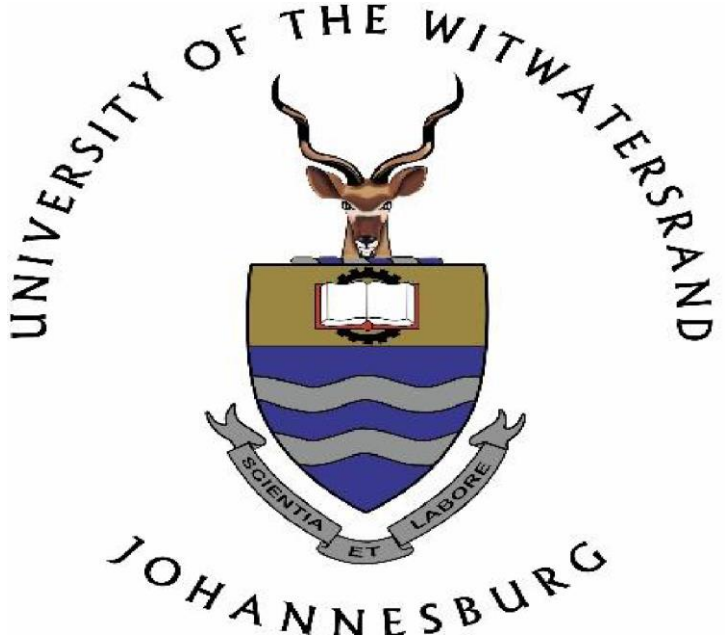
Understanding mycobacterial genome diversity using a cultureomics approach

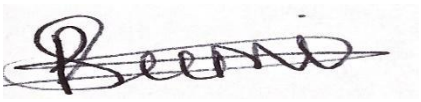
Palesa Seemi (1433351)



A dissertation submitted to the Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, in fulfilment for the degree of Master of Science in Medicine 2022.

DECLARATION

I, Palesa Seemi (student number: 1433351) declare that this dissertation is my own work. It is being submitted for the degree of Master of Science in Medicine at the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination at this or any other University.



15/06/22

Palesa Seemi **Date**

ACKNOWLEDGEMENTS

If it had not been for the Lord who was always on my side, I would have never made it. God, you are worthy of the praise.

To my family, your endless love, support, and guidance have been my anchor. I cannot begin to express my love and gratitude to you all. Mbali 'Mbee' Twala, Sipho 'Sigabops' Masangane and Sechaba 'Dinks' Mazibuko, thank you my friends.

I would like to thank Prof. Bavesh Kana for granting me the opportunity to be a part of the Centre of Excellence for Biomedical Tuberculosis Research (CBTBR) lab. To my supervisors, Dr. Bhavna Gordhan, and Dr. Julian Peters, thank you for the supervision throughout the years. To Dr Nikki Gentle, thank you so much for your willingness to accept me as one of your own students.

Thank you so much to my CBTBR family, Dale Liebenberg, Ofentse Matlhabe, Moagi Shaku, and Christopher Ealand, for always being willing to help me with my experimental work when needed. Your kind hearts are something I will never forget about you all. My fellow MSc friends, WE DID IT! I am thankful to have gone through this journey with people like you. I will forever hold you dear to my heart. God bless you.

ETHICS WAIVER

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

R14/49 Miss Palesa Seemi

HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)

CLEARANCE CERTIFICATE NO. M210415

NAME: (**Principal Investigator**)	Miss Palesa Seemi
DEPARTMENT:	DST/NRF Centre of Excellence for Biomedical TB Research National Health Laboratory
PROJECT TITLE:	Understanding Mycobacterial genome diversity using a culture-omics approach
DATE CONSIDERED:	30/04/2021
DECISION:	Approved unconditionally
CONDITIONS:	
SUPERVISOR:	Dr B. Gordhan and Dr J. Peters
APPROVED BY:	Dr CB Penny, Chairperson, HREC (Medical)
DATE OF APPROVAL:	09/06/2021

This clearance certificate is valid for 5 years from date of approval. Extension may be applied for.

DECLARATION OF INVESTIGATORS

To be completed in duplicate and **ONE COPY** returned to the Research Office Secretary on the Third Floor, Faculty of Health Sciences, Phillip Tobias Building, 29 Princess of Wales Terrace, Parktown, 2193, University of the Witwatersrand. I/we fully understand the conditions under which I am/we are authorized to carry out the above-mentioned research and I/we undertake to ensure compliance with these conditions.Should any departure be contemplated, from the research protocol as approved, I/we undertake to resubmit the application to the Committee. **I agree to submit a yearly progress report.** The date for annual re-certification will be one year after the date of convened meeting where the study was initially reviewed. In this case, the study was initially reviewed in **April** and will therefore be due in the month of **April** each year. Unreported changes to the application may invalidate the clearance given by the HREC (Medical).

Principal Investigator Signature Date

PLEASE QUOTE THE PROTOCOL NUMBER IN ALL ENQUIRIES

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AFB	Acid fast bacilli
ART	Anti-retroviral therapy
ATP	Adenosine triphosphate
BAM	Binary alignment map
BCG	Bacillus Calmette-Guérin
CAS	Central Asian strain
CCD	Charged couple device
CFU	Colony forming units
CRISPR	Clustered regularly interspaced short palindromic repeats
ddNTPs	Di-deoxynucleotides
DR	Direct repeat
EAI	East African Indian
EEA1	Early endosome antigen 1
EPTB	Extrapulmonary tuberculosis
EMB	Ethambutol
ETH	Ethionamide
GC	Guanine-cytosine
HIV	Human immunodeficiency virus
ICT	Information, communication, and technology
INH	Isoniazid
IS6110	Insertion sequence 6110
iTOL	Interactive tree of life
LA	Luria agar
LAM	Latin American Mediterranean
LB	Luria-bertani broth
MDR-TB	Multi-drug resistant TB

MGIT	Mycobacteria growth indicator tube
MIRU-VNTR	Mycobacterial interspersed repetitive unit variable number tandem repeats
MOTT	Mycobacteria other than tuberculosis
MPN	Most probable number
M. tb	*Mycobacterium tuberculosis*
MTBC	*Mycobacterium tuberculosis* complex
NAATs	Nucleic acid amplification tests
NGS	Next generation sequencing
NTM	Nontuberculous mycobacteria
OADC	Oleic acid-albumin-dextrose-catalase
ORFs	Open reading frames
PANTA	Polymyxin B, amphotericin B, nalidixic acid, trimethoprim azlocillin
PBS	Phosphate buffered saline
PE	Proline glutamate
PGRS	Polymorphic guanine-cytosine repetitive sequences
PPE	Proline proline glutamate
PZA	Pyrazinamide
PTPA	Protein tyrosine phosphatase A
qPCR	Quantitative polymerase chain reaction
Rab	Ras-associated binding
RAxML	Random accelerated maximum likelihood
RIF	Rifampicin

RILP	Rab7-interacting lysosomal protein
RPF	Resuscitation promoting factor
SAM	Sequence alignment map
SAMTOOLS	Sequence alignment mapping tools
SDS	Sodium dodecyl sulphate
TB	Tuberculosis
TDS	Targeted deep sequencing
TST	Tuberculin skin test
UGT1A1	Uridine diphosphate glucuronosyltransferase 1A1
USA	United States of America
VBNC	Viable but non-culturable
WGS	Whole genome sequencing
WES	Whole exome sequencing
XDR-TB	Extensively drug resistant tuberculosis

LIST OF SYMBOLS

µl	microliters
cm	centimetres
ml	millilitres
mg	milligrams
nm	nanometres
ng	nanograms
g	grams
%	percentage
℃	degrees Celsius
V	volts
µF	microfarad
Ω	electrical resistance
Xg	centrifugal force
Rpm	revolutions per minute

ABSTRACT

Tuberculosis (TB), caused by the pathogen *Mycobacterium tuberculosis* (*M. tb*), claims millions of lives each year and is one of the leading causes of death worldwide. One of the many challenges to the eradication of this disease is the presence of heterogeneous bacterial populations within the host or infection with different strain types, especially if the metabolically quiescent population harbours drug resistance. The different growth requirements for bacteria complicate the detection of such infections in sputum because the type of growth media used selects for the fitter strain/genotype in culture, resulting in the overrepresentation of actively dividing populations compared to the minority metabolically quiescent, non-replicating populations in sputum. The purpose of this study was to determine whether culturing sputum in axenic media (with and without growth factor supplementation) which promote the enumeration of minority bacterial populations in sputum, revealed genetically distinct genotypes in sputum. This was achieved by comparing the genomic sequence obtained from uncultured sputum to the genomic sequences obtained from the same sputum sample that has been cultured under different growth conditions. This study also aimed to investigate within-host strain dynamics in mixed strain *M. tb* infections using an artificial mixed strain infection model. Results revealed that sputum culture under various growth conditions resulted in an increase in the observed genetic diversity in sputum due to the emergence of different strain types and heterogeneous bacterial populations from the same infecting strain. Furthermore, culturing was shown to be selective of a specific genotype in sputum. The mixed strain infection model required additional time for optimization. Conclusively, this study showed that whole genome sequencing is a useful tool for detecting additional strain types in sputum and that minority populations present in sputum are genetically heterogeneous and require different types of growth media to emerge in sputum.

1. INTRODUCTION

1.1 Background

Tuberculosis (TB) is the leading cause of death globally from a single infectious agent (World Health Organization, 2020). The causative agent for TB is the *Mycobacterium tuberculosis (M. tb)* bacterium, whose success as a pathogen is attributed to its ability to adapt to hostile environmental conditions within its host. In 2020, approximately 10 million infections were reported and 1.5 million lives were claimed by this pathogen (World Health Organization, 2021). The TB epidemic is most widespread in Southeast Asia, Africa, and the Western Pacific region, accounting for 87% of the global incidence in 2019 (World Health Organization, 2020). Developing countries are disproportionately affected by the disease due to limited resources, ill-equipped public health intervention programs, and limited access to reliable diagnostic tests (Castañeda-Hernández and Rodriguez-Morales, 2013). The epidemic is furthered fuelled by HIV/TB co-infection, poor treatment adherence, and the emergence of drug-resistant strains (Millet *et al., 2*013, Pai *et al., 2*016). Multidrug-resistant tuberculosis (MDR-TB) is a type of TB that is resistant to the most potent first-line anti-TB drugs isoniazid (INH) and rifampicin (RIF). Extensively drug resistant TB (XDR-TB) is a form of TB that is resistant to INH, RIF, at least one fluoroquinolone, and one of the injectable agents used in MDR-TB treatment regimens (Migliori *et al., 2*020). Moreover, the current Bacillus Calmette-Guérin (BCG) vaccine is ineffective in protecting against adult pulmonary tuberculosis, highlighting an urgent need for a global effort to strengthen TB prevention strategies.

1.1.1. Global tuberculosis burden

Tuberculosis (TB) was declared a global health emergency by the World Health Organization in 1993 (Nathavitharana and Friedland, 2015). It is estimated that a third of the world's population is latently infected, with a 5-15% risk of developing active disease (Getahun *et al., 2*015, Houben and Dodd, 2016). TB infections are more common in males and this gender bias may be attributed to the differences in social roles, risk behaviours and occupational risk factors that place men at a higher risk of exposure to the *M. tb* bacilli (Nhamoyebonde and Leslie, 2014, Miller *et al., 2*021). For example, stereotypical gender roles that impose the financial responsibility and economic sustenance of the family on males while women stay at home to care for the children increase male exposure as they are more likely to have greater social contacts. Men also work in high incidence settings (such as mines) which may facilitate transmission. TB infection and disease are becoming more widely recognized

during adolescence (10-19 years) and early adulthood (20-24 years) (Osman *et al., 2*021, Reuter *et al., 2*019, Snow *et al., 2*020). The reason for this is not completely understood. However, susceptibility in this age group may be explained by the immunological changes that occur during the transition to adulthood (Osman *et al., 2*021). This is of concern as an increase in severe illness or death from TB in individuals that fall within the economically productive age group (15-54 years) could negatively affect economic growth.

1.1.2. Economic and social determinants of tuberculosis

TB control programmes are centred around medical interventions. However, a multi-sectoral approach that includes measures to address social determinants and TB risk factors, as well as intercepting transmission are necessary to reduce the burden of disease. Factors such as overcrowding and poor ventilation in homes, workplaces, and transportation networks increase the likelihood of *M. tb* exposure (Lönnroth *et al.,* 2009). The stigma associated with TB negatively impacts health-seeking behaviour and poverty restricts access to health care facilities for diagnosis or treatment (Buregyeya *et al., 2*011, Duarte *et al., 2*018, Narasimhan *et al., 2*013). Malnutrition, which is common in high-burdened low socioeconomic settings owing to food insecurity, increases disease susceptibility and severity of disease by lowering natural immunity (Hoyt *et al., 2*019, Lazzari *et al., 2*018). Moreover, smoking, alcohol consumption and diabetes have also been implicated as risk factors for TB disease (Bates *et al., 2*007, Lönnroth *et al., 2*009, Rehm *et al., 2*009, Narasimhan *et al., 2*013) but human immunodeficiency virus (HIV) infection is the most significant risk factor for TB disease (Canetti *et al., 2*020, Duarte *et al., 2*018).

1.1.3. HIV/TB co-infection

Globally, an estimated 4.2 million TB incidence and relapse cases occurred in HIV-infected individuals in 2020 (World Health Organization, 2021). TB infection in HIV-endemic areas presents a clear and ongoing challenge at the patient level as well as health system level. The synergistic relationship of these two diseases results in higher mortality and morbidity owing to the complex immune system dynamics that manifest during coinfection (Bruchfeld *et al., 2*015, Sterling *et al., 2*010). Macrophages play a central role in the innate immune response to various pathogens. Although HIV does not result in the depletion of macrophages, various studies have shown that HIV infection impairs intrinsic macrophage-mediated defences against a variety of intracellular pathogens (Chaturvedi *et al.,*1995, Gazzinelli *et al.,*1995, Kedzierska *et al., 2*000). As a result, HIV-infected individuals are more susceptible to illnesses caused by bacteria that are typically destroyed by mononuclear phagocytes such as

M. tb. Furthermore, HIV reduces the CD4+ T cell count affecting CD4+ T cell immunity against *M. tb*. Although CD8+ T cells and other immune cells are able to defend against *M. tb* infection, their combined effect is not sufficient to compensate for the loss of the dominant CD4+ T cell response (Prezzemolo *et al., 2*014). The ability of CD4+ T cells to detect *M. tb*-infected antigen-presenting cells is crucial in containing the infection (Du Bruyn *et al., 2*021). Furthermore, drug-drug interactions and drug cytotoxicity should be carefully evaluated in the therapeutic management of HIV/TB co-infection. The development of hepatotoxicity and gastrointestinal intolerance was observed in multiple studies that aimed at determining the pharmacokinetics of TB drugs, specifically Rifampicin, and various antiretroviral treatment (ART) drugs (Cerrone *et al., 2*019, Kendall *et al., 2*021, Nabisere *et al., 2*020). Rifamycins upregulate various drug-metabolizing enzymes and transporters, which is of consequence since the induction of these enzymes reduces the plasma concentration and bioavailability of co-administered drugs (Dyavar *et al., 2*020). Rifampicin also upregulates the synthesis of the uridine diphosphate glucuronosyltransferase 1A1(UGT1A1) enzyme which metabolizes drug inhibitors (Dyavar *et al., 2*020). In addition, immune reconstitution inflammatory syndrome (IRIS) may develop resulting in the worsening of pre-existing disease following the initiation of ART hence, a specific need for clinicians to monitor TB-infected patients after initiating ART.

