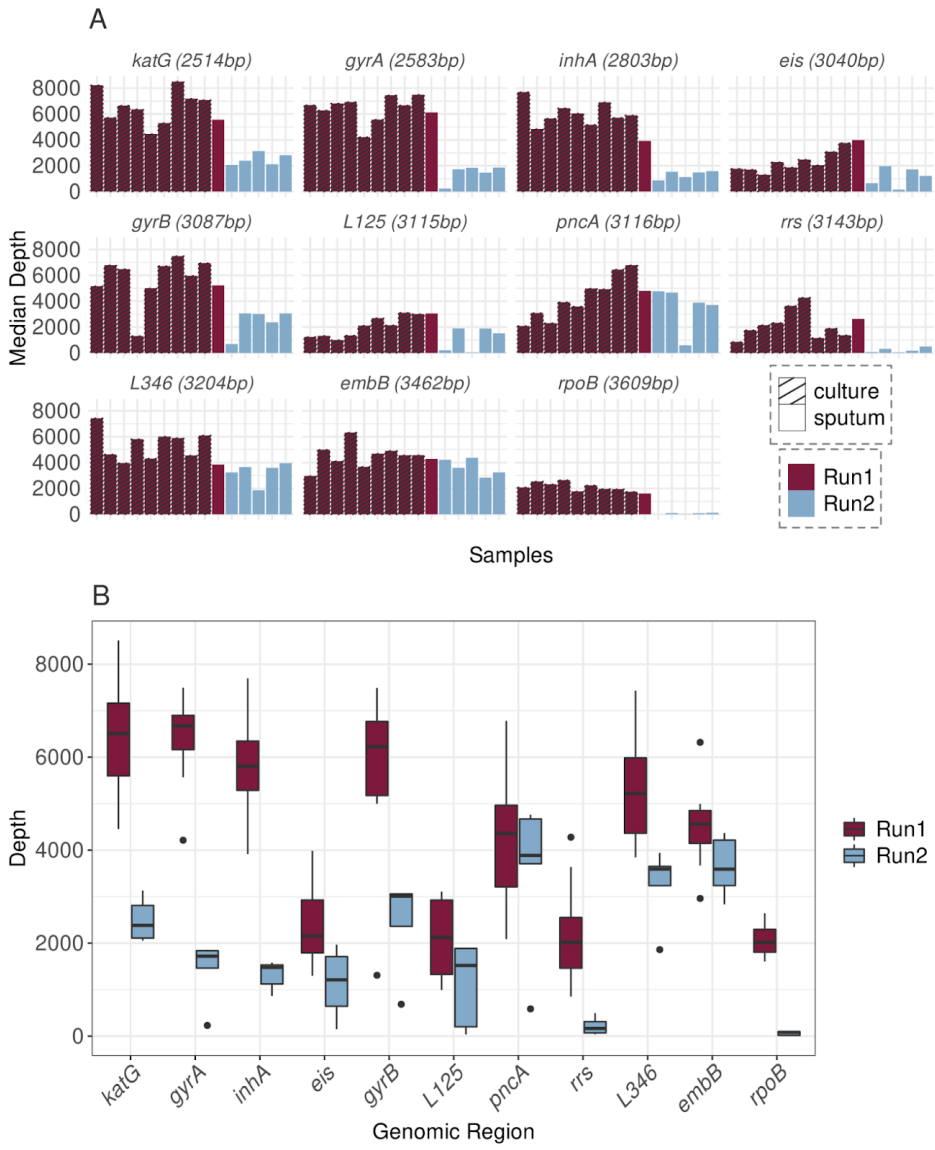


Figure 16: Depth of coverage by gene in each MinION sequencing run. Genes are ordered by length (from the shortest to the longest). **A)** Bar plot representing the median depth by sample and by region of the gene panel. Colours represent the sequencing run and the pattern of the bars represent sample type. **B)** Box plot representing the distribution of sequencing depths of

each genomic region in the fifteen samples. Plots are coloured according to the MinION sequencing run (purple for the first run, blue for the second run).

Calibrating MinION variant calling

A ROC analysis was performed to determine the frequency threshold for MinION Nanopore variant calling optimization, taking Illumina variants as a reference (those passing the filters and at least 0.1 frequency) to set the parameters for future applications. The nature of the vSNPs obtained in MinION and Illumina in all fifteen samples was compared after filtering (i) variants detected in all MinION samples (systematic errors), (ii) triallelic positions (more than one nucleotide change), and (iii) by depth threshold under 50X. While whole-genome sequences were available in the case of Illumina, only positions belonging to the gene panel regions were included in the comparison study. This analysis highlighted 0.4 as the optimum frequency threshold to obtain the highest TPR and TNR to call variants with >0.1 frequency in Illumina (**Figure 17A, Supplementary Table 9**). Based on these cut-offs, calls from MinION were classified as TP if they were detected in both sequencing platforms, FP if they were detected only in MinION reads, and FN if they were detected in Illumina reads and not in MinION. A total of 132 TP, 3 FP, and 14 FN were obtained (**Figure 18A)**, which resulted in a 90.41% TPR, 99.99% TNR, 88.59% agreement, 97.77% precision, 99.99% accuracy, 12.59% error rate, and a 0.94 F1-score. All FN positions were Illumina low-frequency variants missed in MinION. Globally, we found a lower variant frequency in MinION than Illumina (Wilcoxon test, p-value < 0.001) with a median of 4.44 [IQR = 4.51] change in frequency (**Figure 18B**). This difference explains the loss of low-frequency variants in MinION.

Regarding the identification of variants fixed in Illumina (>0.9 frequency) that are typically used for phylogenetic and epidemiological investigations, ROC analysis identified a cut-off value of 0.7 in MinION to maximise TPR and TNR (**Figure 17B**). A strong concordance between both platforms was obtained with an F1-score of 0.99, 98.47% agreement, 99.99% accuracy, 100% TPR, 99.99% TNR, and 1.53% error rate (**Table 5**). Of note, the 0.7 threshold in MinION involved the identification of two FPs.

Same ROC analysis was performed classifying SNPs by sample type for vSNPs and fSNPs obtaining the same cut-off frequencies mentioned above (See **Supplementary Table 12**, **Supplementary Table 13** and **Supplementary Figure 7)**.

Table 5: Comparison of common and discrepant fSNPs

Frequency in MinION	TP	FP	FN	TN	TPR	TNR
0.1	129	1489	0	478248	1	0.9968962
0.2	129	98	0	479592	1	0.9997957
0.5	129	4	0	479678	1	0.9999917
0.4	129	6	0	479676	1	0.9999875
0.3	129	16	0	479666	1	0.9999666
0.6	129	3	0	479679	1	0.9999937
0.7	**129**	**2**	**0**	**479680**	**1**	**0.9999958**
0.8	125	2	4	479680	0.96899	0.9999958
0.9	115	2	14	479679	0.89147	0.9999958

Abbreviations: TP, true positive variants; FN, false negative variants; FP, false positive variants; TN: true negative variants; TPR, true positive rate; TNR, true negative rate.

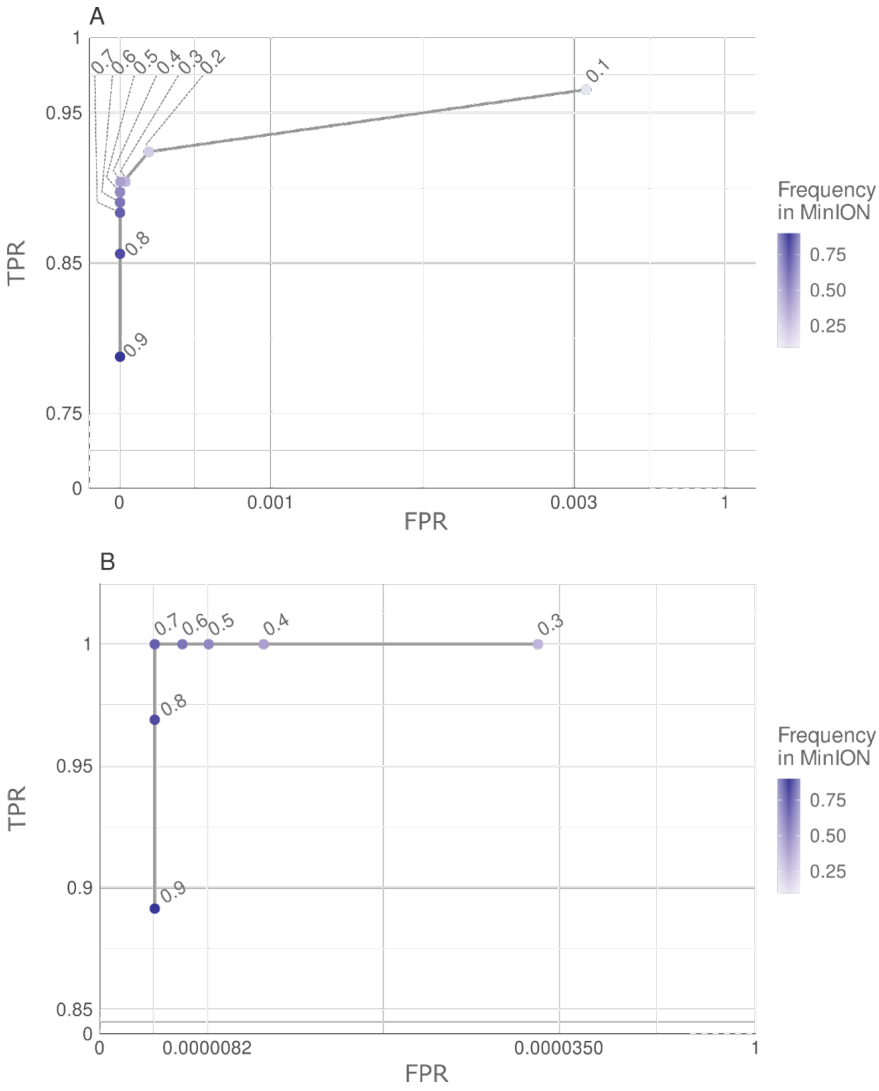


Figure 17: ROC curve used to set the frequency threshold employed to call variants in MinION**.** Points represent the values for recall and false positive rate obtained when applying different frequency values in MinION variant calling. Both axes are truncated. **A**) ROC curve used to set the frequency threshold to call variable SNPs in MinION. Recall and false positive rate value obtained using different variant calling frequency cut-offs for MinION (from 0.1 to 0.9 using

increments of 0.1) and comparing with Illumina variant calls at a 0.1 fixed threshold. **B**) ROC curve used to determine the frequency threshold to call fixed SNPs in MinION. Both axes are truncated. Recall and false positive rate value obtained using different variant calling frequency cut-offs for MinION (from 0.3 to 0.9 using increments of 0.1) and comparing with Illumina variant calls at a 0.9 fixed threshold.

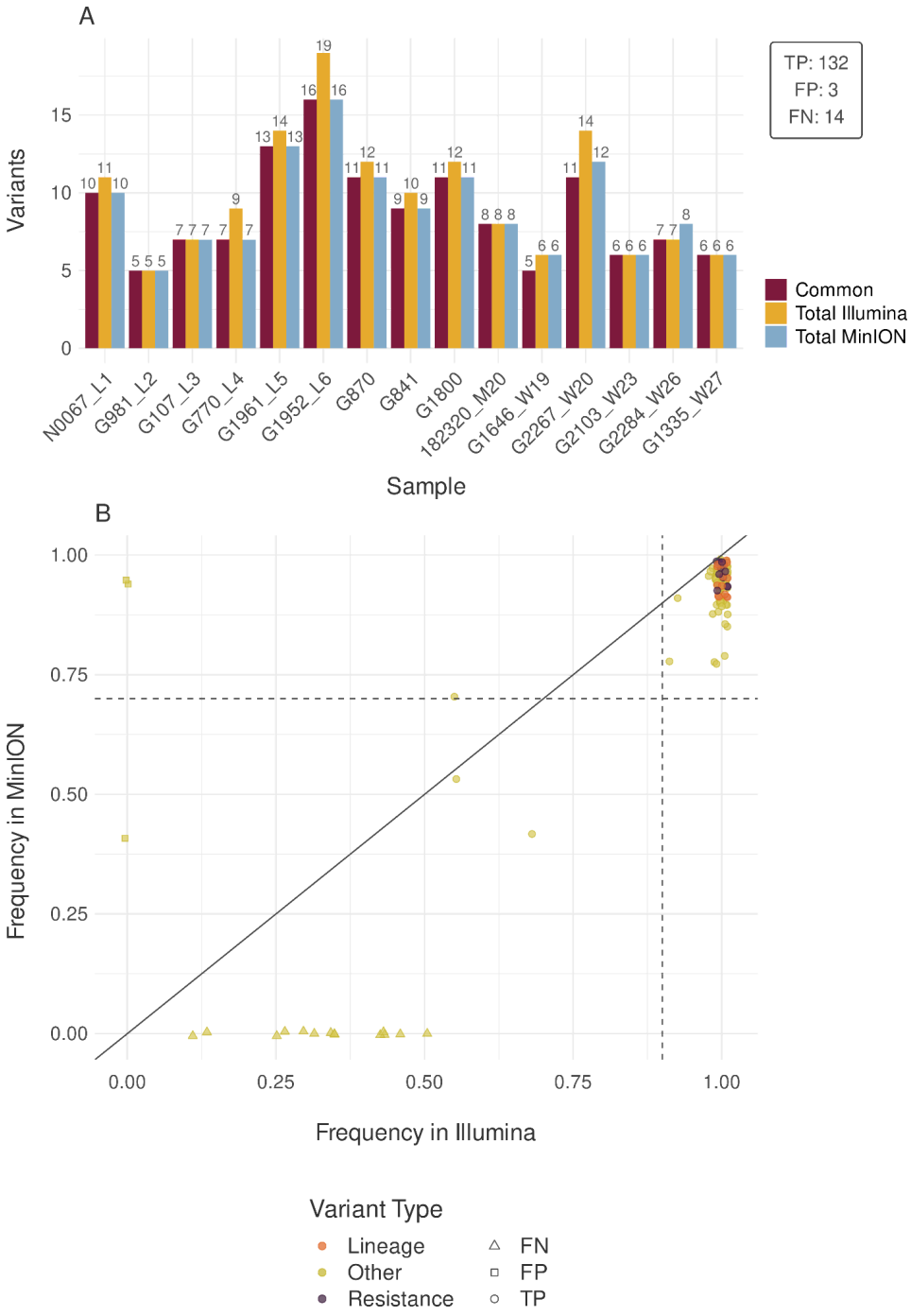


Figure 18: Comparison of variants found in Illumina and MinION. **A**) Number of

variants by sample obtained in MinION and Illumina that passed all quality filters for credible variants defined in methods. Blue bars represent all variants detected in MinION reads, yellow bars represent variants obtained in Illumina reads, and purple bars represent common variants detected by both sequencing platforms. TP, FP, and FN values in the inset box were obtained considering fifteen samples. **B**) Variant frequency correlation between Illumina and MinION reads. Only variants that passed all filters are represented. Points represent variants that appear in both platforms, squares represent discrepant variants only present in MinION reads, and triangles represent variants only present in Illumina. Orange represents phylogenetic variants, purple represents variants associated with antibiotic resistance, and yellow represents other variants. Dashed lines represent the threshold values to define fixed variants for MinION (horizontal) and Illumina (vertical).

Indels analysis

We tried to assess the capability of MinION to call insertions and deletions by calling indels at different frequencies (0.1 to 1 with 0.1 increments) and comparing the results with indels called by Illumina at 0.1. The total amount of indels detected by each sequencing platform is shown in **Table 6**. TP are considered as indels detected by both sequencing platforms. We observed a very low agreement in indel calling between both sequencing platforms suggesting that most of them could be due to sequencing or mapping errors. So that, indel calling still remains a limitation for drug-resistance prediction from MinION data.

Table 6: Indels comparison

Frequency in MinION	Indels MinION	Indels Illumina	TP
Freq 10	14413	16	6
Freq 20	3735	16	6
Freq 30	1159	16	5
Freq 40	356	16	5
Freq 50	81	16	5
Freq 60	49	16	5
Freq 70	19	16	5
Freq 80	17	16	4
Freq 90	1	16	1
Freq 100	0	16	0

Abbreviations: TP - true positive.

Drug resistance and lineage prediction

Regarding phylogenetic and drug resistance-associated variants, 100% of the variants found in Illumina reads were also called by MinION when applying the cut-off frequency obtained in the ROC analysis (0.4). All samples (lineage control samples, sputa, and cultures) could be classified correctly in their correspondent lineage by evaluating lineage-specific variants (**Supplementary Table 8**). From fifteen samples, we obtained one Lineage 1, four Lineage 2, one Lineage 3, seven Lineage 4, one Lineage 5, and one Lineage 6 strains (**Supplementary Table 10**). We determined resistance profiles by evaluating known variants in panel genes associated with drug-resistant phenotypes [181,182] resulting in one sample resistant to RMP (G841), one MDR-TB strain with heteroresistance to INH (G1961), two pre-XDR-TB strains (G1800, G870),

and eleven drug-susceptible strains (**Supplementary Table 10**). Phylogenetic and drug-resistance SNPs detected are shown in **Table 7**. Both lineage classifications and antibiotic resistance predictions by MinION agreed with the results obtained with Illumina WGS, when identifying known DR associated SNPs (0.4 frequency), giving a 100% agreement between both sequencing technologies (See **Supplementary Table 14**). However, the agreement decreased for some regions when examining all variants. For RMP, the agreement was 50% in sputa and 90% in cultures; for SM it was 60% in cultures; for KAN it was 75% in sputa; for L125 the agreement was 54% and 60% in sputa and culture respectively; and for L346 it was 60% in sputa (See **Supplementary Table 14**). These findings provide robust evidence that the gene panel sequenced using MinION can predict resistance profiles and strain lineages associated with known variants, but if novel variants are found additional validation will be required.

Table 7: Phylogenetic and drug resistance-associated variants detected by the gene panel

Sample	Variant	Gene	Antibiotic or Lineage	Frequency MinION	Frequency Illumina	Resistance Profile
G1961_L5	761155CT	*rpoB*	RMP	96.33	100	MDR
G1961_L5	1673425CT	*inhA*	INH	52.94	55.67	
G1961_L5	4247431GA	*embB*	EMB	89.12	100	
G1961_L5	4357657GA	L125	Lineage 5	95.37	100	
G1800	7570CT	*gyrA*	FQ	87.44	100	pre-XDR
G1800	761095TC	*rpoB*	RMP	92.72	100	
G1800	2155168CG	*katG*	INH	98.06	100	
G1800	2715344GA	*eis*	KAN	96.54	100	
G1800	4247429AG	*embB*	EMB	98.49	100	
G1800	4357804TG	L125	Lineage 2	98.17	100	
G870	6750CT	*gyrB*	FQ	95.26	100	pre-XDR
G870	761155CT	*rpoB*	RMP	95.76	99.17	
G870	2155168CG	*katG*	INH	98.11	100	
G870	2715346GA	*eis*	KAN	92.68	100	
G870	4247429AG	*embB*	EMB	98.28	100	
G870	4357804TG	L125	Lineage 2	98.49	100	
G841	761277AT	*rpoB*	RMP	98.67	100	RIF-R
G841	1281771CT	L346	Lineage 4	93.07	100	
N0067_L1	4357773GA	L125	Lineage 1	94.77	100	S
G1335_W27	4357804TG	L125	Lineage 2	97.43	100	S
G981_L2	4357804TG	L125	Lineage 2	98.4	100	S
G107_L3	1281984GA	L346	Lineage 3	97.62	100	S
182320_M20	1281771CT	L346	Lineage 4	92.89	100	S
G1646_W19	1281771CT	L346	Lineage 4	90.75	100	S
G2103_W23	1281771CT	L346	Lineage 4	89.69	100	S
G2267_W20	1281771CT	L346	Lineage 4	90.18	100	S

Sample	Variant	Gene	Antibiotic or Lineage	Frequency MinION	Frequency Illumina	Resistance Profile
G2284_W26	1281771CT	L346	Lineage 4	90.96	100	S
G770_L4	1281771CT	L346	Lineage 4	90.9	100	S
G1952_L6	1281685CG	L346	Lineage 6	84.47	100	S

Abbreviations: FQ, fluoroquinolones; RMP, rifampicin; SM, streptomycin; INH, isoniazid; PZA, pyrazinamide; KAN, kanamycin; EMB, ethambutol; S: Susceptible.

Discussion

The rapid and accurate detection of drug-resistant mycobacterial strains represents a crucial means to ensure treatment effectiveness; however, conventional diagnostic techniques are often time-intensive due to the requirement for *in vitro* culture steps. Amplicon-based approaches coupled with MinION sequencing are a promising strategy towards culture-free characterization of TB bacilli. Amplicons can detect resistant strains in culture-free samples with very low limit-of-detection meaning that drug resistance results can be obtained in around a week which is faster than culture-dependent strategies that can last up to a month. In addition, Illumina MiSeq takes around 72 hours to produce 20GB of output while MinION takes between 24 and 48 hours and the main advantage is that the user can stop it when the output is enough for downstream analysis, as it allows real-time base calling. Here we design an amplicon-based approach aimed to take advantage of MinION long reads. Current targeted amplicon sequencing techniques to identify DR-TB are based on short-read NGS technologies (i.e., Deeplex Myc-TB), which mainly focus on small amplicons (around 300 bp) within

regions containing known drug resistance-associated mutations [103]. Low DNA concentrations and high levels of fragmentation drive the need for small amplicons, particularly in complex samples, but this precludes certain advantages associated with long-read sequencing technologies like Nanopore. MinION long-read technology can sequence entire genes for variant phasing when genomic DNA has high integrity. Notably, the high error rate (as the main associated limitation) has decreased with time [183–185] through improvements in base-calling and variant-calling steps [176].