1.1.4. The end TB strategy

At the 67th World Health Assembly in 2014, the End TB strategy was endorsed to end the global TB epidemic by the year 2035 (World Health Organization, 2015). This strategy aims to achieve a 90% reduction in the global incidence rate, and a 95% reduction in the mortality rate while reducing the financial burden experienced by TB-affected households. The strategy includes a 2020 and a 2025 milestone based on three pillars that focus on improving service delivery for patients, the integration of government and private sector participation, and intensive research and innovation (World Health Organization, 2015). In recent years, the WHO global TB control programme has explored the application of digital health interventions to improve patient care and for public health surveillance (Falzon *et al., 2*016). The rapid development of the information, communication, and technology (ICT) sector, which presents new opportunities with each upgrade, aids in the implementation of such TB control interventions (Lee *et al.,* 2020). This necessitates the need for digital solutions to be sustainable while remaining in sync with the pace of development. Moreover, the limited evidence for the effectiveness of digital health interventions in TB control places a need for

innovative decision-makers (Guo *et al., 2*020). Interestingly, increased TB case detection and treatment initiation has been observed through public-private partnerships (Datta *et al.*, 2017, Do Thu *et al.*, 2020, Shelby *et al.*, 2017). Between 2015 and 2019, the global TB incidence and mortality has decreased by 9% and 14%, respectively. This decrease, however, was not sufficient to meet the strategy's 2020 milestone of a 20% reduction in incidence and a 35% decrease in mortality. In addition, the lack of funding for TB research threatens the realization of the 2035 objective with an investment of US$906 million in 2018, which is 55% less than the yearly target of US$2 billion (Frick and Lessem, 2020). The COVID-19 (coronavirus disease 2019) pandemic which has upended the world and forced governments to restructure and reallocate resources in order to deal with the outbreak, has reversed the progress of the End TB strategy. Molecular diagnostic platforms were reallocated to the COVID-19 response and the travel restrictions that were imposed during the lockdown period had an impact on health seeking behaviour. In countries such as India and South Africa, case detection decreased by more than 50% between March-June 2020 (World Health Organization, 2021). There was a 21% reduction (1.4 million) in the number of people receiving care for TB in 2020 compared to 2019, and for the first time in a decade, TB mortality increased with the greatest impact in high TB burdened countries (Kirby, 2021, Lakoh *et al., 2*021).

1.2. Taxonomy of *M. tb*

It has been hypothesized that *M. tb* evolved from environmental to obligate human pathogens through the adaptation to intracellular environments by organisms that feed on environmental bacteria (such as amoebas) (Jang *et al., 2*008). Genomic rearrangements such as genetic reduction and gene acquisition played a role in the evolution of the *M. tb* complex (MTBC) (Gagneux, 2018). Initially, phenotypic characteristics were used to establish evolutionary relationships between different mycobacteria however, the discriminatory power of phenotypic characteristics was debatable due to overlapping phenotypic characteristics among genotypically distinct organisms (Palomino *et al., 2*007). Advances in molecular techniques have improved the understanding of mycobacterial taxonomic classification.

Actinobacteria is one of the largest phyla with 6 classes, 5 subclasses, 6 orders, and 14 suborders (Ludwig *et al., 2*012). Taxonomically, in the order of Actinomycetales, mycobacteria belong to the family *Mycobacteriaceae* and the genus *Mycobacterium* (Figure 1.1). Currently, over 170 species of *Mycobacterium* have been identified (Forbes, 2017). The

species are further classified into two major groups namely the *M. tb* complex (MTBC) and the non-tuberculous mycobacteria (NTM) (Sykes, 2013).

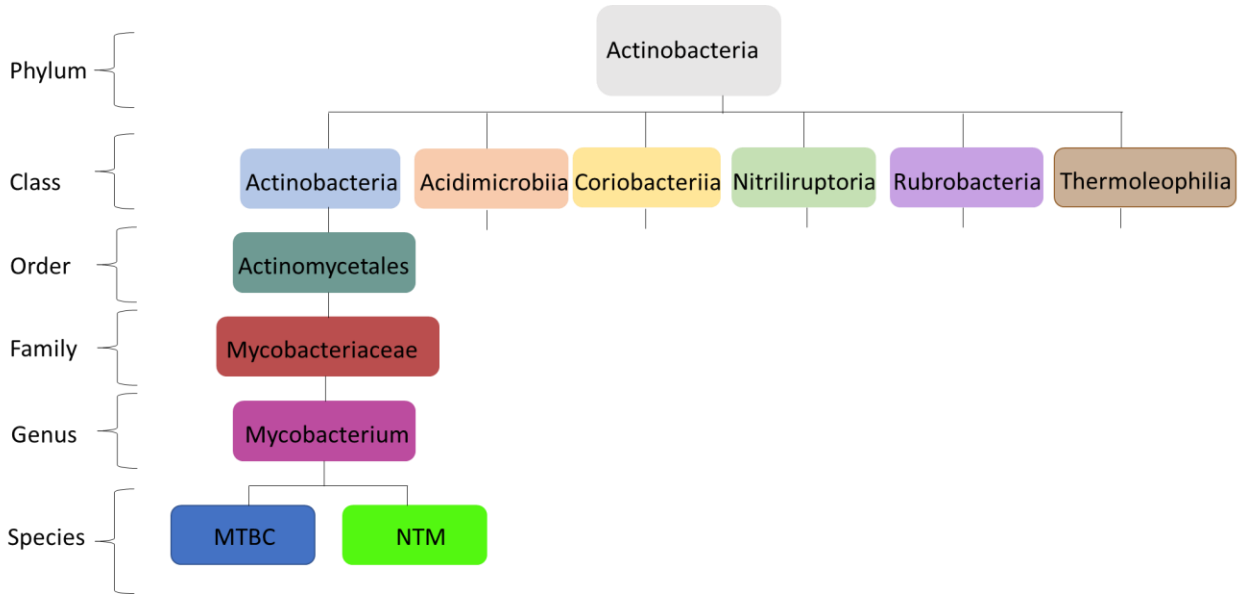


Figure 1.1. Taxonomic classification for the *M. tb* complex and non-tuberculosis mycobacteria species. Adapted from Mohammadipanah and Dehhaghi, 2017.

1.2.1. MTBC complex

The MTBC consists of eight genetically related members of a complex with a 99% nucleotide identity and identical 16s rRNA gene sequence (Beye *et al., 2*018, Böddinghaus *et al.*,1990, Sreevatsan *et al.*,1997). The members of the MTBC are as follows; *M. africanum*, *M. bovis*, *M. bovis BCG*, *M. canetti*, *M. microti*, *M. caprae*, *M. pinnipedii*, and *M. tuberculosis.* These species differ in terms of their host range, morphology, geographical distribution, and pathogenicity (Comas *et al., 2*013, Palomino *et al., 2*007). Members of the MTBC are obligate and opportunistic pathogens responsible for TB disease in both humans and animals. *Mycobacterium africanum* and *M. tuberculosis* are the primary cause of TB disease in humans, while *M. bovis* is the cause of TB disease in wild or domesticated animals (Teppawar *et al., 2*018). Bovine tuberculosis (defined as TB transmitted from animals to humans) is commonly transmitted through the consumption of unpasteurized milk and untreated animal products. In 2019, there were approximately 140, 000 new cases of zoonotic tuberculosis, with Africa accounting for 49.21% of these infections. (World Health Organization, 2020). The high incidence in Africa could be attributed to the reliance on agriculture and subsistence farming for survival. The vaccination of cattle has been employed

as a preventative measure in endemic regions (Hope and Villarreal-Ramos, 2008, Teppawar *et al., 2*018). The only approved TB vaccine for human protection is the Bacille Calmette-Guérin (BCG) vaccine, which is based on the live attenuated *M. bovis* BCG strain. Non-tuberculous mycobacteria cause clinical diseases such as lymphadenitis, skin disease, and disseminated disease in AIDS, but do not cause TB disease (Harris *et al.*, 2009, Horsburgh Jr, 1996).

1.3 Tuberculosis pathogenesis

1.3.1. Transmission and host immune response

Tuberculosis is an airborne disease transmitted through the inhalation of infected aerosol droplets carrying viable tubercle bacilli. Infected droplets are released into the air via coughing or sneezing from an individual with active disease and can remain airborne for several hours (Caws *et al., 2*015). The bacilli travel down the nasal passage to the alveoli where they can face one of 3 fates. They can be (i) eradicated without an immune trace of infection (individual will have a negative tuberculin skin test (TST) result), (ii) engulfed and contained in granulomas (TST+, disease symptom negative), or (iii) cause active disease depending on the competency of the host's immune system (Dheda *et al., 2*016).

1.3.2. Establishment of latent TB infection

The immune cascade following *M. tb* infection begins with the recognition of protein-associated molecular patterns (PAMPs) on the surface of the pathogen by pattern recognition receptors (PRRs) and toll-like receptors (TLRs) on the macrophage surface. This initiates the innate immune response resulting in macrophage phagocytosis (Liu *et al., 2*017). The macrophage then encapsulates the bacteria in its phagosome and fuses with the lysosome to dissolve and eliminate the pathogen. *M. tb* can inhibit macrophage maturation and phagosome-lysosome fusion thus promoting its survival within the host. Studies have attributed this to the release of protein tyrosine phosphatase A (PTPA) by *M. tb* which binds to the H-subunit of the vacuolar-H^+-ATPase, preventing the recruitment of effector EEA1 (early endosome antigen 1) and RILP (Rab7-interacting lysosomal protein) which are specific for Rab 5 and Rab 7 (Ras-associated binding) protein (Langemeyer *et al., 2*018, López de Armentia *et al., 2*016, Mascarello *et al., 2*013, Wong *et al., 2*011). A cell-mediated inflammatory response then recruits proinflammatory cytokines, chemokines, and other immune cells to surround the site of infection resulting in the formation of a granuloma

which consists of *M. tb* infected macrophages in the centre. Infected macrophages are surrounded by epithelioid cells, multinucleated cells, and an outer lymphocyte layer (Guirado and Schlesinger, 2013). The granuloma localizes the infection and prevents dissemination. Failure of the immune system to contain the bacilli results in the release of *M. tb* from the granuloma which spreads within the lungs and causes active pulmonary TB.

1.3.3. Clinical presentations of active disease

Active TB disease can be a result of a primary infection (Primary TB) or a recurring disease (reactivation of latent disease) (post-primary TB). The two are distinguished by their position in the lung and histology. The granuloma, which manifests as a macroscopic lesion in the lower or middle zones of the lung (referred to as the Gohn complex), including hilar and/or paratracheal lymphadenopathy and enlarged lymph nodes, are characteristic of primary disease (Heemskerk *et al., 2*015, Palomino *et al., 2*007). The self-destructive nature of the lesion leaves a sign of calcification and fibrosis which is detectable on chest X-rays as the Renke complex. Post-primary TB manifests as a lipid pneumonia in the periphery of the upper lobes of the lungs (Hunter *et al., 2*014, Welsh *et al., 2*011). Lung cavitation associated with post-primary disease promotes the endobronchial spread of the bacilli resulting in continuous sputum production and elimination (Palomino *et al., 2*007). Symptoms that are associated with the initial development of disease include low-grade fevers, night sweats, coughs lasting for 3 weeks, fatigue, and unexplained weight loss (Getahun *et al., 2*015).

1.4. Tuberculosis treatment and prevention

1.4.1. Chemotherapy

Four first-line drugs, namely isoniazid (INH), ethambutol (EMB), pyrazinamide (PZA) and rifampicin (RIF), have been developed to treat drug-susceptible TB. The duration of the treatment is for a period of six months (Palomino *et al., 2*007). Combined therapy has been proven to be more effective in TB treatment due to the quick selection of drug-resistant bacilli that result from monotherapy leading to treatment failure (Li *et al., 2*021, Petruševska *et al., 2*020). The four drugs are administered during the first two months of treatment known as the initial phase and only two drugs (RIF and INH) are administered for the next four months during the sterilization phase (Nahid *et al., 2*016). Multi-drug resistant TB (MDR-TB) and extensively drug-resistant TB (XDR-TB) are treated over a longer period of up to 24 months. Injectable drugs such as streptomycin, kanamycin, amikacin, and capreomycin are

required for treating MDR-TB together with fluoroquinolones such as levofloxacin, ofloxacin, moxifloxacin, and gatifloxacin (Rich and Jaramillo, 2008). XDR-TB can be treated with bedaquiline, pretomanid, and linezolid which have been associated with improved treatment outcomes (Padayatchi *et al., 2*020, Oelofse *et al., 2*021, Tweed *et al., 2*019). Community-based initiatives to monitor and improve treatment adherence are required as long treatment duration, drug side effects, incorrect doses and poor communication between health care providers and patients have been identified as risk factors for treatment failure (Viney *et al., 2*014).

1.4.2. BCG vaccine

The BCG vaccine is usually administered to infants and young children in high-endemic regions for protection against severe and disseminated tuberculosis disease (Aguirre-Blanco *et al., 2*007). BCG is ineffective in protecting against adult pulmonary TB (Fatima *et al., 2*020, Fine, 1995). The protection offered by the vaccine is limited with studies reporting protection for about 10-20 years (Abubakar *et al., 2*013, Barreto *et al., 2*005, Nguipdop-Djomo *et al., 2*016, Rodrigues *et al., 2*007). The genetic diversity of the different BCG vaccination strains and the varying degrees of T-cell responses elicited by these strains may account for the varying levels of protection (Aguirre-Blanco *et al., 2*007, Ritz *et al., 2*012). There is evidence of a strain-effect relationship that provides a selective advantage to specific strain types. BCG vaccination has been shown to be less protective against typical ('modern') Beijing strains (Moreno-Mendieta *et al., 2*019, Nieuwenhuizen and Kaufmann, 2018, Yang *et al., 2*020). The safety and efficacy of new recombinant BCG vaccine candidates such as MVA85A, M72/AS01E and VPM1002 have been thoroughly tested (Brazier and McShane, 2020, Tait *et al., 2*019, Tameris *et al., 2*013). The results of the M72/AS01E vaccine are encouraging however further validation in a larger study population is required (Tait *et al.*, 2019).

1.5. Diagnosis

1.5.1. Smear microscopy

Smear microscopy is usually the first test performed on sputum samples to detect the presence of acid-fast bacteria (AFB) in low resource settings using either conventional (Ziehl-Neelson or Kinyoun) or fluorescent microscopy (auramine O) (Ryu, 2015). As a diagnostic tool, smear microscopy has a low sensitivity requiring 5000-10 000 AFB per mL of the specimen for a positive result (Palomino *et al., 2*007). This is a disadvantage for specimens with low bacterial loads as they are at risk of false-negative diagnosis (Acharya *et*

*al., 2*020). Moreover, AFB-smear microscopy does not distinguish between live and dead bacteria thus requiring an additional culture step for confirmation.

1.5.2. Culture-based methods

Culture-based methods are more sensitive, requiring 10-1000 viable bacteria per mL of the specimen for a positive result (Palomino *et al., 2*007). Culture sensitivity is demonstrated in documented smear-negative sputum samples that were positive on culture (Lombardi *et al., 2*017, Rasool *et al., 2*019, Sinshaw *et al., 2*019). Bacterial growth is promoted on egg-based (Löwenstein-Jensen) and agar-based solid media (Middlebrook 7H10 and 7H11), as well as in liquid broth (Middlebrook 7H9). The lengthy turnaround time (3-8 weeks) as a result of the slow-growing nature of *M. tuberculosis* is a major limiting factor for culture-based methods (Ryu, 2015). More recently, automated liquid culture systems that fluorometrically detect mycobacterial growth by measuring carbon dioxide production, oxygen consumption, and changes in gas pressure within the culture have become widely available (Cruciani *et al., 2*004, Ryu, 2015). Examples of such systems include the BACTEC 9000 (Becton Dickinson Diagnostic Systems, Sparks, MD), MB/Bact Alert 10 3D (Biomérieux, Durham, NC, USA), and BACTEC MGIT 960 assay (Becton Dickinson Diagnostic Systems, Sparks, MD). Automated systems are limited by their high-cost and inability to differentiate between different mycobacterial species (Cruciani *et al.,* 2004).

1.5.3. Molecular diagnostic tools

Accurate identification of *M. tb* in suspected cases is important for effective case management. Advances in molecular diagnostics have allowed the accurate identification of *M. tb* DNA in specimens using line-probe assays and nucleic acid amplification tests (NAATs). The GeneXpert/RIF assay (Cepheid, Sunnyvale, USA) is the most widely used NAAT endorsed by the WHO as the initial test to diagnose TB simultaneously with RIF resistance in under 2 hours (World Health Organization, 2013). Using this system, mycobacterial DNA is amplified by a polymerase chain reaction in a disposable cartridge and RIF resistance is detected by mutations in the *rpoB* gene which codes for a β sub-unit of an RNA polymerase that is targeted by rifampicin (Mechal *et al.,* 2019). Cartridge usage minimizes biohazard risks and human error. A systematic review and meta-analysis of 106 studies reported a 98% specificity and 95% sensitivity in high burden areas. However, limited sensitivity (46%) has been reported in low burdened regions (Sohn *et al.,* 2014, Li *et al.,* 2017). To overcome the limitation of the GeneXpert MTB/RIF assay in its limited ability to

detect drug resistance, the Xpert MTB/XDR assay (Cepheid, Sunnyvale, USA) has been developed with expanded drug susceptibility testing and a rapid turnaround time (< 90 minutes). The assay is able to detect resistance associated with drugs such as INH, ethionamide (ETH), second-line injectable drugs, and fluoroquinolones in a single run (Penn-Nicholson *et al.,* 2022).

1.6. The *M. tb* genome

1.6.1. Key features

A defining moment in TB research was the completion of the genomic sequencing of the *M. tb* H37Rv strain in 1998 (Cole *et al.,*1998). The genome of the organism is 4,411,529 bp long and codes for approximately 4,000 genes of which 3,924 code for proteins, 50 code for stable RNAs, and the function of the remaining genes is unknown. Four years later, 82 novel protein-coding sequences were identified, with 22 of them having a putative function (Camus *et al.,* 2002). With an overall GC content of 65,59%, several regions with a higher GC content (>80%) distinguished by repeat sequences were identified and related to two novel gene families named the PE (proline-glutamate) and PPE (proline-proline-glutamate) families. These two large multigene families account for 10% of the coding capacity of the *M. tb* H37Rv genome and due to their high GC content and repetitive nature, research on this family of genes is limited.

1.6.2. Genomic polymorphisms

Polymorphic sites (defined as regions in the DNA sequence of an individual that vary from the DNA sequence of the parental strain) in the bacterial genome may be classified into three groups: single nucleotide polymorphisms (SNPs), large sequence polymorphisms (LSPs) and polymorphisms in repetitive sequences (Srinidhi Desikan, 2015). The majority of nucleotide polymorphisms are located in the coding region of the *M. tuberculosis* genome, which has a 90-96% coding capacity (Namouchi *et al., 2*012). SNPs are random nucleotide substitutions resulting in an amino acid change known as non-synonymous SNPs (nSNPs) or those that do not result in an amino acid change or are functionally neutral, known as synonymous SNPs (sSNPS). As a result of their effect on gene functions, nSNPs are under selective pressure. Synonymous SNPs are regarded as selectively neutral, however, there is increasing evidence of their importance in gene regulation (Rolandelli *et al., 2*019, Folkvardsen *et al., 2*018). Large sequence polymorphisms, defined as the insertion or deletion of nucleotide fragments

that are less than 1kb in length into or from genomic DNA, have a varying impact on gene function. LSPs were found to be unique event polymorphisms that occur once in a species and do not display homoplasy (sequence similarity between different species or taxa that is of independent evolutionary origin) (Hirsh *et al., 2*004). This finding suggests that LSPs do not play a significant role in disease pathogenesis as mutations that confer evolutionary advantages are usually under selective pressure during evolution. This hypothesis was supported by Kato-Maeda *et al.* (2001) who found genomic deletions in 93 open reading frames (ORFs) of 19 *M. tb* clinical strains which attenuated their pathogenicity. Contradicting results suggested that LSP-associated indels provide a selective advantage in strains with deletions in the *plcD* gene where patients infected with strains containing this deletion are twice as likely to develop extrapulmonary TB (Yang *et al., 2*005). The identification of positively selected regions in genomes may give insight into adaptive processes and information on essential genes (Yang *et al.,* 2005).

1.6.3. Consequences of genomic polymorphisms

The functional consequence of genomic variations in *M. tb* strains is often the acquisition of traits that may or may not be advantageous for their survival and pathogenicity within the host. Understanding the genomic differences between *M. tb* strains is useful for the determination of the basis of strain-specific characteristics. Differences in SNPs and LSPs are used to classify MTBC lineages, which include the Beijing, Central Asian strain (CAS), East African, Haarlem, Indian (EAI), Latin American Mediterranean (LAM), S-family, T-family, and X-family. Various genotyping methods exploit these genomic polymorphisms as genetic markers for strain differentiation. Polymorphisms that occur in fourteen distinct regions of the genome (known as regions of difference or RD1-14) account for the observed differences in pathobiological features such as virulence, clinical manifestations, immunological responses, and drug resistance (Forrellad *et al., 2*013, Ru *et al., 2*017).

Depending on the deletion of the RD1 region, the strains are further classified as ‘ancient’ or ‘modern’ lineage strains. Different immunological responses were observed between both classes of strains. In addition to an increased risk of transmission, modern lineage strains were associated with a weaker immunological response (De Jong *et al.,* 2008, Hakamata *et al.,* 2020, Portevin *et al.,* 2011). This is of consequence due to the high prevalence of modern clade strains which contributes to the burden of TB disease (Portevin *et al.,* 2011, Wiens *et al.,* 2018). The absence of the RD1 region, which codes for a virulence-related type VII

secretion system, also highlights the clinical utility associated with modern BCG vaccinations. RD1 is absent in all BCG vaccination strains but present in virulent *M. tb* and *M. bovis strains*. (Ru *et al., 2*017, Kozak *et al., 2*011).

Associations between human genetic variations and *M. tuberculosis* genetic diversity provide insights into host-pathogen co-evolution (Gagneux, 2012, McHenry *et al.,* 2020, Uren *et al.,* 2021). *M. tuberculosis* has a low mutation rate of 0.3-0.5 mutations per genome per year and despite this low mutation rate, it has been able to develop resistance against all drugs used against it, indicating the robust adaptation capability of this human-adapted pathogen (Eldholm and Balloux, 2016). Genomic regions associated with drug resistance include *katG* and *inhA* for isoniazid, *pncA* for pyrazinamide, and *embB* for ethambutol. However, mutations in additional drug target genes occurring as a result of antibiotic selection pressure, results in the emergence of strains that are resistant to multiple drugs.

1.6.4. Utilisation of polymorphisms for strain characterisation

Genetic polymorphisms are caused by replication slippage, homologous recombination, and IS6110 (insertion sequence 6110) transposition (Fang *et al.,*1998, Filliol *et al., 2*000, Groenen *et al.,*1993, Hancock, 1995). The IS*6110* transposon is a mobile genetic element belonging to the IS3 family, of which the copy number and location within the genome are used as genetic markers for *M. tb* genotyping (Cave *et al.,*1991, Jagielski *et al., 2*014, Srinidhi Desikan, 2015). To identify different strains using the IS6110 element, genomic DNA is cleaved by the PvuII restriction enzyme, generating fragments that are separated by capillary electrophoresis and hybridised to a membrane for visualisation (Jagielski *et al., 2*014). The pattern of different PvuII fragment sizes is unique per strain lineage. A limitation of this technique is the need for a time-consuming culture step as 2-3 µg of high-quality DNA is required for restriction digestion. Moreover, the discriminatory power of this technique has been demonstrated to be low for strains presenting with a low IS6110 copy number (<6) (Cave *et al.,* 1991, Mathema *et al., 2*006). In such instances, secondary markers such as clustered regularly interspaced short palindromic repeats (CRISPR) and polymorphic guanine-cytosine (GC)-rich repetitive sequences (PGRs) are required to provide sufficient information for strain differentiation (Srinidhi Desikan, 2015).

Mycobacterial interspersed repetitive units (MIRU) are 40-100 base pair (bp) DNA sequences found in 24 independent loci throughout the *M. tb* genome. Previously, the system was based on 12 loci however, other sets of MIRU-VNTR loci were proven to increase

isolate discrimination over the 12-locus system (Alonso-Rodríguez *et al.,* 2008, Iwamoto *et al.,* 2007). This approach utilises variations in the number of repeat sequences occurring within the different loci. The sequences are amplified by polymerase chain reaction (PCR) and the amplicon sizes are determined by capillary electrophoresis. (Nikolayevskyy *et al.,* 2016). The limited resolution of this method has been highlighted in different outbreak investigations where local transmission was ruled out with MIRU-VNTR typing but upon the application of whole genome sequencing (WGS), local transmission was confirmed (Bainomugisa *et al., 2*018, Gardy *et al., 2*011, Pérez-Lago *et al., 2*014). The direct repeat (DR) locus in MTBC strains consists of 10-50 copies of 36 bp DNA sequences interspersed with non-repetitive conserved spacer sequences. Since the spacer sequences are highly conserved and vary in their absence or presence between different strains, the DR locus is used as a target for *in vitro* amplification. The DR locus is amplified by PCR using inversely orientated primers and the PCR products are reverse hybridized to a nylon membrane consisting of 43 covalently bound spacer oligonucleotides derived from the *M. tb* H37Rv, and the *M. bovis* BCG strain (Kamerbeek *et al.,*1997). Resulting hybridization patterns are characteristic of different evolutionary lineages. This method is utilised during Spoligotyping. Thousands of isolates have been analysed using this technique which has provided insight into the global *M. tb* strain diversity and epidemiological insights into the transmission of TB (Brudey *et al., 2*006). Spoligotyping has a lower discriminatory power compared to both IS6110-RFLP and MIRU-VNTR, which is often improved by complementation with either technique (Mathuria *et al., 2*008, Sann *et al., 2*020).

Although these methods target genetic polymorphisms, they investigate less than 1% of the genome (Liu *et al., 2*012). This limitation can be overcome by the application of next generation whole genome sequencing (WGS) which provides the full genetic profile including all potential genomic targets, valuable information on drug resistance, and genome evolution which is otherwise missed by targeting the previously discussed marker sequences (Roetzer *et al., 2*013).

1.7. DNA sequencing

Information on hereditary and biochemical properties within an organism is stored in the order of the nucleic acids in polynucleotide chains. The three-dimensional structure of DNA was solved in 1953 by efforts from Franklin, Watson, and Crick (Watson and Crick, 1953) which led to the development of several technologies aimed at deciphering the genetic code.

1.7.1. First generation sequencing

In 1965, Holley obtained the first ever nucleotide sequence of the alanine tRNA gene from *Saccharomyces cerevisiae* through the enzymatic digestion of the polynucleotide *(*Holley *et al.*, 1965*)*. That same year, Fred Sanger developed an alternative sequencing method referred to as the chain termination or dideoxy method (Sanger *et al.*,1965). In this method, the elongation of the nucleotide chain is terminated through the incorporation of a dideoxynucleotide (ddNTPs) lacking a 3'-hydroxyl group required for extension by DNA polymerase. This produces DNA fragments of different lengths, which are separated by capillary electrophoresis and the sequence is determined on a gel. The nucleotide sequence of the template is indicated by the sequence of the bands in the lanes. The major limitation with this initial approach to sequencing is that the DNA being sequenced cannot exceed 1000 bp, which becomes quite challenging when sequencing large genomes. Sanger sequencing was generally accepted as the 'first-generation' sequencing (FGS) technology.

1.7.2. Second generation sequencing

The second generation of sequencing technologies (also known as next generation sequencing technologies) were introduced in 2005. Second generation sequencing technologies are characterised by their ability to sequence multiple fragments simultaneously, resulting in high throughput data in a short period of time and at a lower cost (Liu *et al.*, 2012).

1.7.2.1. Illumina sequencing

The Illumina sequencing platform is the most successful NGS system claiming a market share of more than 70%, notably with the HiSeq and MiSeq platforms (Kulski, 2016). This technology adopted the 'sequencing by synthesis' method for complementary strand synthesis. The sequencing procedure is divided into three stages: sample preparation, cluster generation, and sequencing (Figure 1.2). A DNA library is produced during sample preparation by fragmenting genomic DNA within a specific size range using sonication, nebulisation, or exonuclease activity (Quail *et al., 2*009). Two types of adapters (oligonucleotides) are then ligated to the ends of the DNA fragments. These adapters consist of a complementary region to the flow cell oligonucleotide for ligation, a sequencing primer binding site, and an index tag for pooling multiple libraries which will be sequenced in a single run (multiplexing). During the second stage, each DNA fragment is isothermally

amplified by bridge PCR, generating clusters of clonal DNA. Approximately one million copies of the original sequence are found in each cluster (Kchouk *et al.*, 2017). Four nucleotides (dATP, dGTP, dCTP, and dTTP), each with a distinct cleavable fluorescent dye and blocking group, are used in the sequencing step (Liu *et al.*, 2012). These allow the reaction to continue one base at a time. The fluorescent dye is photographed with each nucleotide incorporation and then removed, allowing a new cycle to begin. While providing high throughput at lower costs, Illumina sequencing is a highly accurate technique with an overall error rate of less than 1% (Liu *et al.*, 2012, Stoler and Nekrutenko, 2021).