We present a long-amplicon alternative for research applications and demonstrate its utility using diagnostic samples. The strategy proposed in this study involves sequencing entire genes (around 2000 – 3000 bp) which is an advantage to identify all variants in each genomic region but it could compromise the efficiency of the multiplex PCR for some amplicons. We observed differences in depth of coverage between the genes included in our panel being *rpoB* and *rrs* the ones with the lowest depth value. Our data suggest that the length of the amplicon could represent an important bias during multiplex PCR amplification as shorter fragments are preferentially amplified (as seen in **Figure 16** and reported by Tafess *et al.*) [186]. In addition, we can not discard that sample type and between-run differences could also be responsible for the depth inconsistency. In any case, our results showed that the sequencing depth obtained was enough to detect drug resistance associated variants.

For clinical diagnosis, calling susceptibility is as important as calling resistance. This study shows the potential of a long-amplicon panel in predicting susceptibility and drug-resistance from both culture and sputum samples when scanning known markers. We designed a gene panel targeting nine genes

(*gyrA, gyrB, rpoB, rrs, inhA, embB, eis, pncA,* and *katG*) associated with resistance to commonly-used first-line drugs (isoniazid, rifampicin, pyrazinamide, and ethambutol) and injectable anti-TB drugs (streptomycin, kanamycin, and fluoroquinolones). Since 2019, the WHO has updated the treatment for MDR-TB patients by recommending all-oral regimens including new drugs like bedaquiline over injectable drugs and also, some antibiotics, such as streptomycin, fell into disuse in some countries. In this assay, streptomycin was included for historical reasons, but we have to emphasise the flexibility of the amplicon approaches to include or drop amplicons for different regions both for legacy drugs and new drugs by re-setting up the conditions of the multiplex PCR. The major obstacle however, is that many clinically relevant regions for new drugs remain unknown [187].

Beyond the limit of detection of our design, which would require a large number of samples, in this proof-of-concept study we focused on the bioinformatic challenges to identify variation, especially SNPs, associated with drug resistance, when using the Nanopore MinION sequencing technology. Reducing the high number of FP variant calls required a high cut-off frequency (0.4) compared with Illumina (0.1), which compromised recall in our study. Other studies have also shown 100% of agreement in the prediction of drug resistance-associated variants between Illumina and MinION Nanopore when variants detected by MinION at an allele frequency below 0.4 are excluded from the analysis [186,188]. Similarly, two recent studies showed complete concordance between both sequencing platforms when applied a threshold 0.8 to call fixed variants [102,189]. Accurate identification of low-frequency variants is relevant for clinical management [92,190] especially for drugs such as fluoroquinolones, for which frequencies ranging from 14% to 38% have

been reported [191].

Thus, the identification of heteroresistance below 0.4 remains challenging for MinION. MinION is also currently limited to call insertions/deletions due to the high error rate. Insertions/deletions is a type of variation that has been associated with resistance to different drugs. Our analysis shows the limited correlation between indel calls in MinION compared to WGS. Fortunately, Nanopore has recently released new flow cells with an improved chemistry that increases the variant calling accuracy and, therefore, the resolution in low-frequency variants studies [192]. Our proof-of-concept study demonstrates the power of single-molecule sequencing to determine the resistance profile and lineage regardless of sample type and points out the need to overcome variant calling challenges to fully take advantage of its potential for the point-of-care prediction of drug resistance (especially heteroresistance and insertion/deletions).

Ethics

Culture and sputum samples included in this study came from an ongoing surveillance programme of communicable diseases by the General Directorate of Public Health of the Comunitat Valenciana (Spain), and therefore fall outside the mandate of the corresponding Ethics Committee for Biomedical Research. All personal information was anonymised, and no data allowing individual identification was retained.

Authors' contribution

Author contributions are described with CRedit taxonomy. Conceptualization:

IC. Investigation: MTP, GAG, LMV, LZI, CML. Methodology: MTP, GAG, LMV. Resources: MTP, LMV, RB, DN, IC. Software: GAG, CML. Formal analysis: LZI, CML. Data curation: CML. Supervision: ACO, IC. Funding acquisition: IC. Validation: IC, MTP. Visualisation: CML, ACO, IC. Writing original draft: CML, ACO, IC. Writing - reviewing & editing: MTP, ACO, IC. Project administration: IC.

Acknowledgment

This project has been funded by the European Research Council (638553-TB-ACCELERATE, 101001038-TB-RECONNECT) and the Ministerio de Economía, Indústria y Competividad (SAF2016-77346-R, PID2019-104477RB-I00).

Conflicts of interest

The authors declare that there are no conflicts of interest.

General discussion

Mycobacterium tuberculosis (MTB) was discovered in 1882 as the causative agent of one of the top killer infectious diseases, Tuberculosis (TB). After almost 150 years, many efforts have been made to eliminate this disease successfully reducing its incidence in high-income countries [193]. However, there is still work to be done specially in low- and middle-income countries, places where TB represents a big challenge for public health. Despite advancements in molecular techniques, and even when culture produces a significant diagnostic delay, it remains the gold standard for TB diagnosis in many laboratories and primary care centres due mainly to the low costs and technical support required. In addition, almost all the MTB genomic publications have relied on cultures to obtain sufficient bacteria, even those evaluating the diversity in patient granulomas or post-mortem samples [57,76]. However, the impact of culturing on the genetic diversity of diagnostic samples has not been evaluated. Commonly, it has been assumed a bottleneck effect of culturing biassing the growth of those bacterial populations that are less suitable to grow under *in vitro* conditions *[80,118]*. If true, it could also bias the interpretation of all the genomic studies as well as diagnostics. Regarding research studies relying on culture, if only a subset of the total population were able to grow, results would not reflect the overall diversity present in the diagnostic samples which can impact evolutionary inferences with consequences in epidemiological analysis. In addition, in terms of diagnosis, losing a part of the bacterial population could lead to misdiagnosis of resistant strains.

In Chapter 1 of this dissertation we assess this issue and demonstrate that all the genetic diversity present in sputum samples is well mirrored in culture samples. This is particularly true for 61 paired samples from two different datasets originated for this thesis, Mozambique and Georgia, as well as for three additional datasets, already published; two of them originally showed higher diversity in sputa than cultures [65,80,118]. Our findings are also supported by other publications. Votintseva *et al. [62]* and Brown *et al. [66]* also reported no differences between sputum and culture pairs. It is important to notice that they only focused on developing a rapid culture-free resistance profiling approach, and they did not evaluate the whole genetic diversity [62,66]. Furthermore, in our study we detected a significant source of false-positive variation mainly present in enriched-sputa samples. This was driven by supplementary alignments generated due to chimeric reads produced during the PCR amplification of the enrichment and library preparation steps. The low amount of MTB target DNA present in sputa may force the baits to bind nonspecifically and generate these undesired products that can hamper the downstream bioinformatic analysis. We were able to detect them by comparing direct and enriched sputa and realised that they were responsible for the 88.5% of false-positive variants. Importantly, after discarding them, the mean depth and coverage were not affected. For the previously published datasets, we compared the variant calling published versus calling done with our tailored pipeline. We observed no differences in sputum-culture comparisons. Furthermore, when applying our tailored pipeline, we obtained fewer discrepant SNPs in almost all the tested pairs compared to the original publication results. Overall, the comparison of variants between sputum and culture pairs showed less than a 5 non-fixed SNP difference (below 90% frequency), suggesting that the contradictory findings regarding genetic

diversity may be due to differences in the bioinformatic pipelines applied. As previously demonstrated for culture samples' analysis, the bioinformatics pipeline plays an important role in the obtention of high-confidence variants, thus it is of great importance to validate it before being applied [48]. However, currently there is a lack of standardised pipelines for sputa analysis, which complicates the interpretation of results and cross-comparisons within laboratories. Since today, only two bioinformatic pipelines have been published. The first one was developed in 2021 by Cuevas-Córdoba *et al.* and was focused on discarding contaminant reads from targeted-enriched sputa, as the main source of false-positive variation, but not on obtaining high confidence variants [75]. They did not evaluate the whole genomic diversity. The second one was published in 2023 by Heupnik *et al.* and was designed to be implemented in high-burden TB countries for drug-resistance prediction and phylogenetic classification [194]. Nevertheless, they only tested it in primary liquid cultures, not in sputum samples. Hence, as a novelty, the pipeline presented in this dissertation offers a tailored analysis for complex clinical samples and the results obtained reject the hypothetical bottleneck effect on the genetic diversity imposed by culture.

The study conducted in chapter 1 had some limitations. As our main objective was to compare genetic diversity, we selected sputa containing a high initial bacterial load (XpertUltra High or Medium) to obtain high quality sequencing results. Nonetheless, we did not assess the limit of detection of the DNA extraction protocol nor the enrichment approach. What is actually true is that XpertUltra has a very high sensitivity, thus, as it detects two multicopy genes, the results on MTB quantification do not represent the real number of MTB genomes. The limit of detection of WGS compared to XpertUltra bacterial load

will be explored in future projects. Importantly, in this regard, a study published by Nilgiriwala *et al.* in 2023 evaluated the impact of the initial bacterial inoculum on the sequencing depth and coverage for drug susceptibility and typing, in both Illumina and Nanopore sequencing platforms [117]. What they found was that the initial concentration of MTB did not consistently influence the sequencing depth in both protocols. However, they call for an effort on direct sequencing strategies optimization in terms of costs and yield for clinical implementation [117], as we observed in our dataset (we successfully sequenced 61/80 sputa). Additionally, in a previous study conducted in our laboratory, six out of eleven smear-negative sputum samples were accurately sequenced by using the same enrichment approach applied in Chapter 1. Hence demonstrating the potential of culture-free sequencing for drug-resistance prediction and epidemiology [65].

Another limitation of the results presented in Chapter 1 is the lack of evaluation of heteroresistances or mutations with culture fitness cost as there were few samples in our dataset with these characteristics. This is also true when different ecotypes coexist due to mixed or polyclonal infections, especially when they present different growth rates. We only detected one strain with a known resistant conferring mutation at 5-89% frequency in A. Shockey dataset [80]; the rest of resistance and lineage determining mutations were fixed (>90% freq) in both sputa and culture pairs. Heteroresistance happens when both susceptible and resistant subpopulations coexist in a patient due to a mixed infection or to an evolution of a strain when exposed to the selective pressure imposed by the antibiotic [195,196] and it is the precursor of low-level or full-resistant tuberculosis. Therefore, if not accurately identified, it contributes to an increased risk of treatment failure. When the resistant

population appears at a low frequency it is difficult to detect either using molecular techniques or by phenotypic DST. Nowadays, there are no standard techniques for the detection of heteroresistance [196]. WGS is a promising tool for detecting heteroresistance but it is still limited for low or very-low frequency variants as they require a higher deep-sequencing effort which is not yet cost-affordable for diagnosis [196]. These scenarios require further investigation and specific analysis, as addressed by Moreno-Molina *et al. [76]*, but are outside the scope of this thesis. Thus, our findings cannot be extrapolated to polyclonal infections and whether the original proportions of the polyclonal population present in the patient lung lesion are maintained in the diagnostic sputum and the corresponding culture still remains unknown.

Overall, in Chapter 1 we conclude that all the population diversity present in sputum samples is well represented in culture samples in the most common infection scenarios which are those produced by a single strain and do not present fitness-cost mutations for *in vitro* growth. These results have a great impact for research and diagnostics as they suggest that the same results can be obtained with both types of samples. Regarding culture-based genomic studies, our findings support the accuracy/correctness of the results published. In the case of diagnostics where time is a limiting factor, our findings reflect the potential of culture-free sequencing for a prompt and accurate detection of MTB and drug susceptibility prediction.

In addition to its diagnostic value, whole genome sequencing is a useful tool to screen for drug resistance variants that are being spread in each region, i.e. surveillance of drug resistance. This is the main goal of Chapter 2, in which we evaluate the most prevalent DR mutations circulating in Southern Mozambique. In the vast majority of high-burden TB countries of the

sub-Saharan African region, since a few years ago, the frontline tool to diagnose MTB resistant to rifampicin is Xpert/RIF or Xpert/RIF Ultra. However, it imposes a diagnostic gap as it only detects the most common RIF-R variants found in the RRDR region of *rpoB* gene. When Xpert Ultra detects a mutation, the strain is considered resistant to rifampicin and the patient receives the treatment for MDR/RR-TB regimen. However, there are plenty of cases in which a strain harbouring non-common mutations in *rpoB* have been misdiagnosed and spread locally and across borders generating DR-TB outbreaks. This is the case of the *rpoB*_I491F mutation that has been detected in plenty of countries such as Eswatini [93], South Africa [94], Myanmar [197], Botswana [95] or Hong Kong [198]. In 2011, Siu *et al.* suggested the inclusion of this mutation to the Xpert assay as they observed that it was highly contributing to rifampicin resistance and was close to the RRDR region [198]. In Chapter 2, we detected 1 MTB strain from Southern Mozambique harbouring this mutation at low frequency and two strains with another mutation outside the RRDR region (rpoB_V170F). Supporting our findings, in 2024, Barilar *et al.* have also found the rpoB_I491F mutation in 31 strains from the National Tuberculosis Reference Laboratory of Mozambique, suggesting that this RIF-R variant is being spread in different provinces of the country; and, importantly, they cluster with the strains responsible for a significant MDR-TB outbreak in Eswatini [199]. As our results show, there is a high cross-border transmission rate between Mozambique, South Africa and Eswatini. One example is that, over the past 150 years, people from Mozambique have moved to South Africa to work in the mines. Thus, with a high TB incidence reported among miners in both countries, it is imperative to monitor migrant populations at the borders to improve communicable disease control and prevent transmission [200].

Furthermore, the usage of XpertUltra as a DR diagnostic tool also limits the identification of mono-isoniazid resistance cases. We observed an increasing trend of INH-R cases that are not MDR-TB (7.6% in 2018 vs 5.5% in 2014). These numbers are in line with the mean INH-R prevalence obtained in a multi-country study conducted during 2002 and 2018 (7.4% of INF-R among new patients) [201]. This suggests that around 7.5% of new cases have a baseline resistance to isoniazid before starting the treatment. Isoniazid was introduced in 1950 in clinics for TB treatment, 16 years before the release of rifampicin in 1966 [202]. Despite the widespread knowledge of the higher incidence of INH-R compared to RIF-R, no effort has been made to develop a rapid test for isoniazid susceptibility prediction [149]. One reason is that acquisition of INH-R mutations precedes resistance to rifampicin and second-line drugs [203]. Thus, it is assumed that when RIF-R is identified, the strain is treated as a MDR-TB. However, if INH-R is not detected in the prompt diagnosis, a patient receives a pan-susceptible regiment increasing the risk of treatment failure or relapse in mono-resistant cases. Recently, a new cartridge has been released by XpertUltra to identify XDR strains. Hence, it has been designed to be used after the Xpert Ultra MTB/RIF gives a positive result for rifampicin resistance, but not as a first diagnostic assay [98]. In addition, WHO recommends line probe assays such as Hain MDRTBplus that detect both INH and RIF susceptibility, but the main limitation is the challenging access to these tests which is restricted to some reference or centralised laboratories [149]. Nowadays, there are some low and moderate complexity NAAT assays, under review by the WHO, for simultaneous detection of RIF and INH susceptibility. Low complexity assays have been developed in Korea and are SD Biosensor standard M10-TB/RIF/INH (SD Biosensor) and IRON qPCR RFIA kit (Bioneer Corporation); and the moderate complexity NAAT is LiquidArray MTB-XDR

(Hain Lifescience, Germany) [2]. The improvement and final implementation of these tests will allow for the detection of multidrug-resistant or isoniazid monoresistant strains in a single assay, reducing diagnostic costs.

Targeted next generation sequencing (tNGS) is a promising alternative for TB and DR diagnostics as it overcomes the limitation of assessing only specific mutations, includes all genes of interest in a single assay, and allows targets to be updated according to new evidence of association with resistance. In Chapter 3 of this dissertation, we present a gene panel targeting 9 genes associated with resistance to seven drugs, and 2 regions containing the phylogenetic determining positions to predict lineages L1 to L6. This year, in 2024, the WHO has included tNGS as a recommended test to diagnose DR-TB in the update of the diagnosis guidelines instead of phenotypic DST in bacteriologically confirmed cases [52]. They evaluated the accuracy of tNGS on DR-TB prediction by performing a bibliographic review of all the published data and, by comparing tNGS with pDST as reference, they determined a high sensitivity (>93%) for rifampicin, isoniazid, levofloxacin, moxifloxacin and ethambutol; 88% for pyrazinamide. In addition, the specificity obtained was above 96% for all drugs [52]. In these guidelines, they presented three commercial tNGS tests that meet the quality criteria for DR-TB diagnostic: Deeplex Myc-TB (Genoscreen, France), AmPORE-TB (Oxford Nanopore Diagnostics, United Kingdom) and TBseq (Hangzhou ShengTing Medical Technology Co., China). The first one is designed for Illumina and the others for Nanopore sequencing platforms. In addition, there is a tNGS test under revision called DeepChek DST (ABL, France) that targets specific regions associated with resistance to 13 drugs [204,205]. Out of these tNGS tests, Deeplex shows the best performance for 10/15 drugs targeted, including bedaquiline, and it is

the most widely used [52]. Although our panel does not currently include any targets for new drugs, bedaquiline or clofazimine, it is possible to customise the multiplex PCR reaction by easily adding or removing amplicons and adjusting the assay conditions. This flexibility is an advantage over using fixed kits because it allows adjustments depending on the setting and if new targets are discovered.

In Chapter 3 we focused mainly on determining the frequency thresholds for a precise variant calling in MinION. We observed that 40% frequency was the cut-off value to call high-confidence variants in MinION, which is lower than the cut-off used by Cabibbe *et al.* (80% frequency) when they sequenced Deeplex Myc-TB amplicons by MinION [102]. They used a high threshold frequency to prevent false-positive variant calls when they compared with Illumina variant calling [102]. Our panel showed high accuracy in both lineage and drug-resistance prediction when compared to Illumina sequencing results for all regions (100% concordance), including the detection of an INH heteroresistant variant in *inhA* at 50% frequency detected on both platforms that could not be detected if the threshold used was higher. There were no drug-resistant associated mutations found below 40% in samples sequenced with Illumina, thus our results should not be biassed. It should be noted that, due to the error rate of the MinION platform (12% as determined in our results), lowering the frequency threshold for calling would increase the amount of false positive variant calls. In addition, MTB genome has a high GC content, and it is also challenging for MinION. Nowadays, recent advances, like the release of R10 flow cells and the update of the Guppy base caller (Version 6), promise to improve accuracy by up to 10% compared with the previous R9 flow cells, reaching 97-98% accuracy, especially in GC-rich regions [206]. Thus,

improvements in Nanopore MinION chemistry will likely help to decrease the frequency cut-offs used for variant calling.

Regarding clinical relevance, our study highlights the potential of the tNGS panel to detect drug-susceptibility in sputum samples. The advantage of obtaining high depth of coverage enables the enrichment of samples with low MTB load [207] even in smear-negative sputa, as demonstrated by Yu *et al.* [208]. However, the limited amount of diagnostic samples used in our study did not allow us to determine the limit of detection of the technique.

Targeted NGS assays have the WHO approval to be used as a frontline diagnostic tool [52]. However, they are not currently optimised in terms of costs, centralization and supply of the kits, specially to be implemented in the routine diagnostic programmes of low-income high TB-burden countries. Nowadays, there are different on-going trials to evaluate the performance and accuracy of tNGS in low- and medium- resource settings such as India, Georgia, South Africa and Namibia [209,210]. De Araujo and colleagues published in 2023 the steps and challenges of the implementation of tNGS in the TB diagnostics algorithm in Namibia [210]. They explain in detail the whole process they performed to transfer the knowledge from an expert NGS centre applied to clinics to another centre located in a low-income country that belonged to the WHO's list containing 30 countries with a high-TB burden. Importantly, these updates must be supported with external funding and sustainable development plans [210]. This is the case of FIND, which is also focused on getting WHO-recommended tests closer to high-burden TB countries and has funded projects, like Seq&Treat [211], to achieve this goal [52,209]. These pilot investigations can help organisations to understand the feasibility and challenges and guide them to achieve a successful introduction of new

diagnostic algorithms, particularly in low-resource settings [210,211]. Furthermore, by combining tNGS of diagnostic samples with a portable sequencer as MinION, both MTB and drug resistant variants can be identified in real-time, reducing the diagnostic turnaround time [212]. It should be noted that this sequencing technology is still evolving and several complement devices are under development to improve both library preparation and computational requirements for analysis. Among these innovations are Voltrax, an accessory designed to be coupled to the flow cell to automatize the library preparation and avoid this challenging step [213]; the TraxION, an add-on device integrated in the flow cell to perform all the steps from the DNA extraction to sequencing [214]; and MiniT, a GPU computer developed to eliminate the need of a separate computer [212,215]. Regarding the optimization of bioinformatic analysis, different strategies have emerged for the real-time classification of the reads of interest during the sequencing of the metagenomic samples. These approaches involve mapping of the reads to the target reference sequences (e.g.: a reference genome, target genes), while samples are being sequenced [216]. These advancements are expected to facilitate the integration of Nanopore MinION as a front-line tool for DR-TB diagnostics [212].

In Chapters 1 and 3 of this dissertation we focus on sequencing sputum samples as they are the most common sample for active pulmonary tuberculosis diagnosis. Sputa are obtained by expectoration, patients are instructed to cough and collect the sample [217]. Sometimes this expectoration may need to be induced by inhaling a saline spray [218]. Despite being the preferred non-invasive technique, it may not be suitable for everyone as some patients can not produce the sputum. This is the case of young children and

people living with HIV (PLHIV) who have difficulties to expectorate [219,220]. Therefore, alternative diagnostic non-sputum based assays are needed for these populations. Beyond sputum samples, the most promising easy-to-collect alternatives are breath and aerosols, tongue swabs, urine and stool [13,221]. Nathavitharana and colleagues in 2022 collected several ongoing research studies working on developing non-invasive technologies to detect MTB biomarkers in non-sputum samples [221]. While all of those technologies are in early stages, the WHO in 2019 approved and recommended the first biomarker for TB diagnosis in urine, the urine lipoarabinomannan or LAM [222]. The implementation of LAM assays is proceeding really slow [220] and trials of ultrasensitive LAM assays started in 2023 [74,221]. Stool-based assays have also been demonstrated to play an important role on extrapulmonary diagnosis in PLHIV, however their sensitivity (67% from pooling results of 9 studies [221]) still need to be improved for children TB diagnostics [223]. Detection of MTB in aerosols has also been achieved by capturing from face masks or flow tubes containing a filter. Early results have shown promising sensitivity (86.5%) when XpertUltra, which targets multicopy genes, was used to detect MTB. Although, further research is needed to improve its performance [224].

Overall, in this dissertation we present the potential of whole genome and targeted sequencing techniques to detect drug-resistance associated variants and genotype strains in different settings. On the one hand, we demonstrate that the original diversity of sputum samples is well represented in culture samples in the most common scenarios where there is a monoclonal infection and there are no finest-cost mutations affecting strains growth. In this regard, WGS of MTB can still be performed through culture samples for research

purposes such as retrospective studies. This application is presented in Chapter 2, in which we screen for the most common drug-resistance associated mutations circulating in Southern Mozambique. We highlight the importance of these types of analyses to identify DR-TB strains that are "escaping" from frontline diagnostic assays, especially in high-burden TB countries. On the other hand, in diagnostic scenarios where time is a limiting factor, culture-free sequencing from sputum samples is preferable. In this thesis we present two different strategies. In Chapter 1, we employed enrichment approaches from the DNA extraction step, for which we follow a differential lysis based protocol, to the capture of MTB DNA by using RNA probes when needed to obtain high quality MTB genomes. Alternatively, in Chapter 3 we predicted drug-susceptibility in sputum samples by using tNGS tailored for MinION portable platform that is suitable to be used in point-of-care centres.

The integration of WGS to research and clinical settings has represented a substantial advancement to understand the factors involving MTB pathogenesis, drug-resistance and transmission dynamics. Nevertheless, incorporating WGS in the national diagnostic and surveillance algorithms still presents some limitations regarding costs, expertise, in terms of molecular biology and bioinformatics skills, and computational requirements. These challenges are especially pronounced when considering WGS implementation in low and middle income countries that, most of them, are also high-burden TB settings [225]. Capacity-building initiatives are needed to train local people to overcome challenges in genomics research. In addition, the usage of cloud-based computational systems may also be useful for the bioinformatics analysis of the sequences. However, this strategy may be hampered by the limited internet connectivity; hence, employing data-sharing technologies

designed to manage intermittent connections, as BioTorrent, can circumvent this issue [48,225]. Cooperative efforts are essential to take advantage of the potential of WGS as a powerful tool to improve TB research and control strategies worldwide. Importantly, even though our studies did not focus on optimising sequencing for diagnostics, our findings may help or encourage other researchers to streamline these techniques for clinical purposes.

Conclusions

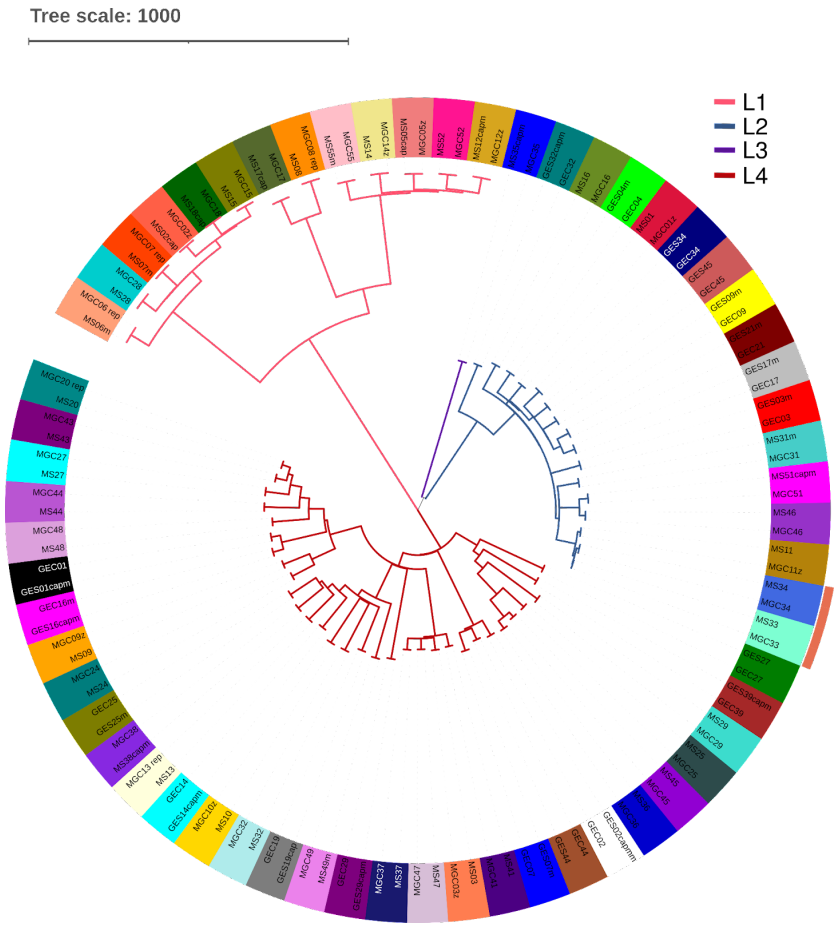
1. Culture mirrors the genetic diversity present in diagnostic sputum samples in terms of both presence and frequency of SNPs in all the datasets analysed.
2. Culture-free WGS is possible by performing both direct and enriched approaches depending on the initial amount of MTB DNA and obtaining high quality sequences in 76% of the sputa sequenced.
3. Enrichment increases MTB DNA by an average of 26.5-fold allowing high quality sequencing of sputum samples with at least 1% of MTB DNA.
4. Including appropriate filtering steps is important to detect and remove false positive variation due to artifacts coming from samples' processing. Removing supplementary alignments reduced discrepant variants between sputum and culture by more than 80% without affecting depth and coverage.
5. Results of sputum-culture diversity represent the majority of scenarios where infections are monoclonal and there are no heteroresistance mutations. However, they can not be extrapolated to polyclonal infections or to strains harbouring mutations with fitness-cost in culture.
6. Whole genome sequencing detects mutations missed by Xpert MTB/RIF Ultra in two strains from southern Mozambique. Borderline resistance mutations in the RRDR region suggest potential misdiagnosis of RIF resistance, leading to ineffective treatments and increased transmission risk.
7. Approximately 10% of cases in 2018 were INH-resistant but undetected by XpertUltra. Limited detection tools for INH resistance in high-burden TB

countries require centralised laboratories, often resulting in these cases being treated as pan-susceptible.

8. There is a high local and inter-country transmission which highlights the importance of close monitoring drug resistance to prevent the spread of resistant strains.
9. Assessment of new molecular tests, that interrogate a larger number of well-characterised mutations and provide a more reliable drug-resistance profile, is necessary.
10. tNGS approaches, currently recommended by the WHO, with MinION provide a rapid, culture-free method to detect drug-resistance. Long-read technology facilitates sequencing of entire genes, enabling a broad range of variants to be identified.
11. Amplicon approaches offer the flexibility to modify the targeted genes by including new targets or removing others that are no longer used, depending on the setting.
12. MinION variant calling shows a high correlation with Illumina variant calling when an appropriate frequency threshold is applied. It is mainly due to the error rate of MinION. Drug resistance variants are identified with both sequencing platforms regardless of the sample type.
13. The high frequency cut-off value required by MinION for variant calling limits the identification of hereroresistant variants below 40%. Ongoing flowcells with an improved chemistry will decrease the error rate and the threshold used for variant calling.
14. WGS and tNGS techniques are effective for detecting DR variants and genotyping MTB strains. It highlights the importance of adapting these techniques for clinical purposes, particularly in settings where rapid diagnostics are crucial.

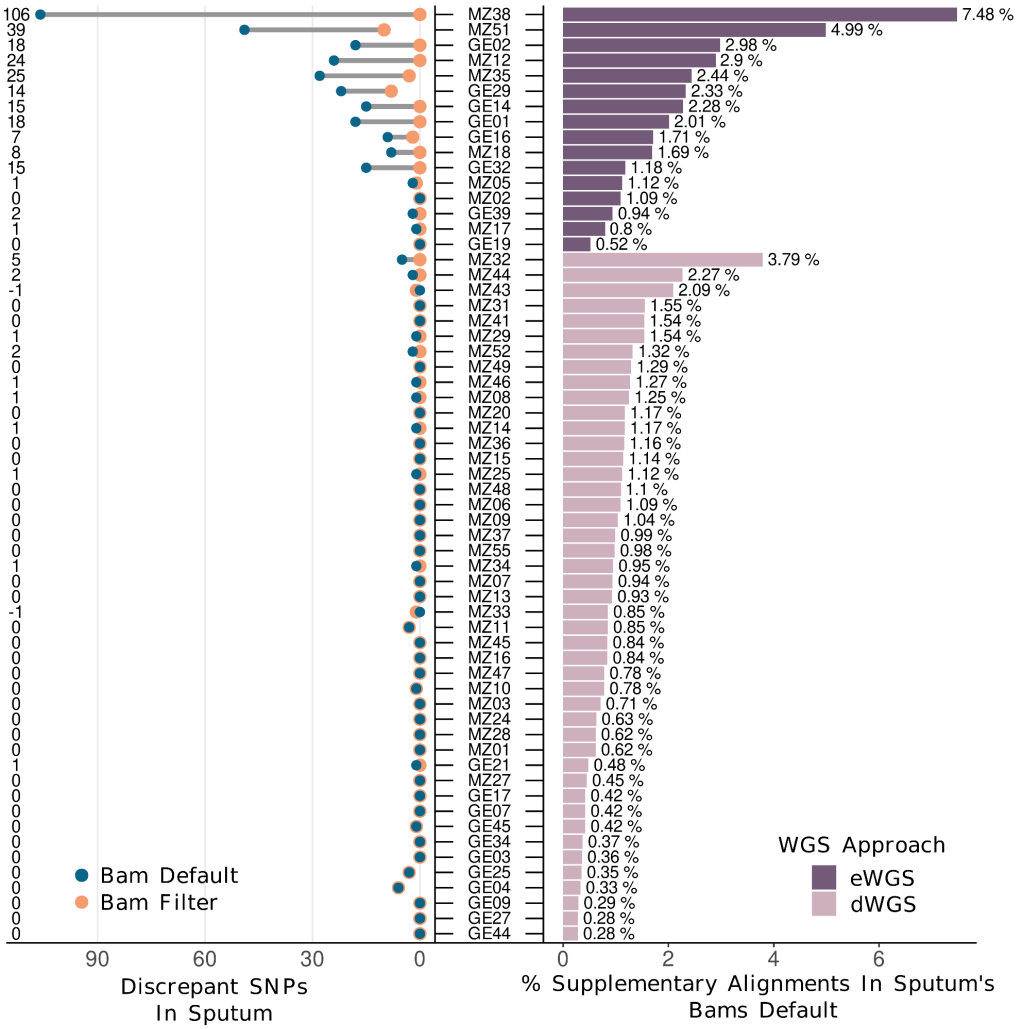
Supplementary Information

Supplementary figures



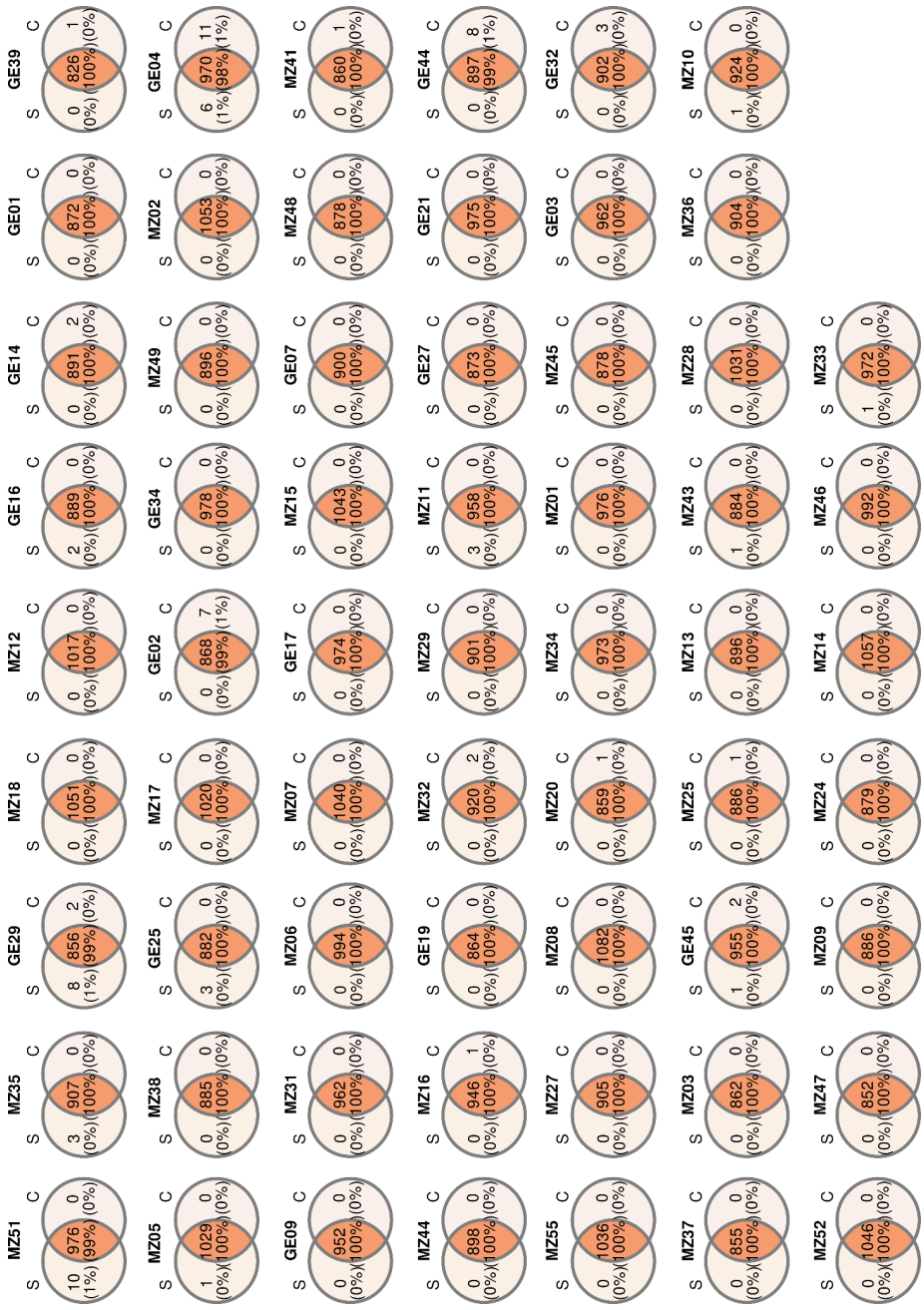
Supplementary Figure 1: Maximum likelihood phylogeny of 61 paired samples. The colour of the branches represents the lineage of the strains. Paired samples are labelled with the same colour. Transmission clusters are represented with an orange square. Each sputum-culture pair is labelled with a different colour. It was created with Itol (v6.8.1) (Letunic and Bork 2021).

Supplementary Figure 2: View of the reads of one enriched sputum mapped against the reference genome. The default BAM file containing the supplementary alignments (labelled in blue) is represented on the left side; and the filtered BAM file on the right side. The position 2389151 is highlighted in yellow in both alignments showing a false variant in the default bam file at 78.95% frequency that disappeared when discarding the supplementary aligned reads. Images extracted from Tablet viewer.

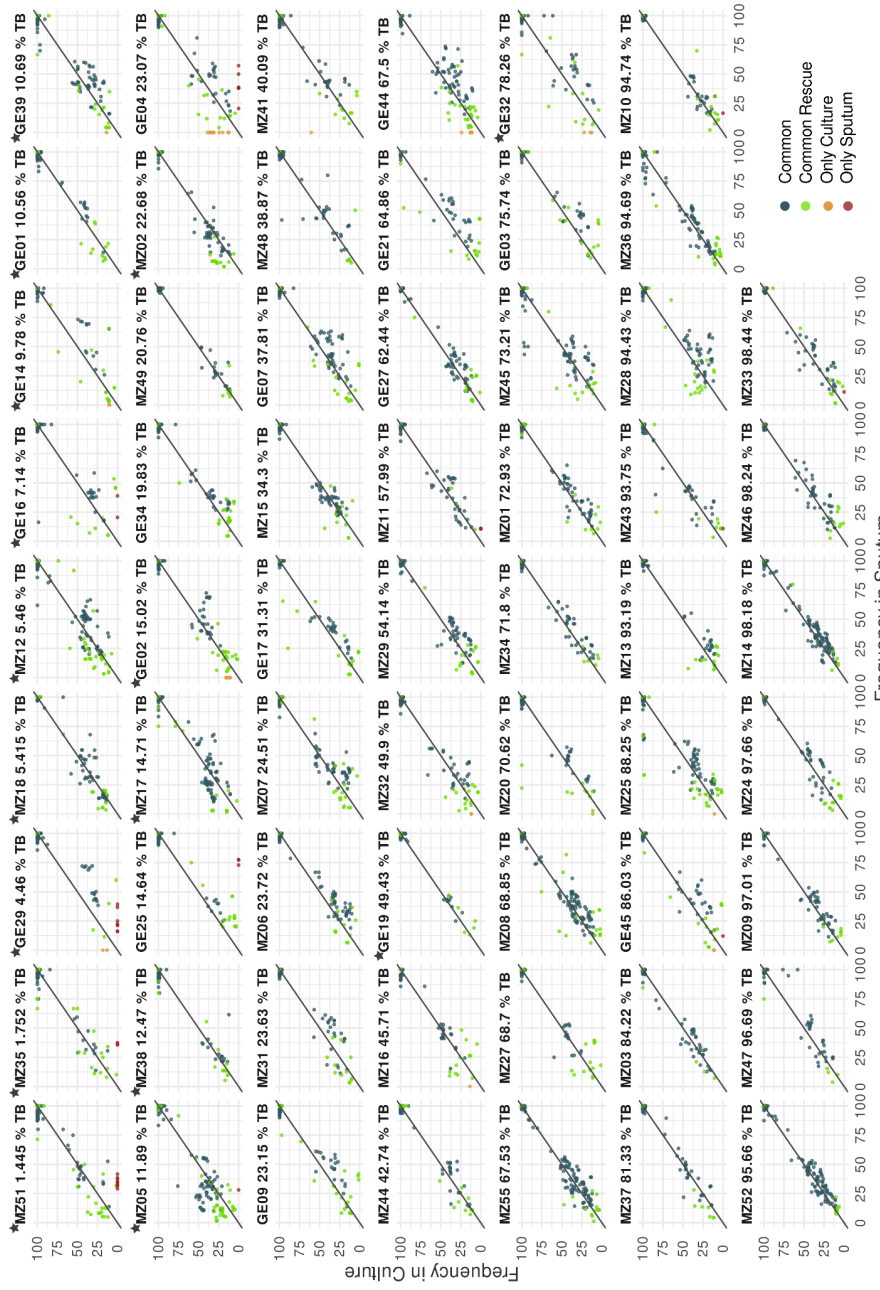


Supplementary Figure 3: Analysis of supplementary alignments of 61 pairs. The right side represents the percentage of supplementary reads in sputum files. Samples are ordered from the highest to the lowest amount of supplementary reads. Colour represents the sequencing approach. Samples sequenced directly appear in light purple and enriched samples in dark purple. The left side of the plot represents the comparison of the amount of discrepant SNPs only appearing in sputum before and after the bam file

filtering. Overlapping points represent a 0 SNP difference. Colours stand for the amount of discrepant variant calls in sputum from bam files before (in blue) and after discarding supplementary reads (in orange). Source data are provided as a Source Data file.