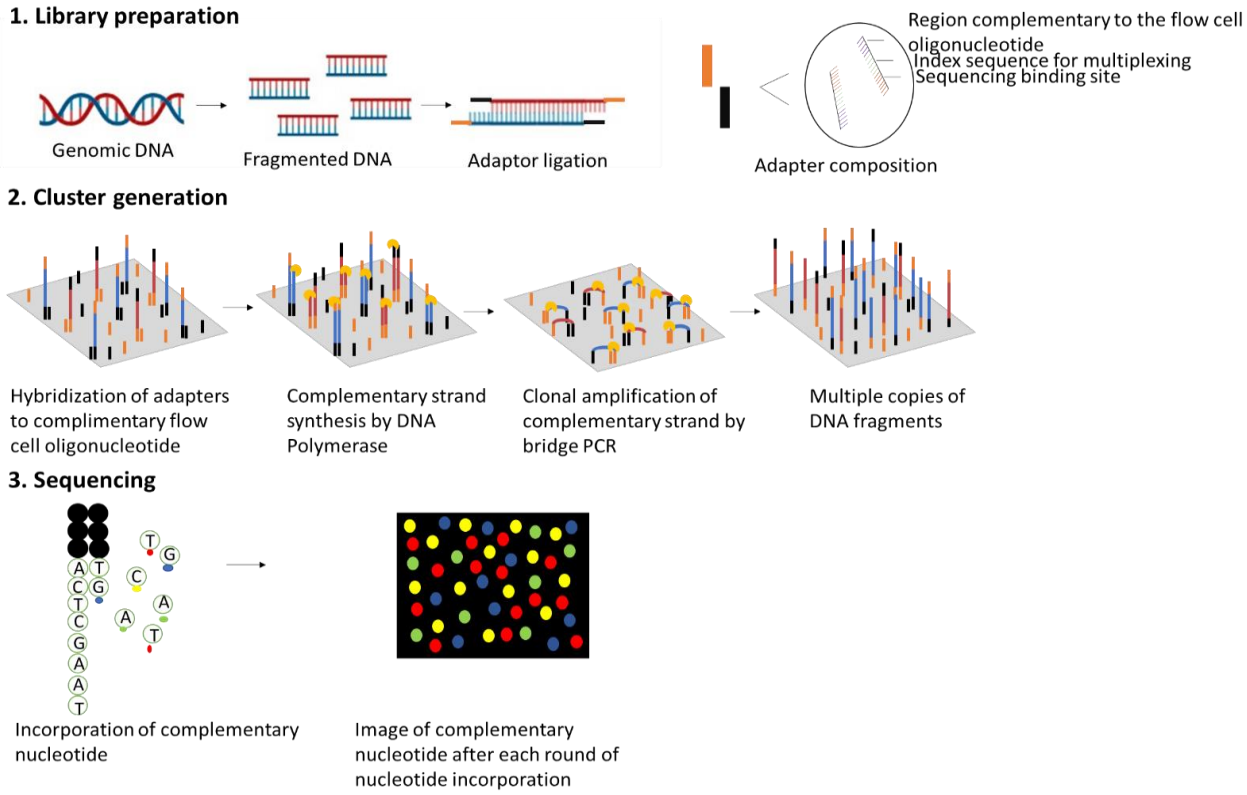


Figure 1.2: Workflow of the Illumina sequencing system. **1)**Adapters containing sequences complementary to the flow cell oligonucleotides are ligated to the fragmented DNA during sample preparation. **2)** DNA libraries are loaded onto the flow cell for amplification using bridge PCR. **3)** A fluorescent signal corresponding to each nucleotide is emitted after incorporation into the DNA strand.

1.7.2.2. Ion torrent sequencing

The Ion Torrent technology translates chemical information into sequence data (Figure 1.3). This technology utilises a microchip with a series of microwells containing a bead. Short DNA fragments are generated using mechanical methods prior to the ligation of adaptors. The ligated fragments bind to oligonucleotides on the surface of the beads for sequencing using emulsion PCR. A proton is released in solution with the incorporation of a nucleotide

and the pH change is detected by a semiconductor transistor (Liu *et al.*, 2012). The sequencing accuracy is reduced in homopolymeric (>6bp) regions as the pH changes, which are then represented as undefined peaks and are difficult to interpret (Kchouk *et al.*, 2017).

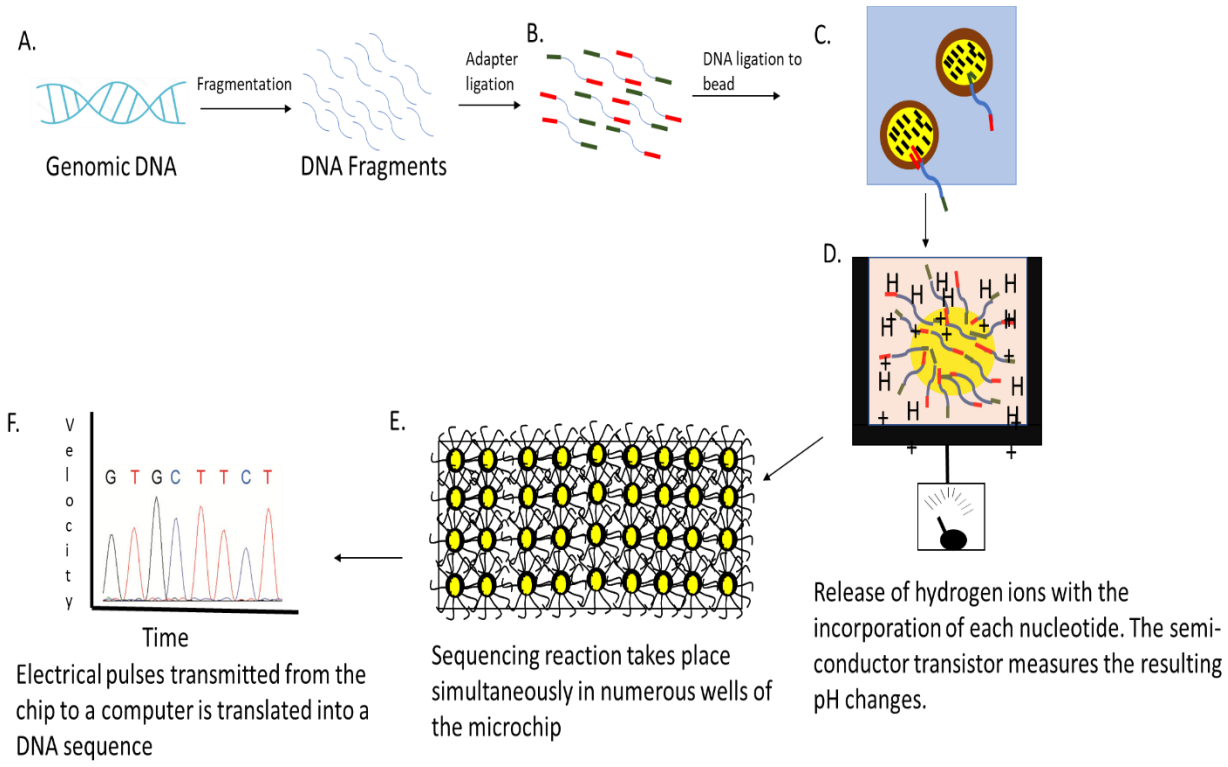


Figure 1.3: Workflow of Ion-torrent sequencing. A-C) Single-stranded DNA fragments created during library preparation are ligated on both ends to adapter sequences. The ligated DNA fragments are captured on microscopic beads bound with complementary oligonucleotides. **D)** Hydrogen ions are released as a by-product with the addition of each nucleotide to the DNA fragment during complimentary strand synthesis. **E)** Clonal amplification of various fragments occurs simultaneously in numerous microwells. **F)** pH changes during each cycle are used to determine the nucleotide sequence.

1.7.3. Applications of Next generation sequencing

Next generation sequencing technologies have shown potential in genomic research and diagnostic settings (Barzon *et al.*, 2011). NGS has been implemented for numerous applications including in transcriptome sequencing at the DNA or RNA level, non-coding RNA expression profiling, targeted gene sequencing, whole exome sequencing (WES), and whole genome sequencing (Hempelmann *et al.*, 2018, Hsiao *et al.*, 2017, Nambot *et al.*, 2018, Tagliani *et al.*, 2021). The sequencing of the entire genomic sequence of organisms (WGS) has assisted in the understanding of bacterial physiological and pathogenic features, outbreak investigations and surveillance, microbial risk assessment of foodborne pathogens, and drug susceptibility predictions (Franz *et al.*, 2016, Genestet *et al.*, 2019b, Ren *et al.*,

*2*003, Tagliani *et al., 2*021, Walker *et al., 2*015). WGS improves clinical management of drug resistance TB (DR-TB) cases by providing information on the patient's full drug-resistance profile. This is useful for guiding individualised clinical decisions for complex resistance cases. In a case study by Dookie *et al.* (2020), additional mutations in the *inhA* coding region, which were not detected by the line probe assay, were revealed by the WGS results of the patient. Patient management was guided by experts based on the individual's profile of resistance (Dookie *et al.*, 2020). Moreover, understanding the genetic changes that result in altered drug responses through the identification of variants in drug target genes has improved the prediction of drug-susceptible phenotypes (Farhat *et al.*, 2016, Papakonstantinou *et al.*, 2021). New epistatic relationships amongst resistance-associated genes and novel mutations to drugs such as cycloserine, para-aminosalicylic acid, and ethionamide were found in a genome-wide association study using 6,465 *M. tb* clinical isolates (Coll *et al.*, 2018). This study revealed new drug targets that will aid in the development of treatment against DR *M. tb* strains.

1.8. Heterogenous *M. tuberculosis* bacterial populations in single sputum samples

M. tb infections were commonly thought to be clonally homogenous and limited to a single strain. However, studies have demonstrated heterogeneity within the infecting bacterial population (Black *et al.*, 2015, Chengalroyen *et al.*, 2016, Dusthackeer *et al.*, 2019, Ford *et al.*, 2012, Mukamolova *et al.*, 2010, Nimmo *et al.*, 2019). Heterogenous bacterial populations within the host can occur as a result of (I) mixed infection with multiple strains, (II) super infection or re-infection with a different strain, or (III) in-host evolution of a single strain by the acquisition of mutations (Figure 1.4).

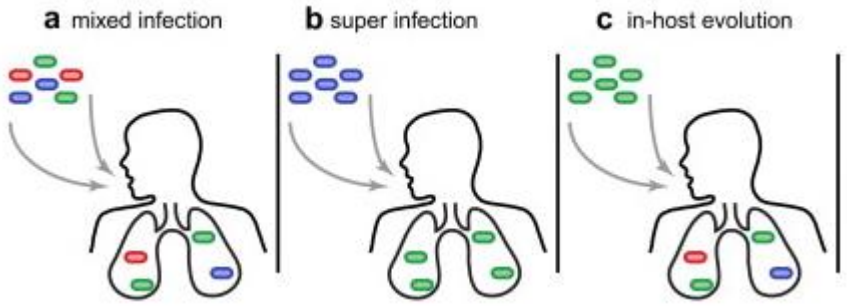


Figure 1.4: Sources of population heterogeneity in *M. tb* infections. A) Depicts simultaneous infection with multiple *M. tb* strains. The different colours represent various *M. tb* strains. **B)** Depicts super infection or re-infection with a distinct strain. **C)** Depicts heterogeneity that arises from mutational events within a single strain infection. Adopted from (Ford *et al., 2*012).

1.8.1. Within host genetic diversity due to mixed strain infection

The belief that *M. tb* infections were caused by a single strain was challenged when different *M. tb* strains were identified from single colonies isolated from a single patient using phage typing (Bates *et al.*,1976, Mankiewicz and Liivak, 1975). Mixed tuberculosis infections are defined as TB disease caused by the presence of two or more clonally distinct strains. Genotyping methods have revealed the clonal complexity of mixed strain infections, with distinct strains identified in different regions of the lungs or at different independent anatomical sites (García de Viedma *et al., 2*003, Kaplan *et al., 2*003).

1.8.1.1. Challenges in the detection of mixed M. tuberculosis infections

Low bacillary loads during latent infections challenge early case identification of mixed strain infections. The lung bacillary load per lesion has been estimated to be around 10^8, which means that the minor strain should be present in large numbers at the time of detection in order to be detected as part of the mixed strain infection (Gillespie, 2002). Shamputa *et al.* (2006) noted that increasing the number of specimens from a single patient increases the likelihood of detecting additional strains, if any, in sputum. This was attributed to the inaccessibility of all pulmonary lesions to the airways at the same time, which impacts the reflection of the true heterogeneity within the patient. In addition, culture has been shown to underestimate the clonal composition and population structure of sputum (Martín *et al.*, 2010, Nimmo *et al.*, 2019). This is of consequence as it can impact treatment and clinical outcomes, especially if the undetected strains possess important traits such as antibiotic resistance and is more virulent. Direct whole genome sputum sequencing could circumvent this issue; however, it is complicated by the presence of low *M. tb* bacterial loads in sputum, the presence of host DNA and other microorganisms (Soundararajan *et al*., 2020).

1.8.2. Within host genetic diversity due to in-host evolution of a single strain

During active replication, a bacterial cell replicates its chromosome once per cell cycle resulting in a population of genetically identical cells. Bacteria are exposed to a variety of microenvironmental conditions that are unfavourable for growth such as acidic pH, nutrient deprivation, low oxygen supply and oxidative stresses. To withstand these stresses, bacteria adopt several adaptive responses such as entering into a reversible state of metabolic quiescence or dormancy.

1.8.2.1. Dormancy

Bacterial dormancy is a non-replicative state characterized by low metabolic activity, in which replicative growth is resumed under favourable environmental conditions (Gengenbacher and Kaufmann, 2012). Dormant bacteria are able to persist in the host for years and have been implicated in latent tuberculosis infection (LTBI) cases (Dusthackeer *et al., 2*019, Gutti *et al., 2*019, Leung *et al., 2*013). It is hypothesized that the progression from LTBI to active disease is caused by the reactivation of these dormant bacteria. Dormant states are achieved by inactive spore formation in bacteria such as *B. subtilis* and *Clostridium* spp (van Vliet, 2015). However, spore formation has not been described in TB. Instead, in TB, dormancy is described by states of reduced growth and metabolism known as persisters, viable but non-culturable (VNBC) and the differentially culturable/detectable tubercle bacilli (DCTB/DDTB).

1.8.2.2. Persisters

Persisters are a sub-population of bacteria that are genetically identical to drug-susceptible bacteria but display phenotypic resistance during exposure to lethal antibiotic doses (Mandal *et al., 2*019). Persisters challenge TB treatment as the reduced metabolic state of the bacteria results in insufficient production of drug targets, rendering antibiotics ineffective (Mandal *et al., 2*019). In addition, the majority of bacterial cells are eradicated during antibiotic exposure. However, a sub-population of the bacteria survives and the population is restored once the antibiotic is removed. Although the exact mechanism of persister formation is unknown, the SOS response to DNA damage has been found to play a significant role in persister generation (Dörr *et al., 2*009, Podlesek and Žgur Bertok, 2020).

1.8.2.3. Viable non-culturable bacteria (VBNC)

VBNC's are distinguished from persisters based on their ability to regain culturability. Unlike persisters that regain culturability once a stressor has been removed (antibiotics), VBNC require time and/or additional supplementation to regain culturability (Ayrapetyan *et al., 2*015). These bacteria are not able to form colonies on solid media but can be resuscitated in liquid media. Culturability of these organisms is enhanced by supplementation with growth factors such as culture filtrate (CF) from growing bacteria and resuscitation-promoting factors (RPFs) (Chengalroyen *et al., 2*016, Mukamolova *et al., 2*010, Nikitushkin *et al., 2*016, Shleeva *et al., 2*002)

1.8.2.4. Differentially culturable tubercule bacilli (DCTB)

Smear-positive *M. tb* sputum is dominated by VBNC cells that have been termed differentially culturable tubercule bacteria (DCTB) (Chengalroyen *et al., 2*016, Mukamolova *et al., 2*010). Different DCTB populations have been identified in sputum, including CF-independent DCTB, Rpf –independent DCTB, and CF-dependent DCTB (Chengalroyen *et al., 2*016). Rpfs are muralytic enzymes secreted into the culture filtrate (CF) of axenically grown wild type *M. tb* and were first identified in *Micrococcus luteus* (Mukamolova *et al.,* 1998). In this family of enzymes, a functional hierarchy has been observed with *rpfB* and *rpfE* ranking above *rpf D* (Kana *et al.,* 2008). RipA (Rpf-interacting protein A) is a binding partner of RpfB, whose enzymatic activity at the septum suggests that the complex may be involved in cell division and potentially during reactivation from dormancy (Hett *et al., 2*007). These findings question current diagnostics that are based on growth assays since non-replicating populations are undetected by standard diagnostic methods, meaning that individuals with these forms of dormant organisms could possibly be misdiagnosed as TB negative. This is of importance in individuals harbouring metabolically quiescent populations with drug resistance. In such cases, the prescribed treatment will fail to eliminate all bacteria present. Furthermore, re-activation of the non-replicating resistant strain could occur at a later stage due to inadequate treatment and the use of treatment regimens that do not account for the dormant resistant strain. It is thus important that all populations present in a sputum sample are represented during diagnosis for effective treatment and clearance of all infecting organisms.