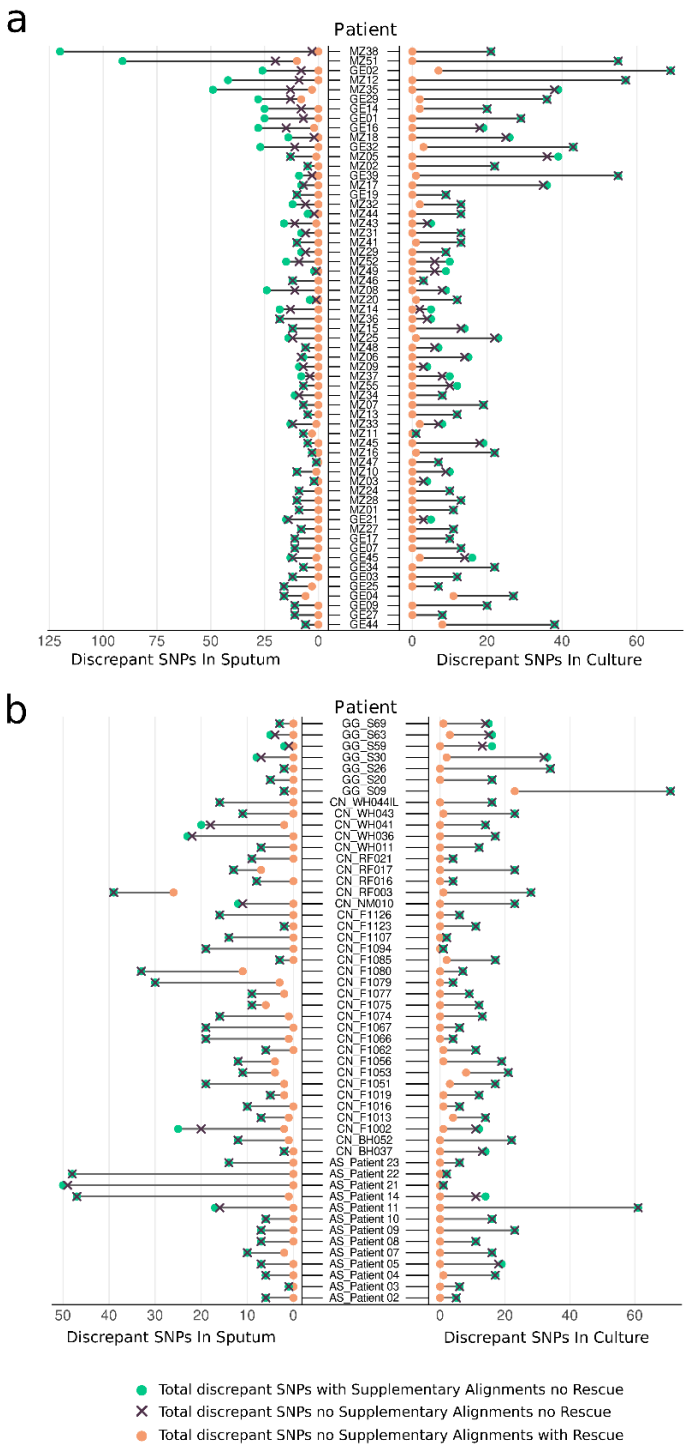


Supplementary Figure 4: Venn Diagrams of SNPs comparisons within sputum (S) - culture (C) pairs. Supplementary alignments have been discarded for this analysis. Source data are provided as a Source Data file.

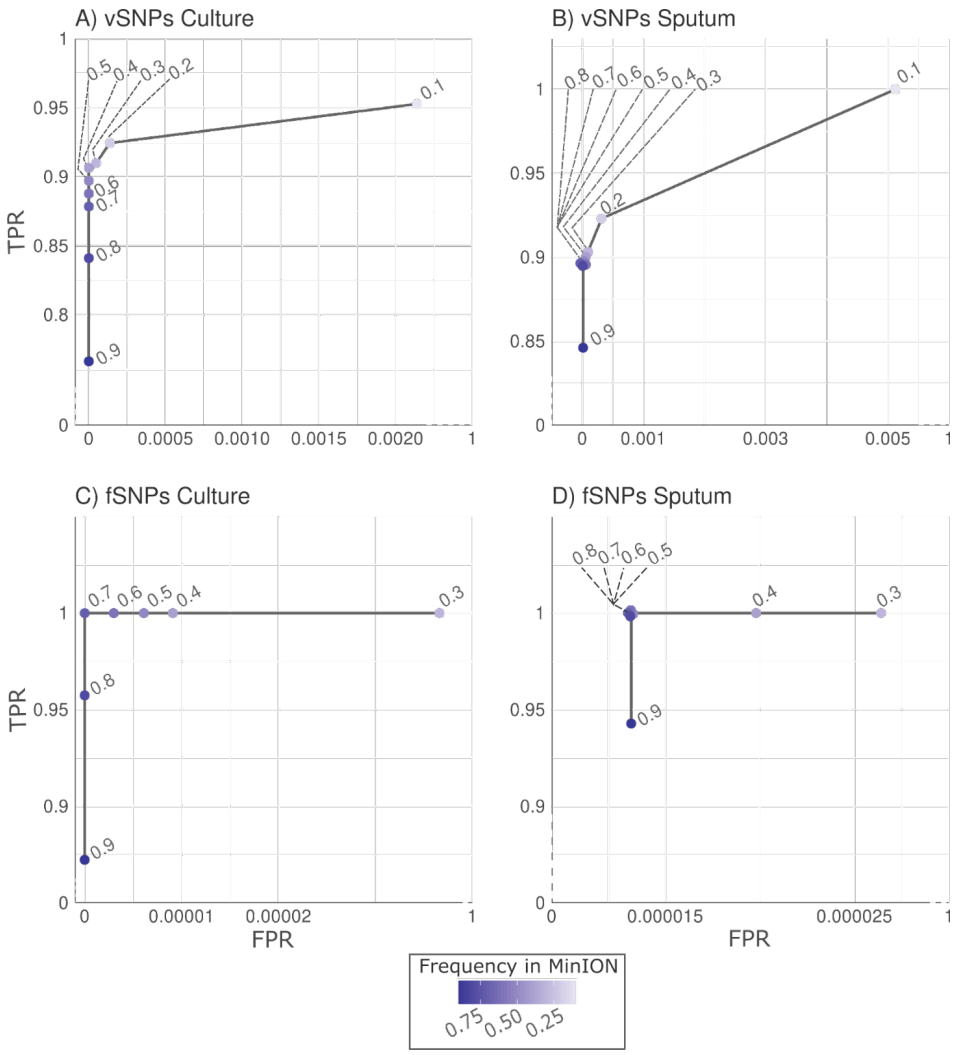


Supplementary Figure 5: Comparison of variant frequency in all 61 sputum-culture paired samples. Pairs are ordered according to the percentage of MTB in the sputum sample (from the lowest to the highest). Colour of the

points represents whether the SNPs are present in both samples (common: blue, common rescued: green), only in the culture (yellow) or only in the sputum (red). The star labelling some pairs indicates that the sputum has been enriched.


a
Patient
MZ38
MZ51
GE02
MZ12
MZ35
GE29
GE14
GE01
GE16
MZ18
GE32
MZ05
MZ02
GE39
MZ17
GE19
MZ32
MZ44
MZ43
MZ31
MZ41
MZ29
MZ52
MZ49
MZ46
MZ08
MZ20
MZ14
MZ36
MZ15
MZ25
MZ48
MZ06
MZ09
MZ37
MZ55
MZ34
MZ07
MZ13
MZ33
MZ11
MZ45
MZ16
MZ47
MZ10
MZ03
MZ24
MZ28
MZ01
GE21
MZ27
GE17
GE07
GE45
GE34
GE03
GE25
GE04
GE09
GE27
GE44
125 100 75 50 25 0
Discrepant SNPs In Sputum
0 20 40 60
Discrepant SNPs In Culture
b
Patient
GG_S69
GG_S63
GG_S59
GG_S30
GG_S26
GG_S20
GG_S09
CN_WH044IL
CN_WH043
CN_WH041
CN_WH036
CN_WH011
CN_RF021
CN_RF017
CN_RF016
CN_RF003
CN_NM010
CN_F1126
CN_F1123
CN_F1107
CN_F1094
CN_F1085
CN_F1080
CN_F1079
CN_F1077
CN_F1075
CN_F1074
CN_F1067
CN_F1066
CN_F1062
CN_F1056
CN_F1053
CN_F1051
CN_F1019
CN_F1016
CN_F1013
CN_F1002
CN_BH052
CN_BH037
AS_Patient 23
AS_Patient 22
AS_Patient 21
AS_Patient 14
AS_Patient 11
AS_Patient 10
AS_Patient 09
AS_Patient 08
AS_Patient 07
AS_Patient 05
AS_Patient 04
AS_Patient 03
AS_Patient 02
50 40 30 20 10 0
Discrepant SNPs In Sputum
0 20 40 60
Discrepant SNPs In Culture
Total discrepant SNPs with Supplementary Alignments no Rescue
Total discrepant SNPs no Supplementary Alignments no Rescue
Total discrepant SNPs no Supplementary Alignments with Rescue

Supplementary Figure 6. Analysis of discrepant SNPs due to supplementary alignments or/and rescue steps obtained in sputum (left) and culture (right). Green points represent the amount of total discrepancies obtained when no filters are applied. Purple X are the number of discrepant SNPs obtained after discarding supplementary alignments; and orange points represent the final number of discrepant SNPs obtained after discarding supplementary alignments and applying the rescue step. **a,** Comparisons in paired samples published in this paper. The order of patients is the same as in **Supplementary Figure 3**. **b,** Comparisons of the published datasets. Patients from A. Shockey's [80] dataset are labelled with an "AS" followed by the patient's code, patients from C. Nimmo's [118] dataset with a "CN", and patients from G. Goig [65] with a "GG". Source data are provided as a Source Data file.



Supplementary Figure 7: ROC curve used to set the frequency threshold employed to call variants in MinION**.** Points represent the values for recall and false positive rate obtained when applying different frequency values in MinION variant calling. Both axes are truncated. **A-B**) ROC curve used to set the frequency threshold to call variable SNPs in MinION. Recall and false positive rate value obtained using different variant calling frequency cut-offs for MinION (from 0.1 to 0.9 using increments of 0.1) and comparing with

Illumina variant calls at a 0.1 fixed threshold. **C-D)** ROC curve used to determine the frequency threshold to call fixed SNPs in MinION. Both axes are truncated. Recall and false positive rate value obtained using different variant calling frequency cut-offs for MinION (from 0.3 to 0.9 using increments of 0.1) and comparing with Illumina variant calls at a 0.9 fixed threshold. A and C represent the analysis made in culture samples, and B and D represent sputum samples.

Supplementary tables

Supplementary Table 1: Information about 61 paired samples sequenced. See Supplementary data 1 in Mariner-Llicer *et al.* [226]

Supplementary Table 2: Sublineage classification of sputum-culture paired samples.

Sputum Sample	Phylogeny Sputum	Culture Sample	Phylogeny Culture	Concordance
GES01	lineage4.3.3	GEC01	lineage4.3.3	OK
GES02	lineage4.2.1	GEC02	lineage4.2.1	OK
GES03	lineage2.2.5	GEC03	lineage2.2.5	OK
GES04	lineage2.2.9	GEC04	lineage2.2.9	OK
GES07	lineage4.2.1	GEC07	lineage4.2.1	OK
GES09	lineage2.2.10	GEC09	lineage2.2.10	OK
GES14	lineage4.10	GEC14	lineage4.10	OK
GES16	lineage4.3.3	GEC16	lineage4.3.3	OK
GES17	lineage2.2.10	GEC17	lineage2.2.10	OK
GES19	lineage4.10	GEC19	lineage4.10	OK
GES21	lineage2.2.10	GEC21	lineage2.2.10	OK
GES25	lineage4.10	GEC25	lineage4.10	OK
GES27	lineage4.1.2.1	GEC27	lineage4.1.2.1	OK
GES29	lineage4	GEC29	lineage4	OK
GES32	lineage2.2.2	GEC32	lineage2.2.2	OK
GES34	lineage2.2.10	GEC34	lineage2.2.10	OK

Sputum Sample	Phylogeny Sputum	Culture Sample	Phylogeny Culture	Concordance
GES39	lineage4.1.1	GEC39	lineage4.1.1	OK
GES44	lineage4.2.1	GEC44	lineage4.2.1	OK
GES45	lineage2.2.10	GEC45	lineage2.2.10	OK
MS01	lineage2.2	MGC01	lineage2.2	OK
MS02	lineage1.1.3	MGC02	lineage1.1.3	OK
MS03	lineage4.4.1.1	MGC03	lineage4.4.1.1	OK
MS05	lineage1.2.2	MGC05	lineage1.2.2	OK
MS06	lineage1.1.3	MGC06	lineage1.1.3	OK
MS07	lineage1.1.3	MGC07	lineage1.1.3	OK
MS08	lineage1.2.2	MGC08	lineage1.2.2	OK
MS09	lineage4.3.2	MGC09	lineage4.3.2	OK
MS10	lineage4.10	MGC10	lineage4.10	OK
MS11	lineage2.2.7	MGC11	lineage2.2.7	OK
MS12	lineage1.2.2	MGC12	lineage1.2.2	OK
MS13	lineage4.10	MGC13	lineage4.10	OK
MS14	lineage1.2.2	MGC14	lineage1.2.2	OK
MS15	lineage1.1.3	MGC15	lineage1.1.3	OK
MS16	lineage2.2	MGC16	lineage2.2	OK
MS17	lineage1.2.2	MGC17	lineage1.2.2	OK
MS18	lineage1.1.3	MGC18	lineage1.1.3	OK
MS20	lineage4.3.4.2.1	MGC20	lineage4.3.4.2.1	OK
MS24	lineage4.3.2.1	MGC24	lineage4.3.2.1	OK
MS25	lineage4.1.1.3	MGC25	lineage4.1.1.3	OK
MS27	lineage4.3.4.2.1	MGC27	lineage4.3.4.2.1	OK

Sputum Sample	Phylogeny Sputum	Culture Sample	Phylogeny Culture	Concordance
MS28	lineage1.1.3	MGC28	lineage1.1.3	OK
MS29	lineage4.1.1.2	MGC29	lineage4.1.1.2	OK
MS31	lineage2.2.7	MGC31	lineage2.2.7	OK
MS32	lineage4.10	MGC32	lineage4.10	OK
MS33	lineage2.2.7	MGC33	lineage2.2.7	OK
MS34	lineage2.2.7	MGC34	lineage2.2.7	OK
MS35	lineage3	MGC35	lineage3	OK
MS36	lineage4.1.1.3	MGC36	lineage4.1.1.3	OK
MS37	lineage4.4.1.1	MGC37	lineage4.4.1.1	OK
MS38	lineage4.10	MGC38	lineage4.10	OK
MS41	lineage4.4.1.1	MGC41	lineage4.4.1.1	OK
MS43	lineage4.3.4.2.1	MGC43	lineage4.3.4.2.1	OK
MS44	lineage4.3.4.2	MGC44	lineage4.3.4.2	OK
MS45	lineage4.1.1.3	MGC45	lineage4.1.1.3	OK
MS46	lineage2.2.7	MGC46	lineage2.2.7	OK
MS47	lineage4.4.1.1	MGC47	lineage4.4.1.1	OK
MS48	lineage4.3.4.1	MGC48	lineage4.3.4.1	OK
MS49	lineage4.10	MGC49	lineage4.10	OK
MS51	lineage2.2.7	MGC51	lineage2.2.7	OK
MS52	lineage1.2.2	MGC52	lineage1.2.2	OK
MS55	lineage1.2.2	MGC55	lineage1.2.2	OK

Supplementary Table 3: Drug-resistance associated SNPs.

Pair	Gene	Rv Gene	Genomic Position	WT codon	Mut Codon	aa Change	Antibiotic	WHO Classif	Freq Sputum (%)	Freq Culture (%)
GE01	katG	Rv1908c	2155168	AGC	ACC	S315T	INH	Assoc w R	100	100
GE02	fabG1_inhA	IG_Rv1482c_Rv1483	1673425	C	T	c-777t	INH, ETH	Assoc w R	100	100
GE04*	rpoB	Rv0667	761155	TCG	TTG	S450L	RIF	Assoc w R	100	100
	katG	Rv1908c	2155168	AGC	ACC	S315T	INH	Assoc w R	100	100
	eis_IG	IG_Rv2416c_Rv2417c	2715369	C	A	g-37t	KAN	Assoc w R	100	100
	embB	Rv3795	4248003	CAG	CGG	Q497R	EMB	Assoc w R	100	100
	ethA	Rv3854c	4326707	TGG	TAG	W256!	ETH	Assoc w R - Interim	100	100
GE29	katG	Rv1908c	2155168	AGC	ACC	S315T	INH	Assoc w R	100	100
GE32**	gyrA	Rv0006	7581	GAC	TAC	D94Y	MXF, LEV	Assoc w R	100	100
	rpoB	Rv0667	761155	TCG	TTG	S450L	RIF	Assoc w R	100	100
	katG	Rv1908c	2155168	AGC	ACC	S315T	INH	Assoc w R	94.74	100
	embB	Rv3795	4248003	CAG	CGG	Q497R	EMB	Assoc w R	100	100
GE39	katG	Rv1908c	2155168	AGC	ACC	S315T	INH	Assoc w R	100	100
MZ07	katG	Rv1908c	2155168	AGC	ACC	S315T	INH	Assoc w R	100	100
MZ09	katG	Rv1908c	2155168	AGC	ACC	S315T	INH	Assoc w R	100	100
MZ12	fabG1_inhA	IG_Rv1482c_Rv1483	1673425	C	T	c-777t	INH, ETH	Assoc w R	100	100
MZ14*	fabG1_inhA	IG_Rv1482c_Rv1483	1673425	C	T	c-777t	INH, ETH	Assoc w R	100	100
	rpoB	Rv0667	761155	TCG	TTG	S450L	RIF	Assoc w R	100	100
	katG	Rv1908c	2155168	AGC	ACC	S315T	INH	Assoc w R	100	100
	pncA	Rv2043c	2288826	GTG	GGG	V139G	PZA	Assoc w R	100	100
	embB	Rv3795	4247431	ATG	ATA	M306I	EMB	Assoc w R	100	98.59

Abbreviations: IG - Intergenic Region, INH - isoniazid, ETH - ethionamide, SM - streptomycin, EMB - ethambutol, ETH - ethionamide, KAN - kanamycin, RMP - rifampicin, FQ - fluoroquinolones, PZA - pyrazinamide, ! - Stop codon

Supplementary Table 4. Accession codes and amount of downloaded sequences included in the analysis.

Country	Study Accession ID	Doi	Samples
Algeria	PRJEB41267	CRyPTIC. San Raffaele Scientific Institute. Algeria	81
Botswana	PRJEB62480	10.3201/eid2905.220796	1421
Botswana	PRJNA436223		11
Botswana	PRJNA670836	10.12688/f1000research.28318.1	28
Congo	PRJEB27847	10.1016/j.ebiom.2018.10.013	324
Congo	PRJEB9545	10.3201/eid2303.160679	122
Congo	PRJNA300846	10.1016/S2666-5247(21)00044-6	33
Djibouti	PRJNA393924	10.1038/s41598-017-17705-3	131
Eswatini	PRJEB37777	10.1186/s13073-020-00793-8	248
Eswatini	PRJEB6273		
Eswatini	PRJEB7281		
Eswatini	PRJEB9680		
Ethiopia	PRJEB9201	10.1016/j.cub.2015.10.061	28
Gambia	PRJEB36076	10.1016/j.tube.2020.101899	442
Ghana	PRJNA616081	10.3389/fmed.2020.00161	434
Ghana	PRJEB4884	10.1038/s41598-018-29620-2	192
Ghana	PRJEB9003		
Ghana	PRJEB23179		
Ghana	PRJEB3223		
Ghana	PRJEB9545		
Ghana	PRJEB6273		
Côte d'Ivoire	PRJNA454477	10.1128/AAC.02175-18	43

Country	Study Accession ID	Doi	Samples
Kenya	PRJEB50767	10.3390/genes13030475	365
Kenya	PRJNA300846	10.1016/S2666-5247(21)00044-6	15
Liberia	PRJEB32589	10.1099/mgen.0.000325	40
Malawi	PRJEB2358	10.7554/eLife.05166	331
Malawi	PRJEB2794	10.7554/eLife.05166	1877
Mozambique	PRJEB27421	10.1099/mgen.0.000844	275
Mozambique	PRJEB50708		
Mozambique	PRJEB49677		
Mozambique	PRJEB61426	published in this paper	337
Mozambique	PRJEB23648	10.1016/j.tube.2018.04.003	13
Nigeria	PRJEB15857	10.1371/journal.pone.0184510	15
Nigeria	PRJNA300846	IeDEA consortium	12
Rwanda	PRJEB43270	10.1016/j.jctube.2022.100299	419
SouthAfrica	PRJEB14199	10.1016/S2213-2600(16)30433-7	151
SouthAfrica	PRJEB43283	10.1099/mgen.0.000815	359
SouthAfrica	PRJNA183624	10.1371/journal.pmed.1001880	212
SouthAfrica	PRJNA300846	10.1016/S2666-5247(21)00044-6	111
SouthAfrica	PRJNA428596	10.1016/S1473-3099(18)30073-2	1081
SouthAfrica	PRJNA670836	10.12688/f1000research.28318.1	22
SouthAfrica	TBARC	TB-ARC - MRC SA	179
Tanzania	PRJEB49562	10.1371/journal.ppat.1010893	676
Tanzania	PRJNA300846	10.1016/S2666-5247(21)00044-6	29
Tanzania	PRJNA670836	10.12688/f1000research.28318.1	390
Tunisia	PRJEB30463	10.3201/eid2503.181370	46
Tunisia	PRJEB39509	10.1016/j.ijid.2020.11.195	29
Tunisia	PRJNA306588	10.1016/j.ijid.2020.11.195	13

Country	Study Accession ID	Doi	Samples
Uganda	PRJEB2424	10.1371/journal.pone.0083012	51
Zimbabwe	PRJEB18529	10.1186/s12916-017-0834-4	26

Supplementary Table 5: Resistance profile of all the strains harbouring at least one mutation included in the WHO Catalog [104].

Patient	INH	RMP	ETH	SM	PZA	EMB	KAN	AMK	CPR	LEV	MXF	BDQ	DLM	Mono-poly-R	poly-R Profile	INH Type
1705225_2	fabG1_inhA_c-777t (fabG1_c-15t)	S	fabG1_inhA_c-777t (fabG1_c-15t)	S	S	S	S	S	S	S	S	S	S	poly-R	INH+ETH	INH noMDR
1726308_5	fabG1_inhA_c-777t (fabG1_c-15t)	S	fabG1_inhA_c-777t (fabG1_c-15t)	S	S	S	S	S	S	S	S	S	S	poly-R	INH+ETH	INH noMDR
1693154_1	fabG1_inhA_g-154a (fabG1_L203L)	S	fabG1_inhA_g-154a (fabG1_L203L)	S	S	S	S	S	S	S	S	S	S	poly-R	INH+ETH	INH noMDR
1689552_2	fabG1_inhA_c-777t (fabG1_c-15t)	S	fabG1_inhA_c-777t (fabG1_c-15t)	S	S	S	S	S	S	S	S	S	S	poly-R	INH+ETH	INH noMDR
1705053_1	fabG1_inhA_c-777t (fabG1_c-15t)	S	fabG1_inhA_c-777t (fabG1_c-15t)	S	S	S	S	S	S	S	S	S	S	poly-R	INH+ETH	INH noMDR
1720088_2	fabG1_inhA_c-777t (fabG1_c-15t)	S	fabG1_inhA_c-777t (fabG1_c-15t)	S	S	S	S	S	S	S	S	S	S	poly-R	INH+ETH	INH noMDR
1758534_7	fabG1_inhA_c-777t (fabG1_c-15t)	S	fabG1_inhA_c-777t (fabG1_c-15t)	S	S	S	S	S	S	S	S	S	S	poly-R	INH+ETH	INH noMDR
1689410_5	fabG1_inhA_g-154a (fabG1_L203L)	S	fabG1_inhA_g-154a (fabG1_L203L)	S	S	S	S	S	S	S	S	S	S	poly-R	INH+ETH	INH noMDR

Patient	INH	RMP	ETH	SM	PZA	EMB	KAN	AMK	CPR	LEV	MXF	BDQ	DLM	Mono-poly-R	poly-R Profile	INH Type
MZ125	fabG1_inhA_c-777t (fabG1_c-15t)	S	fabG1_inhA_c-777t (fabG1_c-15t)	S	S	S	S	S	S	S	S	S	S	poly-R	INH+ETH	INH noMDR
MZ133	fabG1_inhA_c-777t (fabG1_c-15t)	S	fabG1_inhA_c-777t (fabG1_c-15t)	S	S	S	S	S	S	S	S	S	S	poly-R	INH+ETH	INH noMDR
MZ146	fabG1_inhA_c-777t (fabG1_c-15t)	S	fabG1_inhA_c-777t (fabG1_c-15t)	S	S	S	S	S	S	S	S	S	S	poly-R	INH+ETH	INH noMDR
MZ158	fabG1_inhA_c-777t (fabG1_c-15t)	S	fabG1_inhA_c-777t (fabG1_c-15t)	S	S	S	S	S	S	S	S	S	S	poly-R	INH+ETH	INH noMDR
MZ180	fabG1_inhA_c-777t (fabG1_c-15t)	S	fabG1_inhA_c-777t (fabG1_c-15t)	S	S	S	S	S	S	S	S	S	S	poly-R	INH+ETH	INH noMDR
MZ277	fabG1_inhA_c-777t (fabG1_c-15t)	S	fabG1_inhA_c-777t (fabG1_c-15t)	S	S	S	S	S	S	S	S	S	S	poly-R	INH+ETH	INH noMDR
MZ142	fabG1_inhA_c-777t (fabG1_c-15t)	S	fabG1_inhA_c-777t (fabG1_c-15t)	S	S	S	S	S	S	S	S	S	S	poly-R	INH+ETH	INH noMDR
MZ53	fabG1_inhA_c-777t (fabG1_c-15t)	S	fabG1_inhA_c-777t (fabG1_c-15t)	gid_103_del_1_gc_g	S	S	S	S	S	S	S	S	S	poly-R	INH+ETH+SM	INH noMDR
1688828_9	katG_S315T	S	S	gid_472_del_1_gc_g	S	S	S	S	S	S	S	S	S	poly-R	INH+SM	INH noMDR
MZ271	katG_S315T	S	S	rpsL_K43R	S	S	S	S	S	S	S	S	S	poly-R	INH+SM	INH noMDR

Patient	INH	RMP	ETH	SM	PZA	EMB	KAN	AMK	CPR	LEV	MXF	BDQ	DLM	Mono-poly-R	poly-R Profile	INH Type
MZ74	katG_S315T	S	S	gid_103_del_1_gc_g	S	S	S	S	S	S	S	S	S	poly-R	INH+SM	INH noMDR
MZ78	katG_S315T	S	S	gid_103_del_1_gc_g	S	S	S	S	S	S	S	S	S	poly-R	INH+SM	INH noMDR
1669572_6	fabG1_inhA_g-154a (fabG1_L203L)	rpoB_S450L; rpoB_D435F	fabG1_inhA_g-154a (fabG1_L203L)	S	S	S	S	S	S	S	S	S	S	MDR	INH+RMP+ETH	INH MDR
1686630_0	fabG1_inhA_c-777t (fabG1_c-15t)	rpoB_H445D	fabG1_inhA_c-777t (fabG1_c-15t)	rpsL_K43R	S	embB_M306I	S	S	S	S	S	S	S	MDR	INH+RMP+ETH+SM+EMB	INH MDR
1686647_8	katG_S315T	rpoB_S450L	ethA_33_del_1_gc_g	rpsL_K43R	S	embB_M306L	S	S	S	S	S	S	S	MDR	INH+RMP+ETH+SM+EMB	INH MDR
1688851_7	fabG1_inhA_c-777t (fabG1_c-15t)	rpoB_H445D	fabG1_inhA_c-777t (fabG1_c-15t)	rpsL_K43R	S	embB_M306I	S	S	S	S	S	S	S	MDR	INH+RMP+ETH+SM+EMB	INH MDR
1705272_6	katG_S315T	rpoB_S450L	ethA_861_del_1_gt_g	rpsL_K43R	pncA_T135P	embB_D354A	S	S	S	S	S	S	S	MDR	INH+RMP+ETH+SM+PZA+EMB	INH MDR

Abbreviations of Supplementary Table 5: AMI: amikacin, CAP: capreomycin, INH: isoniazid, RIF: rifampicin, ETH: ethionamide, SM: streptomycin, PZA: pyrazinamide, EMB: ethambutol, LEV: levofloxacin, MXF: moxifloxacin; S: susceptible; red: resistant (with the mutation conferring resistance).

Supplementary Table 6: Primer sequences and their corresponding final concentrations for PCR.

Primer	Sequence	PCR Concentration (µM)
katG_F_1	TGCGGCGGGTTGTGGTTGAT	0.1
katG_R_2	GCGCTACGAGTCCAGGGTCG	0.1
gyrA_F_1	AGATGACAGACACGACGTTGCC	0.2
gyrA_R_1	AGCCTAGCTGCCCGATTCCT	0.2
inhA_F_2.1	TTGGCGCCATGGAAGGCAGA	0.1
inhA_R_2.1	TGGCTAGTCGAGCGAACCGC	0.1
gyrB_F_4	TGCGGTTGGCGGCCTATCAA	0.2
gyrB_R_1	CGCAGGGTTGCGTTAGACATCC	0.2
pncA_F_1	ATCCGGACACTTGCCACCCG	0.2
pncA_R_1	GTTGGCGTTGCGTTGGGTCC	0.2
rpoB_F_int	GTCGCATGAAGTGCTGGAAGG	0.3
rpoB_R_int	GAAGTTGACGTCGAGCACGTAAC	0.3
embB_F_1	CCAGACGCCGTTGTCGAGGA	0.6
embB_R_2	GGCCTGGTGCATACCGAGCA	0.6
eis_F_1	AGGGTGACGAGTCCTGGGGT	0.2
eis_R_1	GAAGCGATGAGGTGGGGGCA	0.2
rrs_F_1	ACGAGCGTCCGAAGGCTGTC	0.2
rrs_R_2	GCCATCACCACCCTCCTCCG	0.2
L125_F_3	ACCCGCACTATGCCTGGCTG	0.4
L125_R_3	TGGATGGCGCTCAACGGGAG	0.4
L346_F_1	TCCCGACGGTGCGTGACTTG	0.4
L346_R_3	CGGCAGTGCCAGTTCATGCC	0.4

Supplementary Table 7: Barcode sequences.

Barcode		Sequence
BC01	Barcode_2_263_Fw	ACGATCTACCGCACTGAATCCAACACAAACGATCTAT
	Barcode_2_263_Rv	TAGATCGTTTGTGTTGGATTCAGTGCGGTAGATCGT
BC02	Barcode_3_262_Fw	ACGATCTAAGTCTACGGTGAGTTGACAAACGATCTAT
	Barcode_3_263_Rv	TAGATCGTTTGTCAACTCACCGTAGACTTAGATCGT
BC03	Barcode_1070_914_Fw	ACGATCTAGTGACCGCTACAGCTCATCAACGATCTAT
	Barcode_1070_914_Rv	TAGATCGTTGATGAGCTGTAGCGGTCACTAGATCGT
BC04	Barcode_1313_1025_Fw	ACGATCTAATGTGGTCGGTGATACCGTAACGATCTAT
	Barcode_1313_1025_Rv	TAGATCGTTACGGTATCACCGACCACATTAGATCGT
BC05	Barcode_201_70_Fw	ACGATCTATAGTGGTGAACGGTCTAGAAACGATCTAT
	Barcode_201_70_Rv	TAGATCGTTTCTAGACCGTTCACCACTATAGATCGT
BC06	Barcode_1400_2062_Fw	ACGATCTATAGGATCTGGCAGGTTGTCTACGATCTAT
	Barcode_1400_2062_Rv	TAGATCGTAGACAACCTGCCAGATCCTATAGATCGT
BC07	Barcode_565_556_Fw	ACGATCTAGACTTGCAGAGCTTATGAGAACGATCTAT
	Barcode_565_556_Rv	TAGATCGTTCTCATAAGCTCTGCAAGTCTAGATCGT
BC08	Barcode_579_1859_Fw	ACGATCTATCAGACTAGAACAACGCTCCACGATCTAT
	Barcode_579_1859_Rv	TAGATCGTGGAGCGTTGTTCTAGTCTGATAGATCGT
BC09	Barcode_650_1538_Fw	ACGATCTATATGTTGCGAGCACCATTCGACGATCTAT
	Barcode_650_1538_Rv	TAGATCGTCGAATGGTGCTCGCAACATATAGATCGT
BC10	Barcode_749_1984_Fw	ACGATCTACCGTCAGACACTGAACCAGTACGATCTAT
	Barcode_749_1984_Rv	TAGATCGTACTGGTTCAGTGTCTGACGGTAGATCGT
BC11	Barcode_761_1321_Fw	ACGATCTATACGACCACAAGATATTCGGACGATCTAT
	Barcode_761_1321_Rv	TAGATCGTCCGAATATCTTGTGGTCGTATAGATCGT
BC12	Barcode_1183_2117_Fw	ACGATCTAGTCATAACAGGCACTGTCTTACGATCTAT
	Barcode_1183_2117_Rv	TAGATCGTAAGACAGTGCCTGTTATGACTAGATCGT
BC13	Barcode_1438_1303_Fw	ACGATCTACTCACAAGCGATAGTTCCGGACGATCTAT

Barcode		Sequence
	Barcode_1438_1303_Rv	TAGATCGTCCGGAACTATCGCTTGTGAGTAGATCGT
BC14	Barcode_1609_1571_Fw	ACGATCTATCCAGTGCTGCGACTACGTGACGATCTAT
	Barcode_1609_1571_Rv	TAGATCGTCACGTAGTCGCAGCACTGGATAGATCGT
BC15	Barcode_1696_1697_Fw	ACGATCTAGATTCGGCACACTCTCGCACACGATCTAT
	Barcode_1696_1697_Rv	TAGATCGTGTGCGAGAGTGTGCCGAATCTAGATCGT
BC16	Barcode_1723_1789_Fw	ACGATCTACAATCCAGGCATCCAGAGCCACGATCTAT
	Barcode_1723_1789_Rv	TAGATCGTGGCTCTGGATGCCTGGATTGTAGATCGT
BC17	Barcode_1835_1863_Fw	ACGATCTAATAGAAGTCCTTATGTCTCCACGATCTAT
	Barcode_1835_1863_Rv	TAGATCGTGGAGACATAAGGACTTCTATTAGATCGT
BC18	Barcode_1946_1947_Fw	ACGATCTACGAGTAGTTCAAGGTAGTTCACGATCTAT
	Barcode_1946_1947_Rv	TAGATCGTGAACTACCTTGAACTACTCGTAGATCGT
BC19	Barcode_2081_2107_Fw	ACGATCTACGCACACGTTCCGCTTGCTTACGATCTAT
	Barcode_2081_2107_Rv	TAGATCGTAAGCAAGCGGAACGTGTGCGTAGATCGT
BC20	Barcode_1975_1976_Fw	ACGATCTAAGCGTTACATTTGACAGCATACGATCTAT
	Barcode_1975_1976_Rv	TAGATCGTATGCTGTCAAATGTAACGCTTAGATCGT

Supplementary Table 8: Phylogenetic determining variants.

Lineage	SNP
1	4357773 GA
2	4357804 TG
3	1281984 GA
4	1281771 CT
5	4357657 GA
6	1281685 CG

Supplementary Table 9: Percentage of agreement between Illumina and MinION SNPs at different variant calls in MinION and 0.1 frequency in Illumina.

Frequency in MinION	TP	FP	FN	TN	TPR	TNR
0.1	140	1478	5	478244	0.9655172	0.996919
0.2	134	93	11	479582	0.9241379	0.9998061
0.3	132	13	14	479654	0.9041096	0.9999729
0.4	**132**	**3**	**14**	**479664**	**0.9041096**	**0.9999937**
0.5	131	2	15	479665	0.8972603	0.9999958
0.6	130	2	16	479665	0.890411	0.9999958
0.7	129	2	17	479665	0.8835616	0.9999958
0.8	125	2	21	479665	0.8561644	0.9999958
0.9	115	2	31	479664	0.7876712	0.9999958

Abbreviations: TP, true positive variants; FN, false negative variants; FP, false positive variants; TN: true negative variants; TPR: true positive rate; TNR: true negative rate.

Supplementary Table 10: Resistance profile and phylogenetic classification of samples.

Sample	Antibiotic							Lineage	Agreement	
	FQ	RMP	SM	INH	PZA	KAN	EMB		Resistance Profile	Lineage
N0067_L1	S	S	S	S	S	S	S	L1	100	100
G981_L2	S	S	S	S	S	S	S	L2	100	100
G107_L3	S	S	S	S	S	S	S	L3	100	100
G770_L4	S	S	S	S	S	S	S	L4	100	100
G1961_L5	S	**R**	S	S	S	S	**R**	L5	100	100
G1952_L6	S	S	S	S	S	S	S	L6	100	100
G870	**R**	**R**	S	**R**	S	**R**	**R**	L2	100	100
G841	S	**R**	S	S	S	S	S	L4	100	100
G1800	**R**	**R**	S	**R**	S	**R**	**R**	L2	100	100
182320_M20	S	S	S	S	S	S	S	L4	100	100
G1646_W19	S	S	S	S	S	S	S	L4	100	100
G2267_W20	S	S	S	S	S	S	S	L4	100	100
G2103_W23	S	S	S	S	S	S	S	L4	100	100
G2284_W26	S	S	S	S	S	S	S	L4	100	100
G1335_W27	S	S	S	S	S	S	S	L2	100	100

Abbreviations: FQ, fluoroquinolones; RMP, rifampicin; SM, streptomycin; INH, isoniazid; PZA, pyrazinamide; KAN, kanamycin; EMB, ethambutol; L, lineage; S, susceptible; R, resistant. Strains harbouring a resistance conferring mutation are highlighted in red.

Supplementary Table 11: Sequences of reverse primers for the qPCR of the PCR product.

Primer	Sequence
katG_Rv_qPCR	CGTGGACCTGGTCTTCGGGTC
gyrA_Rv_qPCR	GGCGGAAGCCGGAATCGAAC
inhA_Rv_qPCR	GCCCTGGCTGCGGGTGTATT
pncA_Rv_qPCR	GGGTGTGCTGCCGATGACGA
gyrB_Rv_qPCR	ATCGATGTGCACCCGCCTGG
embB_Rv_qPCR	CAATCAGCCCGGCGATGGTG
rpoB_Rv_qPCR	GGTCTGGACGTCAAGGAGTCCC
L125_Rv_qPCR	CACCCCGGTGATCCACAGCA
L346_Rv_qPCR	GCGCCACGCCGACATATTCC
eis_Rv_qPCR	TCGCGGTGCTGGTGACGG
rrs_Rv_qPCR	GTCCGAGCGTCTGCACCGAG

Supplementary Table 12: ROC analysis of vSNPs in MinION (Illumina frequency >= 0.1).

Freq Minion	TP	FP	FN	TN	TPR Recall	TNR	Sample Type	Agreement (%)
0.1	39	776	0	150696	1	0.9948769	Sputum	4.79
0.2	36	48	3	151408	0.9230769	0.9996831	Sputum	42.86
0.3	35	4	4	151449	0.8974359	0.9999736	Sputum	89.74
0.4	35	3	4	151450	0.8974359	0.9999802	Sputum	92.11
0.5	35	2	4	151451	0.8974359	0.9999868	Sputum	94.59
0.6	35	2	4	151451	0.8974359	0.9999868	Sputum	94.59
0.7	35	2	4	151451	0.8974359	0.9999868	Sputum	94.59
0.8	35	2	4	151451	0.8974359	0.9999868	Sputum	94.59
0.9	33	2	6	151451	0.8461538	0.9999868	Sputum	94.29
0.1	101	702	5	327548	0.9528302	0.9978614	Culture	12.58
0.2	98	45	8	328174	0.9245283	0.9998629	Culture	68.53
0.3	97	9	10	328205	0.9065421	0.9999726	Culture	91.51
0.4	97	0	10	328214	0.9065421	1	Culture	100
0.5	96	0	11	328214	0.8971963	1	Culture	100
0.6	95	0	12	328214	0.8878505	1	Culture	100
0.7	94	0	13	328214	0.8785047	1	Culture	100
0.8	90	0	17	328214	0.8411215	1	Culture	100
0.9	82	0	25	328213	0.7663551	1	Culture	100

Abbreviations: TP, true positive variants; FN, false negative variants; FP, false positive variants; TN: true negative variants; TPR: true positive rate; TNR: true negative rate.

Supplementary Table 13: ROC analysis of fSNPs in MinION (Illumina frequency >= 0.9).

Freq Minion	TP	FP	FN	TN	TPR/Recall	TNR	Sample Type	Agreement
0.3	35	4	0	151452	1	0.9999736	Sputum	89.74358974
0.4	35	3	0	151453	1	0.9999802	Sputum	92.10526316
0.5	35	2	0	151454	1	0.9999868	Sputum	94.59459459
0.6	35	2	0	151454	1	0.9999868	Sputum	94.59459459
0.7	35	2	0	151454	1	0.9999868	Sputum	94.59459459
0.8	35	2	0	151454	1	0.9999868	Sputum	94.59459459
0.9	33	2	2	151454	0.9428571	0.9999868	Sputum	94.28571429
0.3	94	12	0	328214	1	0.9999634	Culture	88.67924528
0.4	94	3	0	328223	1	0.9999909	Culture	96.90721649
0.5	94	2	0	328224	1	0.9999939	Culture	97.91666667
0.6	94	1	0	328225	1	0.999997	Culture	98.94736842
0.7	94	0	0	328226	1	1	Culture	100
0.8	90	0	4	328226	0.9574468	1	Culture	100
0.9	82	0	12	328225	0.8723404	1	Culture	100

Abbreviations: TP, true positive variants; FN, false negative variants; FP, false positive variants; TN: true negative variants; TPR: true positive rate; TNR: true negative rate.

Supplementary Table 14: Recall/TNR of drug-resistance and lineage prediction by gene.