1.9. Knowledge gap

Culturomics encompasses the optimization of culture conditions to allow the growth of all organisms within a sputum sample, thus ensuring maximum representation of all bacterial populations in the sample (Bilen *et al., 2*018). Previous studies have demonstrated that culturing *M. tb* sputum reduces within-sample genetic diversity owing to the selection of fitter strains or random loss of lineages (Nimmo *et al., 2*019, Shockey *et al., 2*019). The effect of sputum culture in growth conditions that promote the growth of minor populations present in sputum on the within-sample genetic diversity has not been investigated. This MSc aims to address this by determining whether axenic culturing of sputum using different growth assays in the absence or presence of growth factors will increase the observed genetic diversity of *M. tb* from sputum samples. Furthermore, since the exact mechanism of strain competition

during mixed strain *M. tb* infections is not completely understood, an *in-vitro* mixed strain infection model will be used to gain insight into the dynamics of the disease at the individual level.

1.10. Aims and objectives.

1.10.1. Aims

This study aimed to determine the extent of the genetic diversity in sputum-derived *M. tb* cultured in different axenic media, with and without growth factor supplementation and to develop an *in-vitro* model of a mixed-strain *M. tb* infection to investigate within-host strain behaviour.

1.10.2. Objectives

Clinical samples were previously grown in different growth conditions for the enumeration of *M. tb* bacterial populations in sputum. The following was conducted to achieve the study aims:

- Re-grow sputum-derived *M. tb* bacteria from stored patient samples previously grown in different growth assay platforms of liquid or solid media (CFU, MPN, and MGIT), un-supplemented or supplemented with different growth factors (CF, RPF-B, RPF-B/RIP-A).
- Extraction of genomic DNA from these bacterial cultures for whole-genome sequencing.
- Identify genotypic differences in the DNA sequence data of the organisms grown in the absence and presence of different growth factors.
- Assess the impact of the different growth factors on the genetic diversity by a comparative analysis of the sequence data from the samples grown under the various culture conditions to the sequence data of the original sputum sample which did not undergo axenic growth.
- Assess the rate at which Beijing and non-Beijing strains outgrow each other in a controlled axenic mixed-strain infection model using real-time quantitative PCR.

2. METHODS AND MATERIALS

2.1. Media and culture conditions.

All *M. tb* strains including the laboratory strain H37Rv, the clinical strains and the barcoded strains used in this study (Table 2.1) were grown in Middlebrook 7H9 broth supplemented with 0.2% glycerol, 0.05% Tween-80 and 10% BD BBL™ Middlebrook Oleic acid-albumin-dextrose-catalase (OADC) at 37 ℃ for 7-10 days or on Middlebrook 7H11 agar supplemented with 0.2% glycerol, 10% OADC enrichment at 37 ℃ for 7-10 days. The solid media was supplemented with 50 µg/ml of Kanamycin (Kan) when necessary. The *E. coli* DH5α strain was grown in Luria-Bertani broth (LB) or on Luria-Bertani agar (LA) at 37℃ with supplementation of both media with 50 µg/ml of kanamycin. The liquid cultures were grown with shaking at 100 rpm. For decontamination of the clinical strains, Middlebrook 7H9 liquid media was supplemented with 2% polymyxin B, amphotericin B, nalidixic acid, trimethoprim and azlocillin (PANTA). Both Middlebrook 7H11 and 7H9 media was incubated at 37℃ for 3 days to ensure sterility before use. The composition of all the above-mentioned media is described in detail in Appendix A, Table A1.

2.2. List of strains used in this study.

Table 2.1- Laboratory and clinical strains used/generated in this study.

Strain	Description	Source
M. tb H37Rv	First isolated in 1905, the strain has been maintained by serial passage of cultures.	(Ioerger *et al.*, 2010)
M. tb clinical samples	The Mycobacterial biomarkers of TB response and success in adult patients: A prospective cohort study (BMG-phase I) was conducted using different growth assays to assess the response patterns of DCTB during standard tuberculosis chemotherapy.	(Peters *et al.*, Manuscript in preparation)
Eschericia coli DH5α	*SupE44 ΔlacU169 hsdR17 recA1 endA1 gyrA96 thi-1 relA1*	Promega, Madison, WI
Barcoded LAM and Beijing clinical strains	Clinical *M. tb* strains carrying plasmids C2-M_pUC57 and C4-S_pUC57.	This study

The *M. tb* clinical strains used in this study were frozen stored samples from the Mycobacterial biomarkers of TB response and success in adult patients: A prospective cohort study (BMG-phase I) (ethics clearance number: M120256) which was conducted at the

Centre of Excellence for Biomedical Tuberculosis Research (CBTBR) (Peters *et al.,* Manuscript in preparation, 2022). In the parent study, sputum from TB infected patients was tested for the presence of differentially culturable tubercule bacteria (DCTB) using the most probable number (MPN) serial dilution assay (Chengalroyen *et al.,* 2016) and the colony forming units (CFU) assay (Mukamolova *et al.,* 2010), and thereafter, using the same patient samples, the resuscitation capabilities of specific growth factors on DCTB were assessed using the Mycobacteria growth indicator tube (MGIT) assay (Ardito *et al.,* 2001). The patients' sputum was tested under a total of nine culture conditions (Figure 2.1). Bacterial cells were harvested from each assay and stored in 1 mL aliquots at -80 °C for up to 2 years. Figure 2.1 summarizes the various culture conditions, and the following text provides a brief explanation of how the different growth assays function, as well as information on the various growth factors that were supplemented in the growth assays. The MPN assay provides a quantitative estimation of viable bacteria in a sample by a serial dilution of the sample in liquid growth media and a Poisson calculation (Figure 2.1 A). The CFU assay estimates the number of viable bacteria in a sample based on the number of colonies that form on an agar plate (Figure 2.1 B). In the MGIT assay, viable bacteria in a sample are measured using a fluorometric system that measures the rate of oxygen consumption in the sample-containing tube. (Figure 2.1 C). Culture filtrate, one of the various growth factors, is a collection of growth stimulatory molecules released into the media by bacteria during growth (Sonnenberg and Belisle, 1997). RPFB (resuscitation promoting factor B) is an enzyme that has been implicated in the resuscitation of *M. tb* from dormancy and RIPA/RPFB is a recombinant enzyme of both RPF-interacting protein A and an RPFB (Nikitushkin *et al.,* 2016). A total of ten patient samples that have been grown using the three growth assays with the different growth factor supplementation were selected at random for further analysis in this study. In addition, each patient's raw sputum sample that had not been cultured in any manner was also included in this analysis.

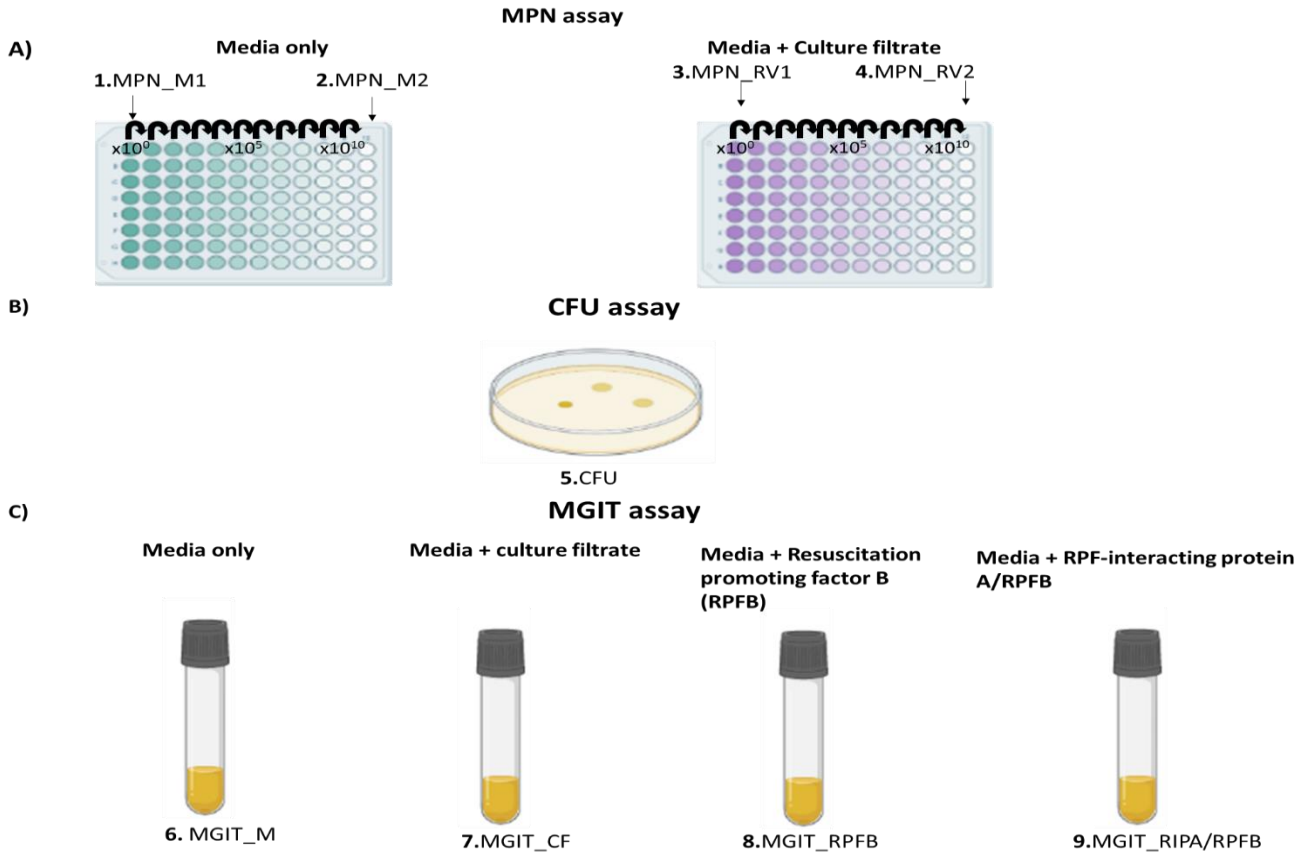


Figure 2.1: Nine culture conditions that were used in the BMG phase I study. **A)** Decontaminated sputum was cultured in the liquid media based MPN (Most probable number assay) in the absence (M1 and M2) and in the presence of culture filtrate supplementation (RV1 and RV2). M1 and RV1 represents the sample in the first well of the plate. M2 and RV2 represent the sample in the last well of the plate. **B)** Decontaminated sputum was streaked directly onto solid media for colony formation. **C)** Sputum was inoculated into MGIT (Mycobacterial growth indicator tubes) tubes containing liquid growth media (M), a mixture of antibiotics and an oxygen quenching fluorescent compound. Three tubes were supplemented with different growth factors, namely culture filtrate (CF), Resuscitation promoting factor B (RPFB) and a recombinant RPF interacting protein A (RIPA)/RPFB.

2.3. Clinical sample purification

2.3.1. Chemical decontamination of the stored sputum samples

Decontamination of the stored sputum samples was carried out using the N-acetyl-L-cysteine-Sodium Hydroxide (NALC-NaOH) decontamination method (Bradner *et al.*, 2013) to digest and kill contaminating bacteria commonly found in sputum samples. The sputum samples stored in 1 mL aliquots were retrieved from the -80°C freezer, thawed and carefully transferred to a 50 mL Falcon tube. To liquefy the viscous sputum and kill contaminating agents, an equal volume (1:1) of N-acetyl-l-cysteine–sodium hydroxide (NALC-NaOH) was added to the specimen and mixed by vortexing 3 times consecutively at maximum speed, for 10 seconds each, with 5-minute intervals in between. The tube was left to stand for 2 minutes to allow the aerosols to settle before the addition of phosphate-buffered saline (PBS) (pH 7.0-

7.4) to the 10 mL mark on the tube to neutralize the pH. The sample was mixed by inversion and then centrifuged at 3000 rpm at 4°C for 15 minutes to pellet the bacterial cells. The supernatant was carefully discarded without dislodging the pellet and 1.5 mL of 2 mm sterile glass beads was added to the pellet. Thereafter, 2 mL of PBS was added and the pellet resuspended to generate a single cell suspension by vortexing 3 times consecutively for 10 seconds at a time. The tube was left to stand for 5 minutes to allow the aerosols to settle and the decontaminated specimen was transferred into a 2 mL Eppendorf tube and stored at -80°C.

2.3.2. Decontamination of the cultured sputum samples

The samples stored from the MPN, CFU and MGIT assay was also taken through a purification step to confirm the presence of *M. tb* and the absence of other contaminating bacteria in the samples. BBL™ polymyxin B, amphotericin B, nalidixic acid, trimethoprim and azlocillin (PANTA) antibiotic mixture was added to the growth media to prevent the overgrowth of contaminants (made as described in Appendix A) (Figure 2.2 A). The 1 mL stored aliquots of the bacterial cells were retrieved from the -80°C freezer, thawed and grown in 9 ml of the PANTA supplemented 7H9 liquid media (Figure 2.2 B) at 37°C for 7-10 days. The bacterial cultures were monitored at 3-day intervals for growth, and once there was sign of bacterial growth, 100 µL of the culture was streaked onto 7H11 media (Figure 2.2 C) and incubated at 37°C for another 7-10 days. This was done to confirm the presence of *M. tuberculosis* in the sample and simultaneously detect the presence of contaminating organisms if any. The remaining liquid culture was stored in 1 mL aliquots at -80 °C as interim stocks for repurification in the event of contamination. To ensure the purity of the samples, a single colony was picked from the 7H11 plates and grown in 9 mL of 7H9 media without PANTA (Figure 2.2 D-E) at 37°C for 7-10 days. When there was observable growth in the liquid culture, 100 µL of the culture was streaked out onto 7H11 agar (Figure 2.1 F) and incubated at 37°C for 7-10 days. One mL aliquots of the remaining culture were stored at -80°C as interim stocks (Figure 2.2 G). Once the bacterial growth on the 7H11 agar plates was contamination-free a single colony from the plate was sub-cultured in 9 mL of Middlebrook 7H9 to OD_{600nm}=0.5-0.8 (Figure 2.2 H-I) to make working stocks which were stored in 1 mL aliquots (Figure 2.2 J) at -80°C. These stocks were used for all downstream experiments.