Prediction	FQ		RMP	SM	INH		PZA	KAN	EMB	L125	L346	Median
Gene	*gyrB*	*gyrA*	*rpoB*	*rrs*	*inhA*	*katG*	*pncA*	*eis*	*embB*			
Total SNPs	45675	38115	46330	38867	41445	37110	46138	44989	51330	42355	47460	44989
TP	11	21	16	3	5	11	9	5	19	23	9	11
TN	45664	38094	46310	38863	41440	37099	46129	44983	51311	42316	47448	44983
FP	0	0	0	1	0	0	0	1	0	4	0	0
FN	0	0	4	0	0	0	0	0	0	12	3	0
Recall all SNPs (95% CI)	1 (0.72-1)	1 (0.84-1)	0.8 (0.56-0.94)	1 (0.29-1)	1 (0.48- 1)	1 (0.72-1)	1 (0.66-1)	1 (0.78-1)	1 (0.82-1)	0.66 (0.48-0.81)	0.75 (0.43-0.95)	1 (0.72-1)
TNR all SNPs (95% CI)	1 (0.99-1)	1 (0.99-1)	1 (0.99-1)	0.99 (0.99-1)	1 (0.99-1)	1 (0.99-1)	1 (0.99-1)	0.99 (0.99-1)	1 (0.99-1)	0.99 (0.98-1)	1 (0.99-1)	1 (0.99- 1)
Agreement all SNPs (%)	100	100	80	75	100	100	100	83.33	100	58.97	75	100
Accuracy all SNPs (%) (95% CI)	100 (99.99-100)	100 (99.99-100)	99.99 (99.99-100)	100 (99.99-100)	100 (99.99-100)	100 (99.99-100)	100 (99.99-100)	100 (99.99-100)	100 (99.99-100)	99.96 (99.94-99.98)	99.99 (99.98-1)	100 (99.99-100)
Sensitivity DR SNPs (95% CI)	1 (0.025-1)	1 (0.025-1)	1 (0.40-1)	No DR SNPs found	No DR SNPs found	1 (0.16-1)	1 (0.025-1)	1 (0.16-1)	1 (0.29-1)	1 (0.54-1)	1 (0.66-1)	1 (0.16-1)
Specificity DR SNPs (95% CI)	1 (0.98-1)	1 (0.97-1)	1 (0.99-1)	1 (0.97-1)	1 (0.88-1)	1 (0.97-1)	1 (0.99- 1)	1 (0.94-1)	1 (0.99-1)	1 (0.91-1)	1 (0.90-1)	1 (0.97-1)

Prediction	FQ		RMP	SM	INH		PZA	KAN	EMB	L125	L346	Median
Gene	*gyrB*	*gyrA*	*rpoB*	*rrs*	*inhA*	*katG*	*pncA*	*eis*	*embB*			
Agreement DR SNPs (%)	100	100	100	100	100	100	100	100	100	100	100	100
Accuracy DR SNPs (%) (95% CI)	100 (98.13-100)	100 (96.55-100)	100 (98.89-100)	NA	NA	100 (97.30-100)	100 (99.84-100)	100 (94.04-100)	100 (98.84-100)	100 (92.13-100)	100 (92.13-100)	100 (97-100)

Abbreviations: TP, true positive variants; FN, false negative variants; FP, false positive variants; TN: true negative variants; TPR: true positive rate; TNR: true negative rate. FQ, fluoroquinolones; RMP, rifampicin; SM, streptomycin; INH, isoniazid; PZA, pyrazinamide; KAN, kanamycin; EMB, ethambutol; L, lineage

Chapter 3. Supplementary Methods

1. Optimization of the single-tube multiplex PCR reaction

A qPCR of the PCR product was performed to set up the optimum concentration of each primer to obtain an equimolar amount of all the amplicons. The qPCR reaction volume was 20uL containing 1X volume Kapa SYBR Fast qPCR master mix (KAPA Biosystems, Roche), 0.2uM primers and 0.2ng of DNA (PCR amplicons). Water was added to reach the final volume. For this test, reverse primers were designed to amplify a small region of each gene (150-200bp) checking the specificity to avoid non-specific interactions (See **Supplementary Table 11** containing the list of the qPCR primer sequences).

2. Barcodes design

We designed a set of 20 different barcodes, maximising the genetic distance between them and supporting the correction of insertions, deletions and substitutions. We used the R package DNABarcodes to create barcodes of 36 bp using the Sequence-Levenshtein metric under the Ashlock heuristic. Each barcode consisted of a unique region of 20bp and two common regions of 8 bp (ACGATCTA) flanking this central sequence (See **Supplementary Table 7** containing customised barcode sequences)**.** Primers with the barcode sequences and their complementary strands were ordered from IDT with HPLC purification.

Barcodes were designed using a R package called DNABarcodes using the Sequence-Levenshtein metric under the Ashlock heuristic model [227]:

mySeqlevSet <- create.dnabarcodes(10, metric="seqlev", euristic="ashlock")

This created 10 bp barcodes. We built our 20 bp barcodes by joining two different barcodes of 10 bp.

3. Enrichment PCR step for lower yield samples

An additional PCR step was performed after the ligation of barcodes, only in samples that didn't yield less than 500ng of gene product in the first PCR. The reaction mix was prepared in a final volume of 50uL containing 10uL HiFi GC buffer 5X, dNTPs 0.3mM each, 0.4U of Kapa HiFi HS polymerase (Kapa Biosystems®), 5uL of each forward barcode 10uM were added to its corresponding sample, 2ng of barcoded DNA and water until reaching the final volume. Thermocycling conditions consist on a denaturation step at 95°C during 3 minutes followed by 15 cycles of amplification as follows: 20 seconds at 98°C, 15 seconds at 65°C (primers annealing), 2 minutes at 72°C (primers extension) and 5 minutes at 72°C (final extension step). Final PCR enriched product was purified with 0.6X volume AMPure XP magnetic beads (following the Agencourt AMPure XP purification protocol) and eluted from the beads with Tris 10mM (pH=8.5) and quantified with Qubit®.

4. Modifications in the MinION library preparation protocol

Incubation time was increased from 10 to 30 minutes in the dA-tailing step, and also in the adapter ligation step to 30 minutes.

5. Base calling

We run guppy basecaller using 'flip flop' algorithm [176] by applying the following parameters:

```
guppy_basecaller --input_path $INPUT \
      --save_path $OUTPUT\
      --verbose_logs \
      --cpu_threads_per_caller 1 \
      --num_callers 20 \
      --config dna_r9.4.1_450bps_flipflop.cfg
```

6. Demultiplexing

To demultiplex the multi-fastq obtained after performing the base calling we run porechop (https://github.com/rrwick/Porechop):

```
python porechop-runner.py -i $FASTQ -b demultiplex/
```

This python script takes a multifasta file containing all the reads and classifies the reads depending on their barcode sequence generating a folder with separated fasta files, one per sample. We modified the Oxford Nanopore barcode sequences by our customised barcode sequences. To demultiplex the reads, default parameters were used: --check-reads 10000, --adapter_threshold 90, --scoring_scheme (match = 3, mismatch = -6, gap open = -5, gap extend = -2), --end_size 150, --min_trim_size 4, --ehd_threshold 75, --extra_end_trim 2, --middle_threshold 85, --min_split_read_size 1000, --extra_middle_trim_good_side 10, --extra_middle_trim_bad_side 100.

7. Mapping

Mapping was done aligning reads to a multi fasta file containing the sequences

of the genome regions included in the panel (extracted from the reference strain NC_000962.3 [125] using minimap2 [177] with the following parameters:

```
# Mapping reads to the reference

minimap2 -ax map-ont $REFERENCE $FASTQ -t20 > $SAM

# Filtering mapped reads by mapping quality

awk '$1 ~ /^@/ || $5 == 60' $SAM > $MAPQ60.sam

# Sorting sam files

samtools sort $MAPQ60.sam > $MAPQ60.sort.sam

# Indexing sam files

samtools view -b ${sample}.MAPQ60.sort.sam -T $REFERENCE >
$MAPQ60.sort.bam
```

8. MinION Variant Calling

Obtention of mpileup files [178]:

```
samtools mpileup -AB -d 1000000 -f $REFERENCE $sort.bam >
$MPILEUP
```

To call SNPs, we execute VarScan [179] on the mpileup files with the following parameters for samples sequenced with MinION: minimum read depth = 50, minimum number of reads supporting a position = 2, minimum base quality at a position to count a read = 10, minimum variant allele frequency threshold = 0.40 and minimum frequency to call homozygote = 0.90.

```
java -Xms10G -Xmx32G -jar VarScan.v2.3.7.jar pileup2snp
$MPILEUP --min-coverage 50 --min-reads2 2 --min-avg-qual
10 --min-freq-for-hom 0.90 --min-var-freq 0.40 > $SNP
```

To call Insertions/Deletions (indels) we execute VarScan [179] on the mpileup files with the following parameters for samples sequenced with MinION:

```
java -Xms10G -Xmx32G -jar VarScan.v2.3.7.jar pileup2indel
$MPILEUP --min-coverage 50 --min-reads2 2 --min-avg-qual
15 --min-freq-for-hom 0.90 --min-var-freq 0.1 --p-value
99e-02 > $INDEL
```

9. Calibration of MinION Variant Calling

Formulas to evaluate True positive rate (TPR), true negative rate (TNR) or recall, precision, agreement, F1-score, accuracy and error rate [172,228]:

$$TruePositiveRate = \frac{TP}{TP+FN}$$

$$TrueNegativeRate = \frac{TN}{TN+FP}$$

$$Precision = \frac{TP}{TP+FP}$$

$$\%Agreement = \frac{TP}{TP+FP+FN} \cdot 100$$

$$F1 - Score = \frac{2 \cdot (TPR \cdot Precision)}{TPR+Precision}$$

$$Accuracy(\%) = \frac{TP+TN}{TP+TN+FP+FN} \cdot 100$$

$$ErrorRateMinION = \frac{FP+FN}{TotalSNPsMinION}$$

References

1. Pai M, Behr MA, Dowdy D, Dheda K, Divangahi M, Boehme CC, et al. Tuberculosis. Nat Rev Dis Primers. 2016;2: 16076.

2. World Health Organization. Global tuberculosis report 2023. World Health Organization; 2023 Nov. Available: https://www.who.int/publications/i/item/9789240083851

3. Abou Jaoude GJ, Garcia Baena I, Nguhiu P, Siroka A, Palmer T, Goscé L, et al. National tuberculosis spending efficiency and its associated factors in 121 low-income and middle-income countries, 2010-19: a data envelopment and stochastic frontier analysis. Lancet Glob Health. 2022;10: e649–e660.

4. Achkar JM, Jenny-Avital ER. Incipient and subclinical tuberculosis: defining early disease states in the context of host immune response. J Infect Dis. 2011;204 Suppl 4: S1179–86.

5. Patterson B, Wood R. Is cough really necessary for TB transmission? Tuberculosis . 2019;117: 31–35.

6. Vynnycky E, Fine PE. The natural history of tuberculosis: the implications of age-dependent risks of disease and the role of reinfection. Epidemiol Infect. 1997;119: 183–201.

7. Behr MA, Edelstein PH, Ramakrishnan L. Revisiting the timetable of tuberculosis. BMJ. 2018;362: k2738.

8. Nguyen HV, Tiemersma E, Nguyen NV, Nguyen HB, Cobelens F. Disease Transmission by Patients With Subclinical Tuberculosis. Clin Infect Dis. 2023;76: 2000–2006.

9. Kradin R l., Deshpande V, Iafrate AJ. 2 - General Principles in the Diagnosis of Infection. In: Kradin RL, editor. Diagnostic Pathology of Infectious Disease (Second Edition). Elsevier; 2018. pp. 3–15.

10. Caulfield AJ, Wengenack NL. Diagnosis of active tuberculosis disease: From microscopy to molecular techniques. J Clin Tuberc Other Mycobact Dis. 2016;4: 33–43.

11. Horne DJ, Royce SE, Gooze L, Narita M, Hopewell PC, Nahid P, et al. Sputum monitoring during tuberculosis treatment for predicting outcome: systematic review and meta-analysis. Lancet Infect Dis. 2010;10: 387–394.

12. Gill CM, Dolan L, Piggott LM, McLaughlin AM. New developments in tuberculosis diagnosis and treatment. Breathe (Sheff). 2022;18: 210149.

13. Pai M, Dewan PK, Swaminathan S. Transforming tuberculosis diagnosis. Nat Microbiol. 2023;8: 756–759.

14. Dorman SE, Schumacher SG, Alland D, Nabeta P, Armstrong DT, King B, et al. Xpert MTB/RIF Ultra for detection of Mycobacterium tuberculosis and rifampicin resistance: a prospective multicentre diagnostic accuracy study. Lancet Infect Dis. 2018;18: 76–84.

15. 3. TB prevention and screening. [cited 17 Jan 2024]. Available: https://www.who.int/teams/global-tuberculosis-programme/tb-reports/global-tuberculosis-report-2023/tb-prevention

16. Sharma SK, Vashishtha R, Chauhan LS, Sreenivas V, Seth D. Comparison of TST and IGRA in Diagnosis of Latent Tuberculosis Infection in a High TB-Burden Setting. PLoS One. 2017;12: e0169539.

17. (gtb) GTP. WHO operational handbook on tuberculosis: module 4: treatment: drug-susceptible tuberculosis treatment. World Health Organization; 24 May 2022 [cited 16 Jan 2024]. Available: https://www.who.int/publications/i/item/9789240050761

18. Vanino E, Granozzi B, Akkerman OW, Munoz-Torrico M, Palmieri F, Seaworth B, et al. Update of drug-resistant tuberculosis treatment guidelines: A turning point. Int J Infect Dis. 2023;130 Suppl 1: S12–S15.

19. Global Plan to end TB 2023-2030. [cited 16 Jan 2024]. Available: https://www.stoptb.org/global-plan-to-end-tb/global-plan-to-end-tb-2023-2030

20. World Health Organization. The end TB strategy. Global Tuberculosis Programme; 2015 Aug. Available: https://www.who.int/publications/i/item/WHO-HTM-TB-2015.19

21. Lönnroth K, Castro KG, Chakaya JM, Chauhan LS, Floyd K, Glaziou P, et al. Tuberculosis control and elimination 2010-50: cure, care, and social development. Lancet. 2010;375: 1814–1829.

22. Kaufmann SHE. Vaccine Development Against Tuberculosis Over the Last 140 Years: Failure as Part of Success. Front Microbiol. 2021;12: 750124.

23. BCG vaccine. React Wkly. 2013;1476: 8–8.

24. Zwerling A, Badar S, Araujo TC, Pennington J, Pai M. BCG world atlas. [cited 17 Jan 2024]. Available: http://www.bcgatlas.org/

25. Mangtani P, Abubakar I, Ariti C, Beynon R, Pimpin L, Fine PEM, et al. Protection by BCG vaccine against tuberculosis: a systematic review of randomized controlled trials. Clin Infect Dis. 2014;58: 470–480.

26. Gong W, Du J, Zhuang L, Wu X. Exploring BCG vaccination as a novel approach to prevent recurrent herpes labialis. EClinicalMedicine. 2023;65: 102279.

27. New TB vaccine research. [cited 17 Jan 2024]. Available: https://www.who.int/teams/global-tuberculosis-programme/research-innovation/vaccines

28. Zentralblatt für Bakteriologie, Parasitenkunde und Infektionskrankheiten: 1 Abt. Medizinisch-hygienische Bakteriologie, Virusforschung und tierische Parasitologie. G. Fischer.; 1896.

29. Meehan CJ, Barco RA, Loh Y-HE, Cogneau S, Rigouts L. Reconstituting the genus Mycobacterium. Int J Syst Evol Microbiol. 2021;71. doi:10.1099/ijsem.0.004922

30. Malone KM, Gordon SV. Mycobacterium tuberculosis Complex Members Adapted to Wild and Domestic Animals. Adv Exp Med Biol. 2017;1019: 135–154.

31. Kolattukudy PE, Fernandes ND, Azad AK, Fitzmaurice AM, Sirakova TD. Biochemistry and molecular genetics of cell-wall lipid biosynthesis in mycobacteria. Mol Microbiol. 1997;24: 263–270.

32. Dulberger CL, Rubin EJ, Boutte CC. The mycobacterial cell envelope - a moving target. Nat Rev Microbiol. 2020;18: 47–59.

33. Gordon SV, Parish T. Microbe Profile: Mycobacterium tuberculosis: Humanity's deadly microbial foe. Microbiology. 2018;164: 437–439.

34. Daffé M, Etienne G. The capsule of Mycobacterium tuberculosis and its implications for pathogenicity. Tuber Lung Dis. 1999;79: 153–169.

35. Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, et al. Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. Nature. 1998;393: 537–544.

36. Coscolla M. Biological and Epidemiological Consequences of MTBC Diversity. Adv Exp Med

Biol. 2017;1019: 95–116.

37. Ngabonziza JCS, Loiseau C, Marceau M, Jouet A, Menardo F, Tzfadia O, et al. A sister lineage of the Mycobacterium tuberculosis complex discovered in the African Great Lakes region. Nat Commun. 2020;11: 2917.

38. Coscolla M, Gagneux S, Menardo F, Loiseau C, Ruiz-Rodriguez P, Borrell S, et al. Phylogenomics of Mycobacterium africanum reveals a new lineage and a complex evolutionary history. Microb Genom. 2021;7. doi:10.1099/mgen.0.000477

39. Bottai D, Frigui W, Sayes F, Di Luca M, Spadoni D, Pawlik A, et al. TbD1 deletion as a driver of the evolutionary success of modern epidemic Mycobacterium tuberculosis lineages. Nat Commun. 2020;11: 684.

40. Brosch R, Gordon SV, Marmiesse M, Brodin P, Buchrieser C, Eiglmeier K, et al. A new evolutionary scenario for the *Mycobacterium tuberculosis* complex. Proc Natl Acad Sci U S A. 2002;99: 3684–3689.

41. Comas I, Coscolla M, Luo T, Borrell S, Holt KE, Kato-Maeda M, et al. Out-of-Africa migration and Neolithic coexpansion of Mycobacterium tuberculosis with modern humans. Nat Genet. 2013;45: 1176–1182.

42. Blouin Y, Hauck Y, Soler C, Fabre M, Vong R, Dehan C, et al. Significance of the identification in the Horn of Africa of an exceptionally deep branching Mycobacterium tuberculosis clade. PLoS One. 2012;7: e52841.

43. Stucki D, Brites D, Jeljeli L, Coscolla M, Liu Q, Trauner A, et al. Mycobacterium tuberculosis lineage 4 comprises globally distributed and geographically restricted sublineages. Nat Genet. 2016;48: 1535–1543.

44. Coll F, McNerney R, Guerra-Assunção JA, Glynn JR, Perdigão J, Viveiros M, et al. A robust SNP barcode for typing Mycobacterium tuberculosis complex strains. Nat Commun. 2014;5: 4812.

45. He C, Cheng X, Kaisaier A, Wan J, Luo S, Ren J, et al. Effects of Mycobacterium tuberculosis lineages and regions of difference (RD) virulence gene variation on tuberculosis recurrence. Annals of Translational Medicine. 2022;10. doi:10.21037/atm-21-6863

46. Ford CB, Lin PL, Chase MR, Shah RR, Iartchouk O, Galagan J, et al. Use of whole genome sequencing to estimate the mutation rate of Mycobacterium tuberculosis during latent infection. Nat Genet. 2011;43: 482–486.

47. Vargas R Jr, Luna MJ, Freschi L, Marin M, Froom R, Murphy KC, et al. Phase variation as a major mechanism of adaptation in Mycobacterium tuberculosis complex. Proc Natl Acad Sci U S A. 2023;120. doi:10.1073/pnas.2301394120

48. Meehan CJ, Goig GA, Kohl TA, Verboven L, Dippenaar A, Ezewudo M, et al. Whole genome sequencing of Mycobacterium tuberculosis: current standards and open issues. Nat Rev Microbiol. 2019;17: 533–545.

49. Marin M, Vargas R, Harris M, Jeffrey B, Epperson LE, Durbin D, et al. Benchmarking the empirical accuracy of short-read sequencing across the M. tuberculosis genome. Bioinformatics. 2022;38: 1781–1787.

50. Di Marco F, Spitaleri A, Battaglia S, Batignani V, Cabibbe AM, Cirillo DM. Advantages of long- and short-reads sequencing for the hybrid investigation of the Mycobacterium tuberculosis genome. Front Microbiol. 2023;14: 1104456.

51. Public Health England. England world leaders in the use of whole genome sequencing to diagnose TB. In: GOV.UK [Internet]. 28 Mar 2017 [cited 4 Mar 2024]. Available: https://www.gov.uk/government/news/england-world-leaders-in-the-use-of-whole-genome-sequencing-to-diagnose-tb

52. (gtb) GTP. WHO consolidated guidelines on tuberculosis: module 3: diagnosis: rapid diagnostics for tuberculosis detection, 3rd ed. World Health Organization; 20 Mar 2024 [cited 2 Apr 2024]. Available: https://www.who.int/publications/i/item/9789240089488

53. Jajou R, de Neeling A, van Hunen R, de Vries G, Schimmel H, Mulder A, et al. Epidemiological links between tuberculosis cases identified twice as efficiently by whole genome sequencing than conventional molecular typing: A population-based study. PLoS One. 2018;13: e0195413.

54. van Embden JD, Cave MD, Crawford JT, Dale JW, Eisenach KD, Gicquel B, et al. Strain identification of Mycobacterium tuberculosis by DNA fingerprinting: recommendations for a standardized methodology. J Clin Microbiol. 1993 [cited 4 Mar 2024]. doi:10.1128/jcm.31.2.406-409.1993

55. Zeng X, Li H, Zheng R, Kurepina N, Kreiswirth BN, Zhao X, et al. Spoligotyping of Mycobacterium tuberculosis Complex Isolates by Use of Ligation-Based Amplification and Melting Curve Analysis. J Clin Microbiol. 2016;54: 2384.

56. Supply P, Mazars E, Lesjean S, Vincent V, Gicquel B, Locht C. Variable human

minisatellite-like regions in the Mycobacterium tuberculosis genome. Mol Microbiol. 2000;36: 762–771.

57. Lieberman TD, Wilson D, Misra R, Xiong LL, Moodley P, Cohen T, et al. Genomic diversity in autopsy samples reveals within-host dissemination of HIV-associated Mycobacterium tuberculosis. Nat Med. 2016;22: 1470–1474.

58. Stucki D, Malla B, Hostettler S, Huna T, Feldmann J, Yeboah-Manu D, et al. Two new rapid SNP-typing methods for classifying Mycobacterium tuberculosis complex into the main phylogenetic lineages. PLoS One. 2012;7: e41253.

59. Walker TM, Camilla LC, Harrell RH, Evans JT, Kapatai G, Dedicoat MJ, et al. Whole-genome sequencing to delineate Mycobacterium tuberculosis outbreaks: a retrospective observational study. Lancet Infect Dis. 2013;13: 137–146.

60. Cancino-Muñoz I, López MG, Torres-Puente M, Villamayor LM, Borrás R, Borrás-Máñez M, et al. Population-based sequencing of Mycobacterium tuberculosis reveals how current population dynamics are shaped by past epidemics. 2022 [cited 4 Mar 2024]. doi:10.7554/eLife.76605

61. Cervera BS, López MG, Chiner-Oms Á, García AM, Cancino-Muñoz I, Torres-Puente M, et al. Fine-grain population structure and transmission patterns of Mycobacterium tuberculosis in southern Mozambique, a high TB/HIV burden area. Microbial Genomics. 2022;8: 000844.

62. Votintseva AA, Bradley P, Pankhurst L, Del Ojo Elias C, Loose M, Nilgiriwala K, et al. Same-Day Diagnostic and Surveillance Data for Tuberculosis via Whole-Genome Sequencing of Direct Respiratory Samples. J Clin Microbiol. 2017;55: 1285–1298.

63. Doyle RM, Burgess C, Williams R, Gorton R, Booth H, Brown J, et al. Direct Whole-Genome Sequencing of Sputum Accurately Identifies Drug-Resistant Mycobacterium tuberculosis Faster than MGIT Culture Sequencing. J Clin Microbiol. 2018;56. doi:10.1128/JCM.00666-18

64. Nimmo C, Doyle R, Burgess C, Williams R, Gorton R, McHugh TD, et al. Rapid identification of a Mycobacterium tuberculosis full genetic drug resistance profile through whole genome sequencing directly from sputum. Int J Infect Dis. 2017;62: 44–46.

65. Goig GA, Cancino-Muñoz I, Torres-Puente M, Villamayor LM, Navarro D, Borrás R, et al. Whole-genome sequencing of Mycobacterium tuberculosis directly from clinical samples for high-resolution genomic epidemiology and drug resistance surveillance: an observational study. Lancet Microbe. 2020;1: e175–e183.

66. Brown AC, Bryant JM, Einer-Jensen K, Holdstock J, Houniet DT, Chan JZM, et al. Rapid Whole-Genome Sequencing of Mycobacterium tuberculosis Isolates Directly from Clinical Samples. J Clin Microbiol. 2015;53: 2230–2237.

67. Ticlla MR, Hella J, Hiza H, Sasamalo M, Mhimbira F, Rutaihwa LK, et al. The Sputum Microbiome in Pulmonary Tuberculosis and Its Association With Disease Manifestations: A Cross-Sectional Study. Front Microbiol. 2021;12. doi:10.3389/fmicb.2021.633396

68. Kent PT, Kubica GP. Public Health Mycobacteriology: A Guide for the Level III Laboratory. 1985.

69. Podbielski, Abele-Horn M, Becker K, Herrmann, Kniehl, Mauch, et al. MIQ 05: Tuberkulose Mykobakteriose: Qualitätsstandards in der mikrobiologisch-infektiologischen Diagnostik. Elsevier Health Sciences; 2019.

70. Doughty EL, Sergeant MJ, Adetifa I, Antonio M, Pallen MJ. Culture-independent detection and characterisation of Mycobacterium tuberculosis and M. africanum in sputum samples using shotgun metagenomics on a benchtop sequencer. PeerJ. 2014;2: e585.

71. Lozano N, Lanza VF, Suárez-González J, Herranz M, Sola-Campoy PJ, Rodríguez-Grande C, et al. Detection of Minority Variants and Mixed Infections in Mycobacterium tuberculosis by Direct Whole-Genome Sequencing on Noncultured Specimens Using a Specific-DNA Capture Strategy. mSphere. 2021;6: e0074421.

72. Sundararaman B, Sylvester MD, Kozyreva VK, Berrada ZL, Corbett-Detig RB, Green RE. A hybridization target enrichment approach for pathogen genomics. MBio. 2023 [cited 5 Mar 2024]. doi:10.1128/mbio.01889-23

73. Goig GA, Blanco S, Garcia-Basteiro AL, Comas I. Contaminant DNA in bacterial sequencing experiments is a major source of false genetic variability. BMC Biol. 2020;18: 1–15.

74. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol. 2014;15: R46.

75. Cuevas-Córdoba B, Fresno C, Haase-Hernández JI, Barbosa-Amezcua M, Mata-Rocha M, Muñoz-Torrico M, et al. A bioinformatics pipeline for Mycobacterium tuberculosis sequencing that cleans contaminant reads from sputum samples. PLoS One. 2021;16: e0258774.

76. Moreno-Molina M, Shubladze N, Khurtsilava I, Avaliani Z, Bablishvili N, Torres-Puente M, et al. Genomic analyses of Mycobacterium tuberculosis from human lung resections reveal a

high frequency of polyclonal infections. Nat Commun. 2021;12: 1–11.

77. Osei-Wusu S, Otchere ID, Morgan P, Musah AB, Siam IM, Asandem D, et al. Genotypic and phenotypic diversity of Mycobacterium tuberculosis complex genotypes prevalent in West Africa. PLoS One. 2021;16. doi:10.1371/journal.pone.0255433

78. Boyce KJ. Mutators Enhance Adaptive Micro-Evolution in Pathogenic Microbes. Microorganisms. 2022;10. doi:10.3390/microorganisms10020442

79. Genestet C, Hodille E, Westeel E, Ginevra C, Ader F, Venner S, et al. Subcultured Mycobacterium tuberculosis isolates on different growth media are fully representative of bacteria within clinical samples. Tuberculosis . 2019;116: 61–66.

80. Shockey AC, Dabney J, Pepperell CS. Effects of Host, Sample, and in vitro Culture on Genomic Diversity of Pathogenic Mycobacteria. Front Genet. 2019;10: 477.

81. Martín A, Herranz M, Ruiz Serrano MJ, Bouza E, García de Viedma D. The clonal composition of Mycobacterium tuberculosis in clinical specimens could be modified by culture. Tuberculosis . 2010;90: 201–207.

82. Koch A, Cox H, Mizrahi V. Drug-resistant tuberculosis: challenges and opportunities for diagnosis and treatment. Curr Opin Pharmacol. 2018;42: 7–15.

83. Lange C, Dheda K, Chesov D, Mandalakas AM, Udwadia Z, Robert Horsburgh C. Management of drug-resistant tuberculosis. Lancet. 2019;394: 953–966.

84. Kim SJ. Drug-susceptibility testing in tuberculosis: methods and reliability of results. Eur Respir J. 2005;25: 564–569.

85. Yusoof KA, García JI, Schami A, Garcia-Vilanova A, Kelley HV, Wang S-H, et al. Tuberculosis Phenotypic and Genotypic Drug Susceptibility Testing and Immunodiagnostics: A Review. Front Immunol. 2022;13. doi:10.3389/fimmu.2022.870768

86. Bemer P, Palicova F, Rüsch-Gerdes S, Drugeon HB, Pfyffer GE. Multicenter evaluation of fully automated BACTEC Mycobacteria Growth Indicator Tube 960 system for susceptibility testing of Mycobacterium tuberculosis. J Clin Microbiol. 2002;40: 150–154.

87. Lee J, Armstrong DT, Ssengooba W, Park J-A, Yu Y, Mumbowa F, et al. Sensititre MYCOTB MIC Plate for Testing Mycobacterium tuberculosis Susceptibility to First- and Second-Line Drugs. Antimicrob Agents Chemother. 2014;58: 11.

88. Palomino J-C, Martin A, Camacho M, Guerra H, Swings J, Portaels F. Resazurin Microtiter Assay Plate: Simple and Inexpensive Method for Detection of Drug Resistance in Mycobacterium tuberculosis. Antimicrob Agents Chemother. 2002 [cited 11 Mar 2024]. doi:10.1128/aac.46.8.2720-2722.2002

89. Cirillo DM, Miotto P, Tortoli E. Evolution of Phenotypic and Molecular Drug Susceptibility Testing. Strain Variation in the Mycobacterium tuberculosis Complex: Its Role in Biology, Epidemiology and Control. 2017; 221–246.

90. Weltgesundheitsorganisation. WHO consolidated guidelines on tuberculosis. Module 3: diagnosis - rapid diagnostics for tuberculosis detection, 2021 update. World Health Organization; 2021.

91. Opota O, Mazza-Stalder J, Greub G, Jaton K. The rapid molecular test Xpert MTB/RIF ultra: towards improved tuberculosis diagnosis and rifampicin resistance detection. Clin Microbiol Infect. 2019;25: 1370–1376.

92. Cancino-Muñoz I, Moreno-Molina M, Furió V, Goig GA, Torres-Puente M, Chiner-Oms Á, et al. Cryptic Resistance Mutations Associated With Misdiagnoses of Multidrug-Resistant Tuberculosis. J Infect Dis. 2019;220: 316–320.

93. Beckert P, Sanchez-Padilla E, Merker M, Dreyer V, Kohl TA, Utpatel C, et al. MDR M. tuberculosis outbreak clone in Eswatini missed by Xpert has elevated bedaquiline resistance dated to the pre-treatment era. Genome Med. 2020;12: 104.

94. Makhado NA, Matabane E, Faccin M, Pinçon C, Jouet A, Boutachkourt F, et al. Outbreak of multidrug-resistant tuberculosis in South Africa undetected by WHO-endorsed commercial tests: an observational study. Lancet Infect Dis. 2018;18: 1350–1359.

95. Modongo C, Barilar I, Wang Q, Molefi T, Makhondo T, Niemann S, et al. Tuberculosis Variant with Rifampin Resistance Undetectable by Xpert MTB/RIF, Botswana. Emerg Infect Dis. 2023;29: 2403–2406.

96. Katamba A, Ssengooba W, Sserubiri J, Semugenze D, William KG, Abdunoor N, et al. Evaluation of Xpert ® MTB/XDR test for susceptibility testing of Mycobacterium tuberculosis to first and second-line drugs in Uganda. medRxiv. 2023. doi:10.1101/2023.04.03.23288099

97. Cao Y, Parmar H, Gaur RL, Lieu D, Raghunath S, Via N, et al. Xpert MTB/XDR: a 10-Color Reflex Assay Suitable for Point-of-Care Settings To Detect Isoniazid, Fluoroquinolone, and

Second-Line-Injectable-Drug Resistance Directly from Mycobacterium tuberculosis-Positive Sputum. J Clin Microbiol. 2021;59. doi:10.1128/JCM.02314-20

98. Bainomugisa A, Gilpin C, Coulter C, Marais BJ. New Xpert MTB/XDR: added value and future in the field. Eur Respir J. 2020;56. doi:10.1183/13993003.03616-2020

99. LifeScience H. GenoType MTBDRsl VER 2.0 instructions for use. Germany: Hain Lifescience. 2015.

100. Bai Y, Wang Y, Shao C, Hao Y, Jin Y. GenoType MTBDRplus Assay for Rapid Detection of Multidrug Resistance in Mycobacterium tuberculosis: A Meta-Analysis. PLoS One. 2016;11: e0150321.

101. Eddabra R, Ait Benhassou H. Rapid molecular assays for detection of tuberculosis. Pneumonia (Nathan). 2018;10: 4.

102. Cabibbe AM, Spitaleri A, Battaglia S, Colman RE, Suresh A, Uplekar S, et al. Application of Targeted Next-Generation Sequencing Assay on a Portable Sequencing Platform for Culture-Free Detection of Drug-Resistant Tuberculosis from Clinical Samples. J Clin Microbiol. 2020;58. doi:10.1128/JCM.00632-20

103. Jouet A, Gaudin C, Badalato N, Allix-Béguec C, Duthoy S, Ferré A, et al. Deep amplicon sequencing for culture-free prediction of susceptibility or resistance to 13 anti-tuberculous drugs. Eur Respir J. 2021;57. doi:10.1183/13993003.02338-2020

104. Walker TM, Miotto P, Köser CU, Fowler PW, Knaggs J, Iqbal Z, et al. The 2021 WHO catalogue of Mycobacterium tuberculosis complex mutations associated with drug resistance: A genotypic analysis. Lancet Microbe. 2022;3: e265–e273.

105. (gtb) GTP. Catalogue of mutations in Mycobacterium tuberculosis complex and their association with drug resistance, 2nd ed. World Health Organization; 15 Nov 2023 [cited 4 Mar 2024]. Available: https://www.who.int/publications/i/item/9789240082410

106. Brankin A, Seifert M, Georghiou SB, Walker TM, Uplekar S, Suresh A, et al. In silico evaluation of WHO-endorsed molecular methods to detect drug resistant tuberculosis. Sci Rep. 2022;12: 17741.

107. Köser CU, Cirillo DM, Miotto P. How To Optimally Combine Genotypic and Phenotypic Drug Susceptibility Testing Methods for Pyrazinamide. Antimicrob Agents Chemother. 2020;64. doi:10.1128/AAC.01003-20

108. Coeck N, de Jong BC, Diels M, de Rijk P, Ardizzoni E, Van Deun A, et al. Correlation of different phenotypic drug susceptibility testing methods for four fluoroquinolones in Mycobacterium tuberculosis. J Antimicrob Chemother. 2016;71: 1233–1240.

109. García-Marín AM, Cancino-Muñoz I, Torres-Puente M, Villamayor LM, Borrás R, Borrás-Máñez M, et al. Role of the first WHO mutation catalogue in the diagnosis of antibiotic resistance in Mycobacterium tuberculosis in the Valencia Region, Spain: a retrospective genomic analysis. Lancet Microbe. 2024;5: e43–e51.

110. Gagneux S. Ecology and evolution of Mycobacterium tuberculosis. Nat Rev Microbiol. 2018;16: 202–213.

111. Gagneux S, Long CD, Small PM, Van T, Schoolnik GK, Bohannan BJM. The competitive cost of antibiotic resistance in Mycobacterium tuberculosis. Science. 2006;312: 1944–1946.

112. Miotto P, Cabibbe AM, Borroni E, Degano M, 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. J Clin Microbiol. 2018;56. doi:10.1128/JCM.01599-17

113. Gehre F, Otu J, DeRiemer K, de Sessions PF, Hibberd ML, Mulders W, et al. Deciphering the growth behaviour of Mycobacterium africanum. PLoS Negl Trop Dis. 2013;7: e2220.

114. Dhillon J, Fourie PB, Mitchison DA. Persister populations of Mycobacterium tuberculosis in sputum that grow in liquid but not on solid culture media. J Antimicrob Chemother. 2014;69: 437–440.

115. Mohamed S, Köser CU, Salfinger M, Sougakoff W, Heysell SK. Targeted next-generation sequencing: a Swiss army knife for mycobacterial diagnostics? Eur Respir J. 2021;57. doi:10.1183/13993003.04077-2020

116. Eshetie S, van Soolingen D. The respiratory microbiota: new insights into pulmonary tuberculosis. BMC Infect Dis. 2019;19: 92.

117. Nilgiriwala K, Rabodoarivelo M-S, Hall MB, Patel G, Mandal A, Mishra S, et al. Genomic Sequencing from Sputum for Tuberculosis Disease Diagnosis, Lineage Determination, and Drug Susceptibility Prediction. J Clin Microbiol. 2023;61: e0157822.

118. Nimmo C, Shaw LP, Doyle R, Williams R, Brien K, Burgess C, et al. Whole genome sequencing Mycobacterium tuberculosis directly from sputum identifies more genetic diversity than sequencing from culture. BMC Genomics. 2019;20: 389.

119. Goossens SN, Heupink TH, De Vos E, Dippenaar A, De Vos M, Warren R, et al. Detection of minor variants in Mycobacterium tuberculosis whole genome sequencing data. Brief Bioinform. 2022;23. doi:10.1093/bib/bbab541

120. Kubica GP, Dye WE, Cohn ML, Middlebrook G. Sputum Digestion and Decontamination with N-Acetyl-L-Cysteine–Sodium Hydroxide for Culture of Mycobacteria. Am Rev Respir Dis. 1963;87: 775–779.

121. Tripathi K, Tripathi PC, Nema S, Shrivastava AK, Dwiwedi K, Dhanvijay AK. Modified Petroff's method: an excellent simplified decontamination technique in comparison with Petroff's method. Int J Recent Trends Sci Technol. 2014;10: 461–464.

122. Somerville W, Thibert L, Schwartzman K, Behr MA. Extraction of Mycobacterium tuberculosis DNA: a question of containment. J Clin Microbiol. 2005;43: 2996–2997.

123. Goig GA, Torres-Puente M, Mariner-Llicer C, Villamayor LM, Chiner-Oms Á, Gil-Brusola A, et al. Towards next-generation diagnostics for tuberculosis: identification of novel molecular targets by large-scale comparative genomics. Bioinformatics. 2020;36: 985–989.

124. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018;34: i884–i890.

125. Comas I, Chakravartti J, Small PM, Galagan J, Niemann S, Kremer K, et al. Human T cell epitopes of Mycobacterium tuberculosis are evolutionarily hyperconserved. Nat Genet. 2010;42: 498–503.

126. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics. 2010;26: 589–595.

127. Koboldt DC, Chen K, Wylie T, Larson DE, McLellan MD, Mardis ER, et al. VarScan: variant detection in massively parallel sequencing of individual and pooled samples. Bioinformatics. 2009;25: 2283–2285.

128. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. Gigascience. 2021;10. doi:10.1093/gigascience/giab008

129. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20: 1297–1303.

130. Structural variants and the SAM format - the long (reads) and short (reads) of it. [cited 22

Nov 2023]. Available: https://cmdcolin.github.io/posts/2022-02-06-sv-sam

131. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med . 2012;22: 276–282.

132. Paradis E, Schliep K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics. 2019. pp. 526–528. doi:10.1093/bioinformatics/bty633

133. Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. Molecular Biology and Evolution. 2018. pp. 1547–1549. doi:10.1093/molbev/msy096

134. Mann BC, Jacobson KR, Ghebrekristos Y, Warren RM, Farhat MR. Assessment and validation of enrichment and target capture approaches to improve Mycobacterium tuberculosis WGS from direct patient samples. J Clin Microbiol. 2023;61: e0038223.

135. Use of targeted next-generation sequencingto detect drug-resistant tuberculosis. [cited 28 Jul 2023]. Available: https://www.who.int/publications/i/item/9789240076372

136. Van Deun A, Decroo T, Aung KJM, Hossain MA, Gumusboga M, De Rijk WB, et al. Mycobacterium tuberculosis borderline rpoB mutations: emerging from the unknown. Eur Respir J. 2021;58. doi:10.1183/13993003.00783-2021

137. Pandey P, Bhatnagar AK, Mohan A, Sachdeva KS, Samantaray JC, Guleria R, et al. Mycobacterium tuberculosis polyclonal infections through treatment and recurrence. PLoS One. 2020;15: e0237345.

138. Mukamolova GV, Turapov O, Malkin J, Woltmann G, Barer MR. Resuscitation-promoting factors reveal an occult population of tubercle Bacilli in Sputum. Am J Respir Crit Care Med. 2010;181: 174–180.

139. Chengalroyen MD, Beukes GM, Gordhan BG, Streicher EM, Churchyard G, Hafner R, et al. Detection and Quantification of Differentially Culturable Tubercle Bacteria in Sputum from Patients with Tuberculosis. Am J Respir Crit Care Med. 2016;194: 1532–1540.

140. Samo Gudo P, Cuna Z, Coelho E, Maungate S, Borroni E, Miotto P, et al. Is multidrug-resistant tuberculosis on the rise in Mozambique? Results of a national drug resistance survey. Eur Respir J. 2011;38: 222–224.