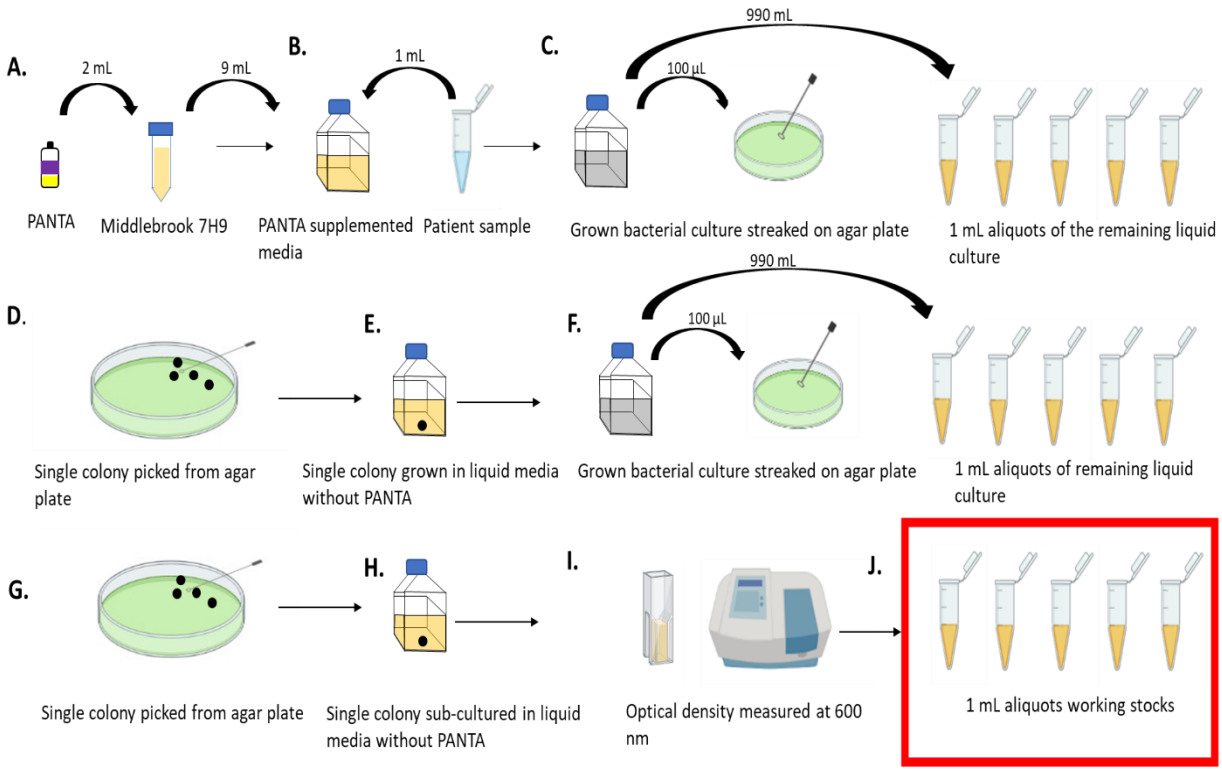


Figure 2.2: Schematic illustration of the purification process of the *M. tb* samples. Bacterial cells were grown in PANTA supplemented media to eliminate contaminants (A-C). Bacterial cells were re-grown without the antibiotic mixture to ensure the absence of contaminants the sample (D-F). Pure isolates were stored in 1 ml aliqotes at -80°C (G-J).

2.4. Large-scale genomic DNA isolation using the Cetyl trimethylammonium bromide (CTAB/NaCl) method

The CTAB/NaCl method (Hosek *et al.*, 2006) was used to isolate genomic DNA (gDNA) from the decontaminated sputum and the purified cultured sputum samples. The *M. tb* H37Rv strain was used for the optimisation of bacterial DNA extractions. Once optimised, gDNA was isolated from the clinical samples for downstream analysis. For the growth of *M. tb* H37Rv and the *M. tb* clinical samples, 1 mL of the respective freezer stocks were thawed and grown in 9 mL of Middlebrook 7H9 liquid media at 37 °C to an OD_{600nm}=1-2. The cultures were transferred to a 50 mL Falcon tube and the bacterial cells were pelleted at 4317 x g for 10 minutes. The pellets were resuspended in 500 µL of 1x TE buffer to prevent nucleic acid degradation and transferred into a 2 mL O-ring screw-cap tube. Following resuspension in 1x TE buffer, the cells were heat-killed at 75°C for 30 minutes. The tubes were left to cool for 2 minutes before the addition of 50 µL of lysozyme (10 mg/ml) to the suspension. The tubes were incubated at 37°C for 1 hour to break down the bacterial cell wall. Following lysozyme treatment, 70 µL of 10% SDS and 6 µL of proteinase K (10 mg/ml) was added to each tube, mixed by inversion, and incubated at 65°C for 1 hour to allow for protein degradation.

Following this, 100 µL of pre-warmed 5 M NaCl (65°C) and 80 µL of prewarmed CTAB/NaCl (65°C) were added to each tube and incubated at 65°C for 1 hour to facilitate the separation of polysaccharides that may interfere with the isolation of the gDNA. An equal volume (900 µL) of chloroform: isoamyl alcohol (24:1 v/v) was added and mixed by inversion before centrifugation at 12470 x g for 10 minutes to separate cell debris into the organic phase and the gDNA into the aqueous phase. Without disturbing the protein interphase, the top aqueous layer was transferred into a sterile 1.5 mL O-ring screw cap tube with 450 µL of isopropanol to precipitate the DNA. DNA was precipitated by centrifugation at 12470 x g for 20 minutes. The gDNA pellet was washed with 500 µL of cold 70% ethanol followed by centrifugation at 12470 x g for 5 minutes to remove any remaining salts from the precipitation step. The supernatant was discarded and the pellet was dried at 60 °C for 10 minutes in a SpeedVac (Thermo Fischer Scientific™) and resuspended in 100 µL of 1x TE buffer. The DNA was allowed to dissolve at 65°C for 10 minutes. The concentration, purity, and integrity of the extracted gDNA was determined for genotyping and whole-genome sequencing.

2.5. Quantification and visualisation of genomic DNA

2.5.1. Quantification of genomic DNA

The Nanodrop ND-1000 spectrophotometer (Thermo Fischer Scientific™) was used to determine the concentration and purity of the extracted gDNA. The absorbance of the sample at 260 nm was used to quantify the DNA, and the purity was assessed by the DNA to protein ratio, which is determined by the absorbance at 260 nm and 280 nm, respectively. Samples with an A_{260}/A_{280} ratio between 1.8-2.0 were considered as pure.

2.5.2. Analysis of gDNA by agarose gel electrophoresis

A 1% agarose gel was used to determine the integrity of the extracted gDNA. The gel was prepared by dissolving 0.3 g of SEAKEM® Agarose gel powder in 30 mL of fresh 1x TAE (Tris, Acetic acid, EDTA). The agarose mixture was then boiled in a microwave, in 30-second intervals (maximum of 90 seconds), until it was completely dissolved. The bottom of the flask was placed under cold running water for about 40-60 seconds, or until the temperature of the flask was about 50 °C. A volume of 3 µL of SYBR™ safe DNA gel stain was added to the agarose solution and mixed by swirling. The solution was then poured into a gel casting tray and allowed to solidify at room temperature for 15-20 minutes. The solidified gel was transferred to an electrophoresis chamber which was filled with 1x TAE buffer until

it covered the gel. A volume of 10 µL of the DNA sample was mixed with 5 µL of loading dye on a piece of parafilm and loaded into the wells of the gel. The DNA was electrophoresed at 90V for 45 minutes. DNA visualisation was done under UV light using the SynGene Gbox (Bio-Rad) and the Snapgene version 7.12 acquisition software.

2.6. Genotyping of clinical samples

2.6.1. Spoligotyping

The spoligotyping technique was performed to identify the lineages of the clinical samples using the gDNA that was extracted in section 2.4 above. Spoligotyping is a PCR-based genotyping technique that detects the presence or absence of 43 spacers in the direct repeat (DR) genomic region of the strains belonging to the *M. tb* complex (Driscoll, 2009). Biotinylated PCR products are hybridised to complementary membrane-bound oligonucleotides and are detected by a chemical reaction with streptavidin and electrochemiluminescence (ECL) reagents.

2.6.1.1 Polymerase Chain reaction (PCR)

Polymerase chain reaction was performed to amplify the DR regions in the extracted gDNA from each clinical sample. The forward and reverse primers (Table 2.2), as well as the BCG and *M. tb* H37RV controls, were supplied with the spoligotyping kit from Mapmygenome (Hyderabad, India). The primers were reconstituted by the addition of 1 mL of nuclease-free water (Hyclone) and stored in 200 µL aliquots for future use. The PCR reactions were conducted in a final volume of 25 µL by the addition of 6.5 µL of Hyclone nuclease-free water, 12.5 µL of ReadyMix™ Taq PCR reaction mix, 2 µL of DRa primer, 2 µL of DRb primer and 2 µL of the extracted gDNA. The PCR tubes were spun briefly using a mini centrifuge (Bio-Rad laboratories) to ensure all the reagents were mixed and were then placed in a thermal cycler for the amplification process (Bio-Rad laboratories). The cycling conditions used for the PCR reaction were as follows: initial denaturation (95°C for 3 minutes), second denaturation (94°C for 1 minute), annealing of primers (55°C for 1 minute), and extension (72°C for 30 seconds). This cycle was repeated 30 times and the final extension was at 72°C for 10 minutes. Once complete, the biotinylated PCR products were stored at 4°C, for no longer than 7 days, before hybridisation was carried out (Driscoll, 2009).

Table 2.2: PCR primers used for amplification of the direct repeat sequences in the clinical strains

Primer	Primer sequence (5' to 3')	Annealing temperature	Storage temperature	Reference
DRa*(5'biotin)	GGTTTTGGGTCTGACGAC	55°C	4°C	(Kamerbeek *et al.*, 1997)
DRb	CCGAGAGGGGACGGAAAC	55°C	-20 °C	

2.6.1.2. Hybridisation and detection

To dilute the PCR products, 150 µL of 2X SSPE/0.1% SDS was added to the stored PCR tubes and denatured at 99°C for 10 minutes using the thermal cycler (Figure 2.3 A-B). The tubes were immediately placed on ice to prevent the DNA from re-annealing. The spoligotpying membrane from the spoligotyping kit supplied by Mapmygemone (Hyderabad, India) was prepared with pre-warmed 2X SSPE/0.1% SDS at 60°C for 10 minutes. Thereafter, the membrane was placed on a foam cushion in the mini blotter ensuring that the loading wells ran parallel to the lines on the membrane and that the side with the snipped edge was positioned on the bottom left of the blotter. Excess fluid was wiped off before tightly sealing the mini blotter with plastic screws. Residual fluid was removed by aspiration from the slots of the mini blotter before loading 170 µL of the denatured PCR products onto the membrane (Figure 2.3 C-D). The H37Rv and BCG control samples and a no-DNA template control (NTC) were loaded into the first three wells, followed by the PCR products from the clinical samples in the remaining wells. After loading the PCR products, the mini blotter was transferred to the hybridisation oven at 60°C for 1 hour. Thereafter, the samples were aspirated from the slots and the membrane was removed from the mini blotter. The membrane was then washed twice at 60°C for 5min with 2X SSPE/0.5% SDS to remove unbound probes (Figure 2.3 E).

After the wash step, the membrane was placed in a roller bottle with a mixture of 40 mL of 2x SSPE/0.5% SDS and 10 µL of Streptavidin-peroxidase conjugate (500U/mL) and incubated at 42°C for 1 hour in the hybrisation oven. Through a chemical reaction, the streptavidin-peroxidase binds to the biotinylated membrane-bound PCR products for detection. The membrane was washed twice with 2x SSPE/0.5% SDS by incubation at 42°C for 5 minutes and then once more at room temperature for another 5 minutes. Approximately,

10 mL of both detection solution I and II from the ECL detection kit (GE Healthcare) were added and swirled for 90 seconds to enhance the chemiluminescent detection of the bound products. The solutions were discarded immediately and the membrane was placed in a clear plastic sleeve which was exposed to three light-sensitive X-ray film in a dark room for 40 minutes.

The X-rays were individually developed in the darkroom in the following order: developing solution, water, fixer solution, and then water. Each wash step was performed for 30 seconds. The X-rays were hung to dry before analysis while the membrane was stripped by incubation in 10% SDS at 80 °C for 1 hour and stored in 200mM EDTA at 4°C for reuse. Once dry, the spoligotype patterns on the sheets were analysed by the presence or absence of a small black square on the developed X-ray film spotted at each DR loci (Figure 2.3 F). Different patterns/combinations indicate the genotype of the strains. The spoligotyping pattern of each sample was determined using the SPOTCLUST spoligotyping website (https://tbinsight.cs.rpi.edu/runspotclust.html) (Figure 2.3 G). The binary value of 1 was assigned to the presence of the spacer, while the binary value of 0 was assigned to indicate the presence of a spacer.

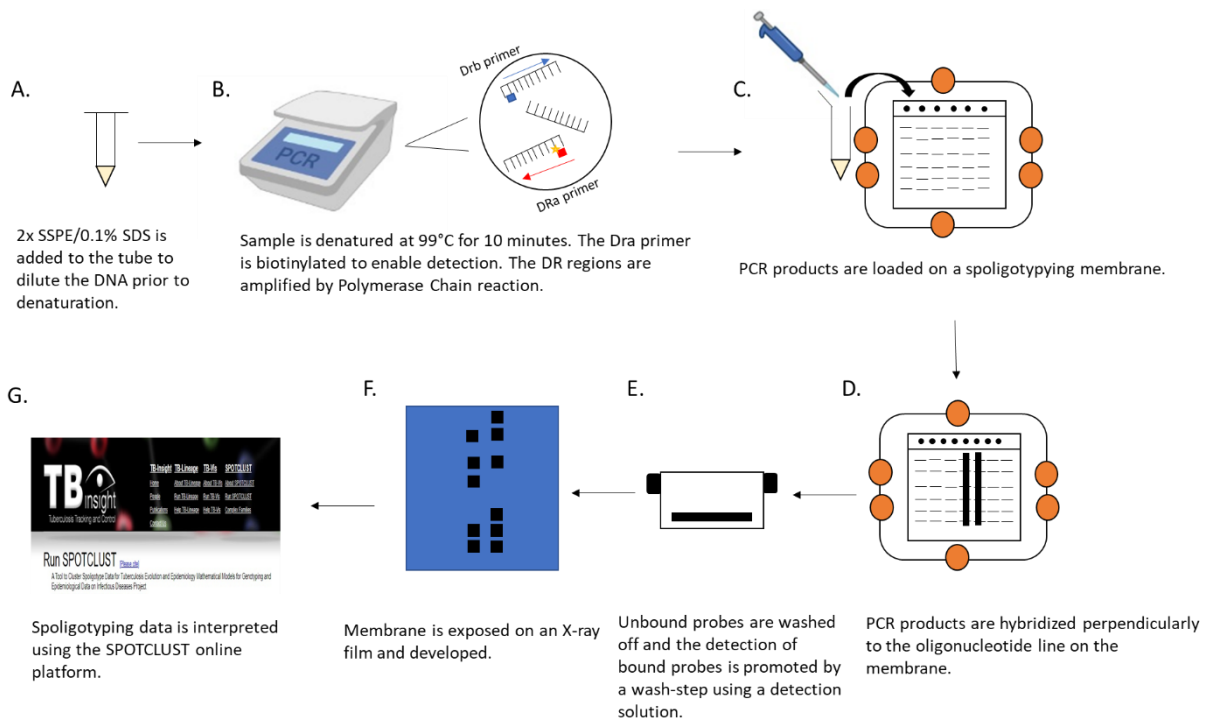


Figure 2.3: Schematic illustration of the spoligotyping technique. A) PCR products were diluted and **B)** denatured using a thermal cycler. **C-D)** The denatured PCR products were hybridised on membrane bound oligonucleotides and **E-F)** the membrane was developed for visualisation of the hybridisation pattern. **G)** The resulting pattern was interpreted using SPOTCLUST (https://tbinsight.cs.rpi.edu/run_spotclust.html) to determine the strain type (Vitol *et al.*, 2006).

2.7. Next generation sequencing

To obtain the genome sequence of the mycobacteria analysed in this study, the genomic DNA (gDNA) extracted from the various samples was sent for whole genome sequencing at Rutgers University in Columbia, Newark. The sequencing was performed on the Illumina HiSeq platform where the DNA fragments were sequenced in the 3'-5' and 5'-3' direction, producing paired end reads of 150 base pairs each (bp). The FastQ whole genome sequence data files were shared via a secure file transfer protocol (FTP) (OneDrive) of the University.

2.7.1. Bioinformatics analysis of whole genome sequencing data

Whole genome sequencing data were analysed using a custom-made data analysis pipeline. The data analysis pipeline utilises various open-source next generation sequencing (NGS) analysis tools combined with Python commands. The workflow for the analysis of the whole genome sequences is summarised in Figure 2.4.

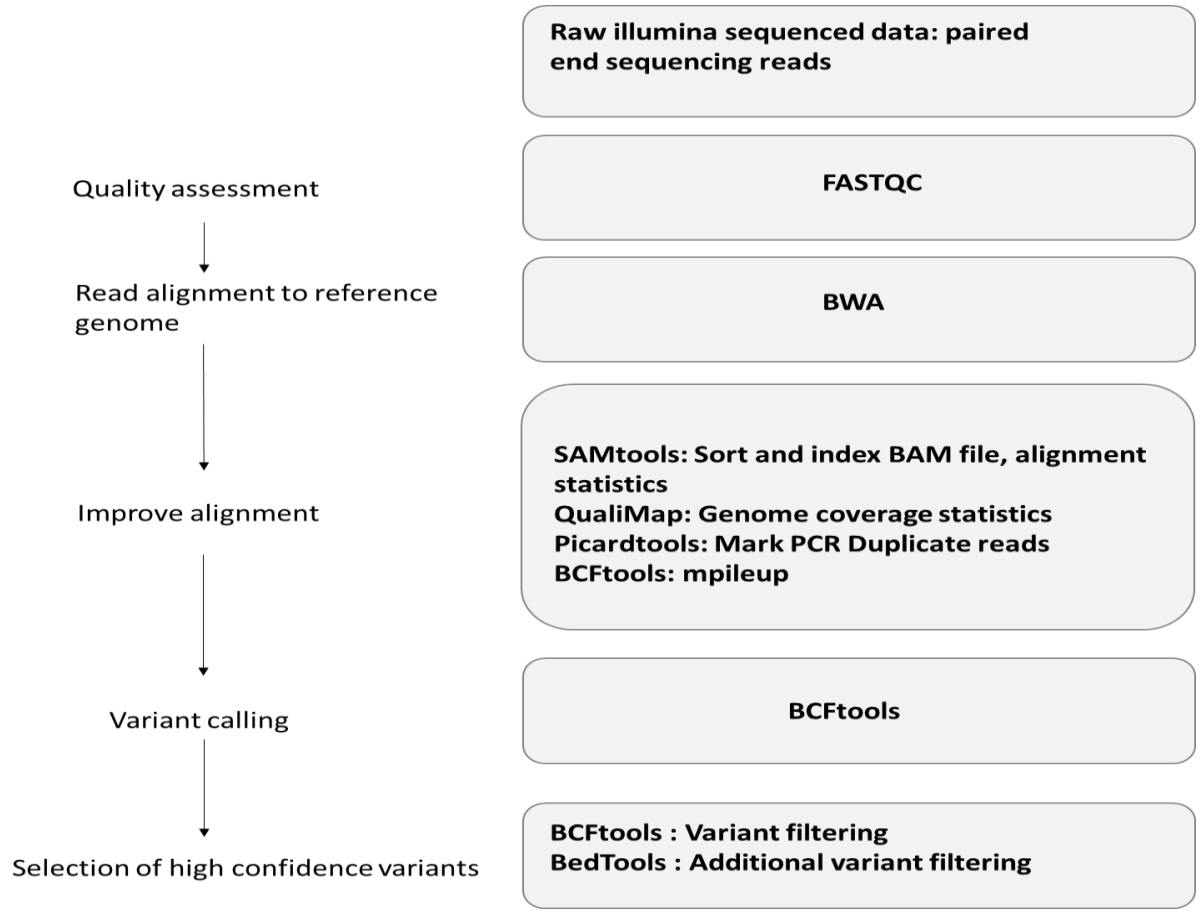


Figure 2.4: Workflow of the computational analysis of the raw sequencing data and the bioinformatics tools that were used at each step. Specialised bioinformatic tools were utilised for each aspect of the analysis. The tools are listed in the grey chart. **Abbreviations**: BWA, Burrows-Wheeler Aligner; SAMtools, Sequence Alignment/Map tools; PCR, polymerase chain reaction.

2.7.1.1. Read mapping and variant detection

The raw reads were used to infer *in-silico* strain spoligotypes using the SpolPred program (a C++ program used to predict strain spoligotypes from raw sequencing reads) (Coll *et al., 2*012). Thereafter, (as depicted in Figure 2.4) the quality of the paired-end reads was assessed using the FastQC v0.11.9 toolkit (Andrews, 2010). All bases had a Phred quality score > 30, thus read trimming was not required. Reads were mapped against the *M. tb* H37Rv reference genome (Genbank accession number: NC_000962.3) using BWA MEM v0.7.17-r1188 (Li and Durbin, 2009). SAMtools v1.10 (Li *et al., 2*009) was used for SAM to BAM file conversion, retrieval of basic alignment statistics, and for sorting and indexing the chromosomal co-ordinated alignment file for downstream analysis. Samples with a coverage breadth < 95% were excluded from the analysis. Genome coverage statistics were obtained using QualiMap v2.2.2 (Okonechnikov *et al., 2*016). Duplicate reads were marked using Picard Tools v2.26.2 (Picard, 2018) and variants, including single nucleotide polymorphisms (SNPs) and small insertion/deletions (indels), were called using BCFtools v.1.10.2 (Danecek *et al.,* 2014). A high confidence variant file was generated after retaining variants with a base quality ≥ 30, and a read depth > 10. Of these variants, those found in repetitive families of genes (PE, PPE, and PE-PGRS gene families), possible transposons, insertion sequences and phage genes were removed using BEDtools v2.27.1 (Quinlan and Hall, 2010).

2.7.2. Detection of mixed infections using the number of heterozygous SNPs from the WGS data

To detect mixed infections in sputum, the heterozygous sites method by Sobkowiak *et al.* (2018) was employed. Heterozygous base calls are indicative of mixed infections since variants identified at specific genomic sites from different strains results in more than one allele call at the specific site. The number of heterozygous SNPs (which includes small insertions and deletions) were identified from each sputum sample using BCFftools v1.10.2 (Danecek *et al., 2*014) with heterozygous base calls enabled. Variant filtering was carried out as above (section 2.7.1). An allele frequency of 1 was used for the heterozygous variants. For classification of mixed infections, sputum samples with ≥ 20 heterozygous sites were considered as potential mixed infections (Sobkowiak *et al., 2*018). An additional condition of a > 1.5 % heterozygous to total SNP proportion was required for classification of samples with between 11-19 heterozygous SNPs as a mixed infection.

2.7.3. Phylogenetic Analysis

Extracted variant sites (allele frequency = 1) were concatenated and used to construct a maximum-likelihood phylogenetic tree to assess the genetic relatedness of the clinical samples based on their genetic distance. The variants were converted to FASTA format using a Perl script https://github.com/decarlin/stuartlab-scripts/blob/master/perl/Tools/tab2fasta.pl and all sequences were aligned using MAFFT v7.490 (Katoh *et al., 2*002) prior to tree building with the random accelerated maximum likelihood (RAxML) tool (Stamatakis, 2014), based on the GTRCAT substitution model using 1000 bootstrap replicates. The tree was visualised using the online tool iTOL (https://itol.embl.de/) and pairwise SNP distances were calculated using snp-dists v0.8.2 (https://github.com/tseemann/snp-dists).

2.7.4. Metagenomic analysis

Metagenomic analysis was performed on the raw sputum samples to assess the proportion of contaminating reads using Kraken2 (Wood and Salzberg, 2014). Kraken assigns each read to a specific taxon by linking genomic substrings (k-mers) to the lowest common ancestor of the taxa with a similar k-mer content. Reads were classified against a database of reference sequences available at RefSeq (https://www.ncbi.nlm.nih.gov/refseq/), which includes protozoa, archaea, fungi, bacteria, viruses, and human genomes. Pavian v4.1.0 (https://fbreitwieser.shinyapps.io/pavian/) was used to compare the classification results from each sample.

2.7.5. Variant detection in drug targets and in regulatory genes that are responsible for the expression of important housekeeping genes

The variants were annotated using SnpEff v5.0 (Cingolani *et al.,* 2012) which annotates and predicts the effect of each variant in the genome. The variants were manually inspected from the BAM files using Interactive genome viewer (IGV) v2.11.0 (Robinson *et al.,* 2011).

2.7.6. Data analysis

GraphPad prism8 software was used to construct the graphs in this study. The statistical significance between the depth of coverage between the two data groups was assessed using the Mann-Whitney test for unpaired samples, and the statistical significance between the depth of coverage between the three data groups was assessed using a one-way ANOVA test on GraphPad.

2.7.7. *In-vitro* mixed strain *M. tuberculosis* infection model

To understand strain behaviour in competitive environments, an artificial in vitro model that mimics infection with mixed strains of *M. tb* was developed. A competitive environment was created by growing each strain in different ratios and under the same growth conditions to investigate within-host strain behaviour. The mixed model of infection was tested on clinical isolates of LAM and Beijing stains due to the high prevalence of these strains in South Africa. Furthermore, Beijing and Non-Beijing strains have been shown to be responsible for most clinical mixed infections (Maguga-Phasha *et al., 2*017, Victor *et al., 2*004, Warren *et al., 2*004). Two Beijing and two LAM clinical strains identified by spoligotyping (section 2.6.1) were selected for analysis in the mixed model of infection.

2.7.8. Genomic insertion of genetic barcodes into the Beijing and Lam strains

In order to identify the Beijing and Lam strains under mixed culture conditions, unique DNA sequences (barcodes) were designed for each strain and inserted into the multiple cloning site of the pUC57-Kan integrating vector (Cheo *et al.*, 2004). Genetic tags (DNA barcodes) for individual *M. tb* strains were designed and cloned into the multiple cloning site of the pUC57_Kan integrating vector. The plasmid containing barcode 1 was denoted as pUC57_C2M, whilst the plasmid containing barcode 2 was denoted as pUC57_C4S. The 50 bp barcodes were inserted into the bacterial chromosomes via a site-specific hybrid recombination event at the phage attachment site (attB) of the target genome by electroporation. The cloning strategy is demonstrated in Appendix C, Figure 1. All barcode labelled plasmids used in this study (Table 2.3) were supplied by GenScript Biotech (USA). These plasmids were propagated in *E. coli* cells and the purified plasmid was then electroporated into the *M. tb* clinical strains.

Table 2.3 - Cloning vectors used for the generation of barcoded strains

Vector	Characteristics	Barcode number	Barcode sequence (5' to 3')	Vector size (bp)	Reference
C2-M_pUC57-Kan	Mycobacterial integrating vector; *oriE, Km^r, attP*	1	CACCGACTTCTTGTTCCCGAATGACACGCGTCCCCCTGCGGGTAGGTCGC	2951	This study
C4-S_pUC57-Kan	Mycobacterial integrating vector; *oriE, Km^r, attP*	2	CTGCCTGGCAGCTCTAATCTCTCGCCACGGCCGTAAAAGCCTCCAAGAGT	2951	This study

oriE = origin of replication, Km^r=kanamycin resistant, attB= attachment site to host. 2.7.8.1. Preparation of competent *Escherichia coli* cells by $CaCl_2$

Chemically competent *E. coli* DH5α cells were prepared using the calcium chloride method (Chang *et al.,* 2017). A starter culture was prepared by inoculating 1 mL of *E. coli* DH5α into 5 mL of 2X TY (Tryptone yeast) and grown at 37°C with shaking overnight. The culture was diluted 1/200 in 25 mL of 2XTY and grown at 37°C with shaking to an OD_{600nm}= 0.3-0.6. Bacterial cells were harvested by centrifugation at 4317 x g at 4°C for 5 minutes. Thereafter, the cells were resuspended in 1 culture volume of ice cold $MgCl_2$ (0.1 M) to alter membrane permeability by charge neutralisation. After incubation on ice for 60 seconds, the cells were harvested as above, and resuspended in ½ culture volume of cold $CaCl_2$ (0.1 M). This was followed by incubation on ice for 30-60 minutes to increase the competency for efficient uptake of DNA. The cells were harvested once more and resuspended in 1/10 culture volume of $CaCl_2$ (0.1 M) and used immediately for transformation of the plasmids.

2.7.8.2. Transformation of chemically competent *Escherichia coli*

The chemically competent *E. coli* cells were mixed with 100 ng/µL of plasmid DNA and incubated on ice for 10 min. The cells were transformed by heat shock at 42°C for 90 seconds. Thereafter, 900 µL of pre-warmed 2X TY was added to the cells and incubated at 37°C for 1 hour to allow for recovery and phenotypic expression. The cells were then spread on 50 µg/ml of kanamycin supplemented LA plates and incubated at 37°C overnight. Several transformants were selected to screen for the presence of the respective plasmids following isolation from the *E. coli* cells.

2.7.8.3. Small scale plasmid DNA extraction from *E. coli* DH5α

The QIAprep Spin Miniprep Kit (QIAGEN, USA) was used for small scale plasmid DNA extraction following the manufacturer's instructions. The selected *E. coli* transformants were grown in 5 mL of LB media supplemented with 50 µg/ml of kanamycin at 37°C overnight. The cells were harvested by centrifugation at 12 470 x g for 5 minutes and resuspended in 250 µL of Buffer P1 (with RNase A). Cell lysis was promoted by the addition of 250 µL of Buffer P2 followed by the addition of 350 µL of Buffer N3. The suspension was mixed by inversion. Thereafter, the cell mixture was centrifuged at 12 470 x g for 10 minutes to separate the plasmid DNA from the cell debris. The supernatant containing the plasmid DNA was transferred to a QIAprep 2.0 spin column and centrifuged at 12 470 x g for 60 seconds to harvest the DNA. The flow through was discarded and the DNA was washed by the addition of 500 µL of PB Buffer and centrifugation at 12 470 x g for 60 seconds to remove any remaining salts from the precipitation step. The wash step was repeated with a further 750 µL of PE Buffer and centrifugation at 12 470 x g for 60 seconds. After discarding the flow

through, the columns were centrifuged at 12 470 x g for 60 seconds to remove any residual wash buffer. The spin column was transferred to a sterile 1.5 mL microtube and 25 µL of Buffer EB was added to the column. The column was incubated at room temperature for 60 seconds before eluting the plasmid DNA by centrifugation at 12 470 x g for 60 seconds. The isolated plasmid DNA was checked for genomic integrity by restriction enzyme analysis.