141. Chakravorty S, Simmons AM, Rowneki M, Parmar H, Cao Y, Ryan J, et al. The New Xpert MTB/RIF Ultra: Improving Detection of Mycobacterium tuberculosis and Resistance to Rifampin in an Assay Suitable for Point-of-Care Testing. MBio. 2017;8.

doi:10.1128/mBio.00812-17

142. Valencia S, Respeito D, Blanco S, Ribeiro RM, López-Varela E, Sequera VG, et al. Tuberculosis drug resistance in Southern Mozambique: results of a population-level survey in the district of Manhiça. Int J Tuberc Lung Dis. 2017;21: 446–451.

143. García-Basteiro AL, Miranda Ribeiro R, Brew J, Sacoor C, Valencia S, Bulo H, et al. Tuberculosis on the rise in southern Mozambique (1997-2012). Eur Respir J. 2017;49. doi:10.1183/13993003.01683-2016

144. Bernardo EL, Nhampossa T, Clouse K, Carlucci JG, Fernández-Luis S, Fuente-Soro L, et al. Patterns of mobility and its impact on retention in care among people living with HIV in the Manhiça District, Mozambique. PLoS One. 2021;16: e0250844.

145. Kadura S, King N, Nakhoul M, Zhu H, Theron G, Köser CU, et al. Systematic review of mutations associated with resistance to the new and repurposed Mycobacterium tuberculosis drugs bedaquiline, clofazimine, linezolid, delamanid and pretomanid. J Antimicrob Chemother. 2020;75: 2031–2043.

146. Roberts LW, Malone KM, Hunt M, Joseph L, Wintringer P, Knaggs J, et al. Repeated evolution of bedaquiline resistance in Mycobacterium tuberculosis is driven by truncation of mmpR5. bioRxiv. 2022. p. 2022.12.08.519610. doi:10.1101/2022.12.08.519610

147. Walker TM, Lalor MK, Broda A, Ortega LS, Morgan M, Parker L, et al. Assessment of Mycobacterium tuberculosis transmission in Oxfordshire, UK, 2007-12, with whole pathogen genome sequences: an observational study. Lancet Respir Med. 2014;2: 285–292.

148. Feliciano CS, Namburete EI, Rodrigues Plaça J, Peronni K, Dippenaar A, Warren RM, et al. Accuracy of whole genome sequencing versus phenotypic (MGIT) and commercial molecular tests for detection of drug-resistant Mycobacterium tuberculosis isolated from patients in Brazil and Mozambique. Tuberculosis . 2018;110: 59–67.

149. Sulis G, Pai M. Isoniazid-resistant tuberculosis: A problem we can no longer ignore. PLoS Med. 2020;17: e1003023.

150. World Health Organization. Global Tuberculosis Report 2019. 2019.

151. Hoang TTT, Nguyen NV, Dinh SN, Nguyen HB, Cobelens F, Thwaites G, et al. Challenges in detection and treatment of multidrug resistant tuberculosis patients in Vietnam. BMC Public Health. 2015;15: 1–10.

152. World Health Organization. Global tuberculosis control: a short update to the 2009 report. World Health Organization; 2009. Report No.: WHO/HTM/TB/2009.426.

153. Lee A, Xie YL, Barry CE, Chen RY. Current and future treatments for tuberculosis. BMJ. 2020;368. doi:10.1136/bmj.m216

154. Marks SM, Hirsch-Moverman Y, Salcedo K, Graviss EA, Oh P, Seaworth B, et al. Characteristics and costs of multidrug-resistant tuberculosis in-patient care in the United States, 2005–2007. 2016 [cited 15 Dec 2020]. doi:10.5588/ijtld.15.0575

155. World Health Organization. Rapid communication: key changes to treatment of multidrug-and rifampicin-resistant tuberculosis (MDR/RR-TB). World Health Organization; 2019. doi:WHO reference number: WHO/CDS/TB/2019.26

156. Roelens M, Battista Migliori G, Rozanova L, Estill J, Campbell JR, Cegielski JP, et al. Evidence-based Definition for Extensively Drug-resistant Tuberculosis. Am J Respir Crit Care Med. 2021. doi:10.1164/rccm.202009-3527OC

157. Mirzayev F, Viney K, Linh NN, Gonzalez-Angulo L, Gegia M, Jaramillo E, et al. World Health Organization recommendations on the treatment of drug-resistant tuberculosis, 2020 update. Eur Respir J. 2021;57. doi:10.1183/13993003.03300-2020

158. The Mycobacterium avium Complex and Slowly Growing Mycobacteria. Molecular Medical Microbiology. Academic Press; 2015. pp. 1669–1678.

159. World Health Organization. The use of molecular line probe assay for the detection of resistance to isoniazid and rifampicin: policy update. World Health Organization; 2016.

160. Mäkinen J, Marttila HJ, Marjamäki M, Viljanen MK, Soini H. Comparison of two commercially available DNA line probe assays for detection of multidrug-resistant Mycobacterium tuberculosis. J Clin Microbiol. 2006;44: 350–352.

161. Ninan MM, Gowri M, Christopher DJ, Rupali P, Michael JS. The diagnostic utility of line probe assays for multidrug-resistant tuberculosis. Pathog Glob Health. 2016;110: 194–199.

162. Osei Sekyere J, Maphalala N, Malinga LA, Mbelle NM, Maningi NE. A Comparative Evaluation of the New Genexpert MTB/RIF Ultra and other Rapid Diagnostic Assays for Detecting Tuberculosis in Pulmonary and Extra Pulmonary Specimens. Sci Rep. 2019;9: 16587.

163. Bunsow E, Ruiz-Serrano MJ, López Roa P, Kestler M, Viedma DG, Bouza E. Evaluation of GeneXpert MTB/RIF for the detection of Mycobacterium tuberculosis and resistance to

rifampin in clinical specimens. J Infect. 2014;68: 338–343.

164. Sanchez-Padilla E, Merker M, Beckert P, Jochims F, Dlamini T, Kahn P, et al. Detection of drug-resistant tuberculosis by Xpert MTB/RIF in Swaziland. N Engl J Med. 2015;372: 1181–1182.

165. López B, Aguilar D, Orozco H, Burger M, Espitia C, Ritacco V, et al. A marked difference in pathogenesis and immune response induced by different Mycobacterium tuberculosis genotypes. Clin Exp Immunol. 2003;133: 30–37.

166. Ford CB, Shah RR, Maeda MK, Gagneux S, Murray MB, Cohen T, et al. Mycobacterium tuberculosis mutation rate estimates from different lineages predict substantial differences in the emergence of drug-resistant tuberculosis. Nat Genet. 2013;45: 784–790.

167. Colman RE, Schupp JM, Hicks ND, Smith DE, Buchhagen JL, Valafar F, et al. Detection of Low-Level Mixed-Population Drug Resistance in Mycobacterium tuberculosis Using High Fidelity Amplicon Sequencing. PLoS One. 2015;10: e0126626.

168. Rinder H, Mieskes KT, Löscher T. Heteroresistance in Mycobacterium tuberculosis. Int J Tuberc Lung Dis. 2001;5: 339–345.

169. Magi A, Semeraro R, Mingrino A, Giusti B, D'Aurizio R. Nanopore sequencing data analysis: state of the art, applications and challenges. Brief Bioinform. 2018;19: 1256–1272.

170. Xu Y, Lewandowski K, Lumley S, Pullan S, Vipond R, Carroll M, et al. Detection of Viral Pathogens With Multiplex Nanopore MinION Sequencing: Be Careful With Cross-Talk. Front Microbiol. 2018;9: 2225.

171. Schmidt K, Mwaigwisya S, Crossman LC, Doumith M, Munroe D, Pires C, et al. Identification of bacterial pathogens and antimicrobial resistance directly from clinical urines by nanopore-based metagenomic sequencing. J Antimicrob Chemother. 2017;72: 104–114.

172. Bainomugisa A, Duarte T, Lavu E, Pandey S, Coulter C, Marais BJ, et al. A complete high-quality MinION nanopore assembly of an extensively drug-resistant Mycobacterium tuberculosis Beijing lineage strain identifies novel variation in repetitive PE/PPE gene regions. Microb Genom. 2018;4. doi:10.1099/mgen.0.000188

173. Public health teams in Madagascar pioneer use of portable, real time DNA sequencing to fight drug-resistant tuberculosis (TB). [cited 15 Dec 2020]. Available: http://nanoporetech.com/about-us/news/public-health-teams-madagascar-pioneer-use-portable-real-time-dna-sequencing-fight

174. Tyler AD, Mataseje L, Urfano CJ, Schmidt L, Antonation KS, Mulvey MR, et al. Evaluation of Oxford Nanopore's MinION Sequencing Device for Microbial Whole Genome Sequencing Applications. Sci Rep. 2018;8: 10931.

175. Cancino-Muñoz I, Gil-Brusola A, Torres-Puente M, Mariner-Llicer C, Dogba J, Akinseye V, et al. Development and application of affordable SNP typing approaches to genotype Mycobacterium tuberculosis complex strains in low and high burden countries. Sci Rep. 2019;9: 15343.

176. Wick RR, Judd LM, Holt KE. Performance of neural network basecalling tools for Oxford Nanopore sequencing. Genome Biol. 2019;20: 129.

177. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34: 3094–3100.

178. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25: 2078–2079.

179. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, et al. VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. 2012;22: 568–576.

180. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics. 2009;25: 1754–1760.

181. Feuerriegel S, Schleusener V, Beckert P, Kohl TA, Miotto P, Cirillo DM, et al. PhyResSE: a Web Tool Delineating Mycobacterium tuberculosis Antibiotic Resistance and Lineage from Whole-Genome Sequencing Data. J Clin Microbiol. 2015;53: 1908–1914.

182. ReSeqTB. [cited 21 Jan 2021]. Available: https://platform.reseqtb.org/main/home.html

183. Minervini CF, Cumbo C, Orsini P, Anelli L, Zagaria A, Specchia G, et al. Nanopore Sequencing in Blood Diseases: A Wide Range of Opportunities. Frontiers in Genetics. 2020. doi:10.3389/fgene.2020.00076

184. Orsini P, Minervini CF, Cumbo C, Anelli L, Zagaria A, Minervini A, et al. Design and MinION testing of a nanopore targeted gene sequencing panel for chronic lymphocytic leukemia. Sci Rep. 2018;8: 11798.

185. Ammar R, Paton TA, Torti D, Shlien A, Bader GD. Long read nanopore sequencing for detection of and variants and haplotypes. F1000Res. 2015;4: 17.

186. Tafess K, Ng TTL, Lao HY, Leung KSS, Tam KKG, Rajwani R, et al. Targeted-Sequencing Workflows for Comprehensive Drug Resistance Profiling of Mycobacterium tuberculosis Cultures Using Two Commercial Sequencing Platforms: Comparison of Analytical and Diagnostic Performance, Turnaround Time, and Cost. Clin Chem. 2020;66: 809–820.

187. Organization WH, Others. Meeting report of the WHO expert consultation on the definition of extensively drug-resistant tuberculosis, 27-29 October 2020. 2021. Available: https://apps.who.int/iris/bitstream/handle/10665/338776/9789240018662-eng.pdf

188. Greig DR, Jenkins C, Gharbia S, Dallman TJ. Comparison of single-nucleotide variants identified by Illumina and Oxford Nanopore technologies in the context of a potential outbreak of Shiga toxin–producing Escherichia coli. Gigascience. 2019;8. doi:10.1093/gigascience/giz104

189. Chan WS, Au CH, Chung Y, Leung HCM, Ho DN, Wong EYL, et al. Rapid and economical drug resistance profiling with Nanopore MinION for clinical specimens with low bacillary burden of Mycobacterium tuberculosis. BMC Res Notes. 2020;13: 444.

190. Vargas R, Freschi L, Marin M, Epperson LE, Smith M, Oussenko I, et al. In-host population dynamics of complex during active disease. Elife. 2021;10. doi:10.7554/eLife.61805

191. Rigouts L, Miotto P, Schats M, Lempens P, Cabibbe AM, Galbiati S, et al. Fluoroquinolone heteroresistance in Mycobacterium tuberculosis: detection by genotypic and phenotypic assays in experimentally mixed populations. Sci Rep. 2019;9: 11760.

192. Karst SM, Ziels RM, Kirkegaard RH, Sørensen EA, McDonald D, Zhu Q, et al. High-accuracy long-read amplicon sequences using unique molecular identifiers with Nanopore or PacBio sequencing. Nat Methods. 2021;18. doi:10.1038/s41592-020-01041-y

193. Lei Y, Wang J, Wang Y, Xu C. Geographical evolutionary pathway of global tuberculosis incidence trends. BMC Public Health. 2023;23: 755.

194. Heupink TH, Verboven L, Sharma A, Rennie V, de Diego Fuertes M, Warren RM, et al. The MAGMA pipeline for comprehensive genomic analyses of clinical Mycobacterium tuberculosis samples. PLoS Comput Biol. 2023;19: e1011648.

195. Faye LM, Hosu MC, Oostvogels S, Dippenaar A, Warren RM, Sineke N, et al. The Detection of Mutations and Genotyping of Drug-Resistant Mycobacterium tuberculosis Strains Isolated from Patients in the Rural Eastern Cape Province. Infect Dis Rep. 2023;15: 403–416.

196. Andersson DI, Nicoloff H, Hjort K. Mechanisms and clinical relevance of bacterial

heteroresistance. Nat Rev Microbiol. 2019;17: 479–496.

197. Mon AS, Ei PW, Htwe MM, Nyunt MH, Win SM, Nyunt WW, et al. First Detection of Mycobacterium tuberculosis Clinical Isolates Harboring I491F Borderline Resistance rpoB Mutation in Myanmar. Antimicrob Agents Chemother. 2022;66: e0092522.

198. Siu GKH, Zhang Y, Lau TCK, Lau RWT, Ho P-L, Yew W-W, et al. Mutations outside the rifampicin resistance-determining region associated with rifampicin resistance in Mycobacterium tuberculosis. J Antimicrob Chemother. 2011;66: 730–733.

199. Barilar I, Fernando T, Utpatel C, Abujate C, Madeira CM, José B, et al. Emergence of bedaquiline-resistant tuberculosis and of multidrug-resistant and extensively drug-resistant Mycobacterium tuberculosis strains with rpoB Ile491Phe mutation not detected by Xpert MTB/RIF in Mozambique: a retrospective observational study. Lancet Infect Dis. 2024;24: 297–307.

200. Watsa K, Cossa H, Manhiça INA, Smines L. Critical health screening for Mozambique miners crossing the border into South Africa. In: World Bank Blogs [Internet]. World Bank Group; [cited 28 Mar 2024]. Available: https://blogs.worldbank.org/en/africacan/critical-health-screening-mozambique-miners-crossing-border-south-africa

201. Dean AS, Zignol M, Cabibbe AM, Falzon D, Glaziou P, Cirillo DM, et al. Prevalence and genetic profiles of isoniazid resistance in tuberculosis patients: A multicountry analysis of cross-sectional data. PLoS Med. 2020;17: e1003008.

202. Hassounah IA, Shehata NA, Kimsawatde GC, Hudson AG, Sriranganathan N, Joseph EG, et al. Chapter 11 - Designing and testing single tablet for tuberculosis treatment through electrospinning. In: Grumezescu AM, editor. Fabrication and Self-Assembly of Nanobiomaterials. William Andrew Publishing; 2016. pp. 335–365.

203. Cohen KA, Manson AL, Desjardins CA, Abeel T, Earl AM. Deciphering drug resistance in Mycobacterium tuberculosis using whole-genome sequencing: progress, promise, and challenges. Genome Med. 2019;11: 45.

204. 6. TB research and innovation. [cited 4 Apr 2024]. Available: https://www.who.int/teams/global-tuberculosis-programme/tb-reports/global-tuberculosis-report-2023/tb-research-and-innovation

205. DeepChek® Assay 13-Plex KB Drug Suscptibility Testing V1. In: ABL Diagnostics S.A

[Internet]. [cited 4 Apr 2024]. Available: https://www.abldiagnostics.com/products/deepchek-assay-13-plex-kb-drug-suscptibility-testing-v1/

206. Ni Y, Liu X, Simeneh ZM, Yang M, Li R. Benchmarking of Nanopore R10.4 and R9.4.1 flow cells in single-cell whole-genome amplification and whole-genome shotgun sequencing. Comput Struct Biotechnol J. 2023;21: 2352–2364.

207. Su J, Lui WW, Lee Y, Zheng Z, Siu GK-H, Ng TT-L, et al. Evaluation of Mycobacterium tuberculosis enrichment in metagenomic samples using ONT adaptive sequencing and amplicon sequencing for identification and variant calling. Sci Rep. 2023;13: 5237.

208. Yu G, Shen Y, Yao L, Xu X. Evaluation of Nanopore Sequencing for Diagnosing Pulmonary Tuberculosis Using Negative Smear Clinical Specimens. Infect Drug Resist. 2024;17: 673–682.

209. Walker TM. Web Annex D. 18. Review of the diagnostic accuracy of targeted next-generation sequencing technologies for detection of drug resistance among people diagnosed with TB. WHO consolidated guidelines on tuberculosis Module 3: diagnosis--rapid diagnostics for tuberculosis detection. 2024; 508.

210. de Araujo L, Cabibbe AM, Mhuulu L, Ruswa N, Dreyer V, Diergaardt A, et al. Implementation of targeted next-generation sequencing for the diagnosis of drug-resistant tuberculosis in low-resource settings: a programmatic model, challenges, and initial outcomes. Front Public Health. 2023;11: 1204064.

211. Seq&treat. In: FIND [Internet]. 11 Dec 2022 [cited 4 Apr 2024]. Available: https://www.finddx.org/what-we-do/projects/seqtreat/

212. Mongan AE, Tuda JSB, Runtuwene LR. Portable sequencer in the fight against infectious disease. J Hum Genet. 2020;65: 35–40.

213. Nanopore. VolTRAX: a versatile, programmable, portable device for multiplexed amplification and library preparation. Available: https://nanoporetech.com/sites/default/files/s3/posters/pdf/VolTRAX%20A3.pdf

214. Rekimoto J. Traxion. ACM SIGGRAPH 2014 Emerging Technologies. New York, NY, USA: ACM; 2014. doi:10.1145/2614066.2614079

215. Wang Y, Zhao Y, Bollas A, Wang Y, Au KF. Nanopore sequencing technology, bioinformatics and applications. Nat Biotechnol. 2021;39: 1348–1365.

216. Kovaka S, Fan Y, Ni B, Timp W, Schatz MC. Targeted nanopore sequencing by real-time mapping of raw electrical signal with UNCALLED. Nat Biotechnol. 2021;39: 431–441.

217. Individual, Family Health. Instructions for collecting sputum for TB (tuberculosis). [cited 5 Apr 2024]. Available: https://www.health.state.mn.us/diseases/tb/basics/factsheets/sputum.html

218. Weiszhar Z, Horvath I. Induced sputum analysis: step by step. Breathe. 2013;9: 300–306.

219. Gunasekera KS, Vonasek B, Oliwa J, Triasih R, Lancioni C, Graham SM, et al. Diagnostic Challenges in Childhood Pulmonary Tuberculosis-Optimizing the Clinical Approach. Pathogens. 2022;11. doi:10.3390/pathogens11040382

220. Broger T, Koeppel L, Huerga H, Miller P, Gupta-Wright A, Blanc F-X, et al. Diagnostic yield of urine lipoarabinomannan and sputum tuberculosis tests in people living with HIV: a systematic review and meta-analysis of individual participant data. Lancet Glob Health. 2023;11: e903–e916.

221. Nathavitharana RR, Garcia-Basteiro AL, Ruhwald M, Cobelens F, Theron G. Reimagining the status quo: How close are we to rapid sputum-free tuberculosis diagnostics for all? EBioMedicine. 2022;78: 103939.

222. Organization WH, Others. The use of lateral flow urine lipoarabinomannan assay (LF-LAM) for the diagnosis and screening of active tuberculosis in people living with HIV: policy guidance. World Health Organization; 2015. Available: https://apps.who.int/iris/bitstream/handle/10665/193633/9789241509633_eng.pdf

223. MacLean E, Sulis G, Denkinger CM, Johnston JC, Pai M, Ahmad Khan F. Diagnostic Accuracy of Stool Xpert MTB/RIF for Detection of Pulmonary Tuberculosis in Children: a Systematic Review and Meta-analysis. J Clin Microbiol. 2019;57. doi:10.1128/JCM.02057-18

224. Williams CM, Abdulwhhab M, Birring SS, De Kock E, Garton NJ, Townsend E, et al. Exhaled Mycobacterium tuberculosis output and detection of subclinical disease by face-mask sampling: prospective observational studies. Lancet Infect Dis. 2020;20: 607–617.

225. Vogel M, Utpatel C, Corbett C, Kohl TA, Iskakova A, Ahmedov S, et al. Implementation of whole genome sequencing for tuberculosis diagnostics in a low-middle income, high MDR-TB burden country. Sci Rep. 2021;11: 1–9.

226. Mariner-Llicer C, Goig GA, Torres-Puente M, Vashakidze S, Villamayor LM, Saavedra-Cervera B, et al. Genetic diversity within diagnostic sputum samples is mirrored in the culture of

Mycobacterium tuberculosis across different settings. Nat Commun. 2024;15: 7114.

227. Buschmann T. DNABarcodes: an R package for the systematic construction of DNA sample tags. Bioinformatics. 2017; btw759.

228. Lalkhen AG, McCluskey A. Clinical tests: sensitivity and specificity. Contin Educ Anaesth Crit Care Pain. 2008;8: 221–223.