2.7.8.4. Restriction digestion of plasmid DNA

Restriction enzymes cleave DNA at specific sequences. The isolated plasmids were therefore screened by several different restriction enzymes to confirm if the constructs were in the correct conformation and retained integrity. All restriction enzymes used in this study were purchased from New England Biolabs (NEB) or Roche Applied Sciences. Clone Manager Suite 9 software (Scientific and Educational Software, USA) was used to select the enzymes and predict the expected fragment sizes with the respective restriction enzyme. Digestions were carried out in a final volume of 20 µL per reaction which contained 2 µL buffer, 1 µL enzyme, 9 µL nuclease free water and 8 µL of DNA. The DNA was digested at 37°C for 1 hour. The DNA fragments were separated on a 1% agarose gel and visualised as described in section 2.5.2. Once the plasmids were confirmed to be correct, 1 ml of freezer stocks of the respective cultures were made in 66% glycerol (1:1 v/v).

2.7.8.5. Large scale plasmid DNA extraction from *E. coli* DH5α

The NucleoBond® Xtra Midi/Maxi plasmid purification kit (Macherey-Nagel, Germany) was used for large scale plasmid DNA isolation and purification following the manufacturer's instructions. The glycerol stocks of the *E. coli* cells containing the plasmid was inoculated in 5 mL of LB media supplemented with 50 µg/ml of kanamycin and grown at 37°C overnight. The preculture was diluted 1/1000 in 100 mL of LB media supplemented with 50 ug/ml of Kanamycin and grown at 37°C with shaking overnight. The stationary phase cells were harvested by centrifugation at 12 470 x g at 4°C for 10 min and the pellet was resuspended in 8 mL of Buffer RES (with RNase A). Thereafter 8 mL of Lysis Buffer LYS was added to the suspension and mixed by inversion and the tubes were incubated for 5 minutes at room temperature to allow for cell lysis. The pH of the solution was neutralised by the addition of 8 mL of Buffer NEU (neutralising buffer) to denature the proteins. Prior to loading the lysate to the NucleoBond® column filter unit, the filter was equilibrated with 12 mL of Buffer EQU. Once the DNA was bound, the column was washed twice with 8 mL of Buffer WASH (wash buffer). The plasmid DNA was eluted with 5 mL of Buffer ELU (elution buffer). The eluate was collected in a 50 mL Falcon tube to which 3.5 mL of isopropanol was added. The

plasmid DNA was precipitated by centrifugation at 12 470 x g for 15 minutes. The DNA pellet was washed with 2 mL of 70% ethanol by centrifugation at 15 000 x g for 5 minutes. The supernatant was discarded and the pellet was dried using a SpeedVac (Thermo Fischer Scientific^{TM}) at 60°C for 10 minutes. The pellet was resuspended in 50 µL of 1x TE buffer and used for transformation in *M. tb*.

2.7.8.6. Electroporation of plasmid DNA into *Mycobacterium tuberculosis*

The purified plasmids containing the barcodes were electroporated (Fiedler and Wirth, 1988) into the two LAM and two Beijing *M. tb* strains that were selected for the model. Electroporation was performed for the integration of the genomic barcodes into the chromosomal genome. A 1 mL freezer stock of each strain that was stored from the decontamination process in section 2.3.2 was grown in 9 mL of 7H9 media at 37 ˚C to OD_{600nm}= 0.5-0.8. Then 1 mL of the preculture was inoculated in 50 mL of 7H9 media and grown at 37 ˚C to OD_{600nm}= 0.8-1. Glycine (1.5%) was then added to the cultures and incubated at 37˚C for 16 hours to weaken the cell wall and enhance transformation efficiency. After this incubation period, the bacterial cells were harvested by centrifugation at 4317 x g for 10 min and the pellet washed by resuspension in 40 mL of 10% glycerol and centrifugation at 4317 x g for 10 min. The wash step was repeated 2 more times before the pellet was resuspended in 2 mL of 10% glycerol. The competent cells were used immediately for electroporation. An amount of 5 µg of the plasmid DNA was UV irradiated in a UV stratlinker 1800 (Stratagene^{R}) at 100 mJ/cm^2. The irradiated DNA was added to 400 µL aliquots of the competent *M. tb* cells and the suspension was transferred to a 0.2 cm electroporation cuvette (Bio-Rad laboratories). The cells were pulsed using the BIO-RAD GENE PULSER XCell system using the following parameters: Voltage = 2500V, Time constant= 25 µF, Resistance = 1000Ω and Distance = 0.2 cm. Eight hundred µL of 7H9 media was immediately added to the electroporated cells and the suspension was transferred to a 1.5 mL O-ring screw cap tube. The cells were incubated at 37°C for 16 hours to allow for recovery and phenotypic expression, before spreading them on Middlebrook 7H11 agar plates containing the appropriate antibiotics. The clinical samples were spread on Middlebrook 7H11 agar plates supplemented with kanamycin, while the control samples were spread on Middlebrook 7H11 agar plates supplemented with hygromycin. The plates were incubated at 37 °C for 2-5 weeks before scoring the colonies. Once colonies emerged for each of the clinical strains containing the barcoded plasmids and the control samples, multiple clones

were picked and grown in 9 mL of 7H9 media to an OD_{600}nm= 0.5 to make 1 ml aliquots of working stocks for downstream experiments.

2.7.9. Isolation of genomic DNA

The stored 1 mL aliquots of the transformed *M. tb* cells were grown in 9 mL of 7H9 media at 37°C to an OD_{600nm}= 0.8-1. Genomic DNA was isolated from these cultures using the CTAB/NACL method as described in section 2.4. The extracted DNA was quantified, and the integrity was assessed as described in section 2.5. The DNA was used for the confirmation of the stable integration of the plasmids into the chromosome at the phage attachment site

2.7.10. DNA amplification for the confirmation of chromosomal DNA integration by PCR

Polymerase Chain Reaction amplifies a specific region of a DNA target. This procedure was utilised for the confirmation of the strain genotype. Primers that are specific for the DNA barcodes (Table 2.4) were supplied by Inqaba Biotech Company. A gradient PCR was carried out for primer optimisation. The PCR reactions were made to a volume of 20 µL with the following components: 2 µL of 10x PCR Buffer, 1.6 µL of 25 mM $MgCl_2$, 4 µL of 5x GC-Rich Buffer, 0.4 µL of 10mM dNTPs, 2 µL (10 µM) forward and reverse primers each, 2 µL extracted DNA (concentration of 50 ng/µL), 0.16 µL of Roche Faststart Taq DNA polymerase and 5.84 µL of Hyclone nuclease-free water. A non-template control was included in the PCR reaction to rule out non-specific amplification or DNA contamination.

Table 2.4: PCR primers used for the amplification of target sequences

Name	Primer sequence 5'→3'	5' modification	Length (bp)
Forward primer	GAGGAGAAGACTGCTGACCT	None	20
Reverse primer	CAGGCCTTGCTGTTCCC	None	17

The reactions were carried out using the BioRad laboratories thermo cyler at the following conditions: Initial denaturation at 95°C for 4 minutes for heat activation of DNA polymerase, denaturation of the double stranded genomic DNA at 95°C for 30 seconds, annealing of primers to the DNA target (barcodes) at 70°C,68°C,66°C,64°C and 62 °C for 30 seconds, initial elongation for complementary strand synthesis at 72°C for 10 seconds and a final elongation step at 72°C for 7 minutes.

2.7.11. Analysis of PCR products by high percentage agarose gel electrophoresis

A 4% agarose gel was used to confirm the integration of the DNA barcodes into the *M. tb* chromosome. For preparation of the 4% gel, 1.2g of the agarose gel powder was dissolved in 30 mL of 1x TAE. The gel was prepared and visualised as described in section 2.5.2.

3. RESULTS

3.1. Isolation of genomic DNA

The success of molecular investigations is based on the quality and purity of the DNA used; hence, the isolation of high-quality DNA is an important step. The objective of the first component of the study was to re-grow *M. tb* strains from previously stored patient samples for the isolation of genomic DNA. The total sample size of 100 consisted of sputum samples from ten patients, each of which had been grown in nine different culture conditions (Figure 2.1), in addition to the uncultured sputum sample. Of these samples, only 96 were available for use in this study as four samples (59161 CFU, 59162 CFU, 59165 CFU and 59152 RPF-AB) were not stored in the parent study. Genomic DNA was isolated from the available samples following sample decontamination and purification processes (as described in section 2.3) and the quantity and purity (as measured by the A260/280 ratio) of the DNA was assessed. Spectrophotometric quantification and purity assessment results are summarised in Appendix D, Table 1. The DNA concentration varied for the various samples (range 2.90 - 651.01 ng/µL) with high concentrations observed for samples previously grown using the MPN assay (average = 178.3 ng/µl). As expected, the concentrations of DNA extracted directly from sputum samples was the lowest (average= 81.1 ng/µl). This could be attributed to varying sputum quality and sample diversity, both of which have an effect on DNA concentration (Riethmueller *et al.*, 2008). With the criterion that pure DNA samples should have an A_{260}/A_{280} ratio between 1.8-2.00, 30 (31%) samples did not fall within this desired range. A ratio of < 1.8 was obtained for these samples suggesting the presence of protein, or other contaminants that absorb strongly at 280 nm (Wilfinger *et al.*,1997). These samples were analysed on an agarose gel to confirm the presence of these contaminants prior to purification processes.

3.2. Qualitative evaluation of DNA extracted from mycobacterial cultures

The integrity of the isolated genomic DNA was evaluated using agarose gel electrophoresis. A solid band on the agarose gel confirmed the presence of quality nucleic acid in the sample, whereas the absence of a band indicated a failed extraction. All samples displayed same sized high molecular weight genomic DNA bands (Figure 3.1) with the exception of 3 samples that were culture negative and had failed DNA extractions (indicated by asterisk in Figure 3.1). The lack of growth for these three samples during culture indicated that the bacterial cells had lost viability. However, DNA extraction was attempted on these samples to obtain the genomic DNA since DNA should still be present in the samples even after cells have died.

Band intensity differed between samples depending on the DNA quantity and streaking of the DNA in a few samples suggest the presence of sheared DNA or could be a consequence of a high concentration of DNA. For samples that did not achieve the desired A_{260}/A_{280} ratio, there was no evidence of impurities in the DNA on the gel suggesting that the nanodrop is not always an accurate indicator of DNA purity. Even though the ratio of these samples did not fall within the desired range, the extracted DNA appears acceptable and the samples were not further purified. The DNA was used for downstream analysis.

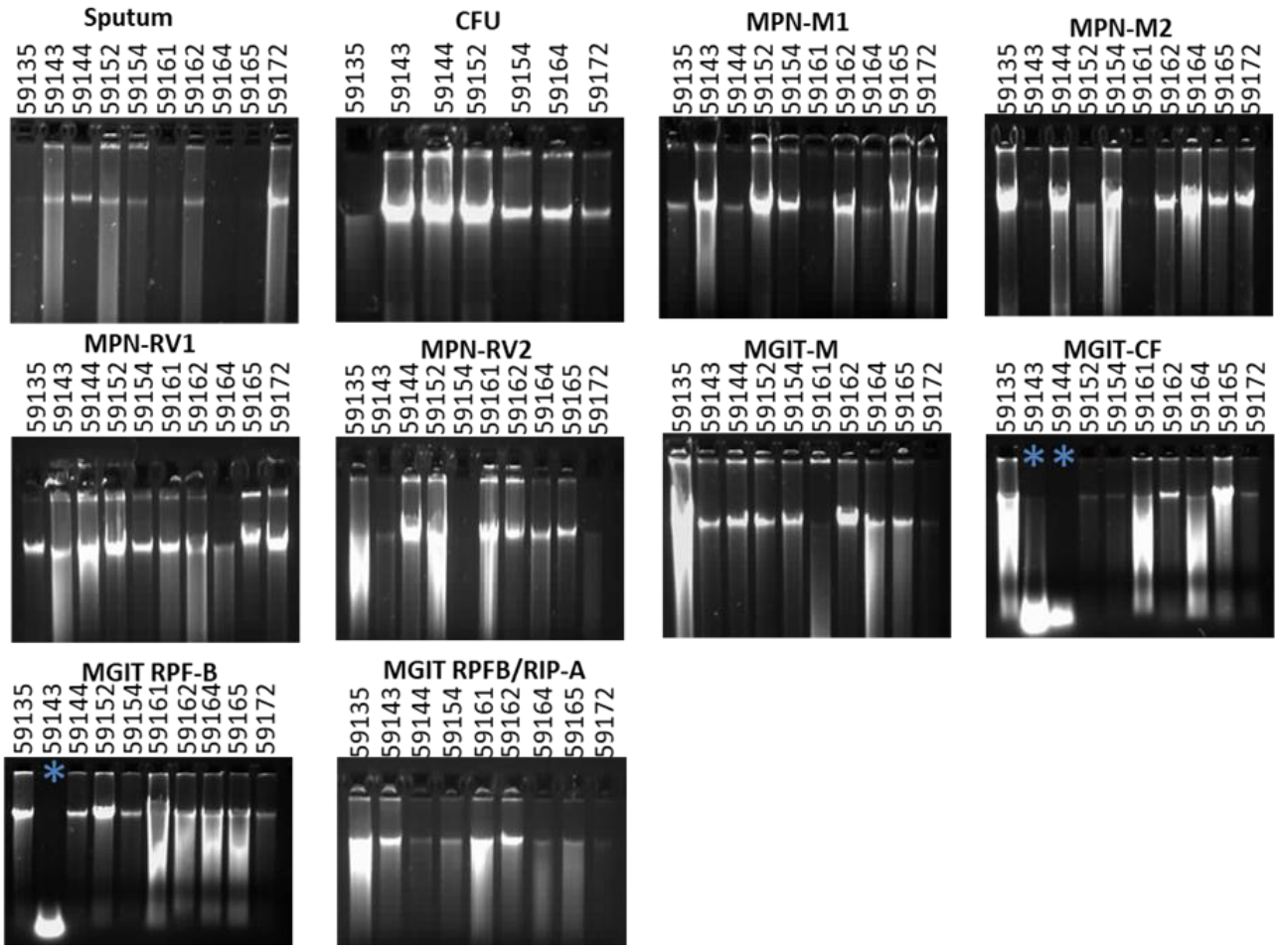


Figure 3.1: Agarose gel electrophoresis results of extracted genomic DNA from direct sputum and sputum cultured under various growth conditions. Ten µL of each sample was electrophoresed on a 1% agarose gel at 90V for 45 minutes. Stained DNA on the agarose gels was visualised using a UV gel imager. Blue asterisks represent samples with a failed DNA extraction which were excluded from further analysis.

3.3. Spoligotyping analysis of *M. tuberculosis* clinical samples

The *M. tb* clinical samples from the study population were genotyped using the isolated DNA for the classification of strain types. Spoligotyping differentiates strains based on the presence or absence of the spacer sequence in the direct repeat region of the *M. tb* genome (as

described in section 2.6). The strain type distribution of the 93 *M. tb* samples was dominated by lineage 4 strains (58%), but strains from lineages 3 (18%) and 2 (24%) were also present. An unusual instance of an infection with the laboratory H37Rv strain (Lineage 4) was observed in 4 (4/93) of the samples. Strain type distribution by lineage and sub lineage is summarised in Figure 3.2 A and the spoligotyping data is shown in Figure 3.2B.

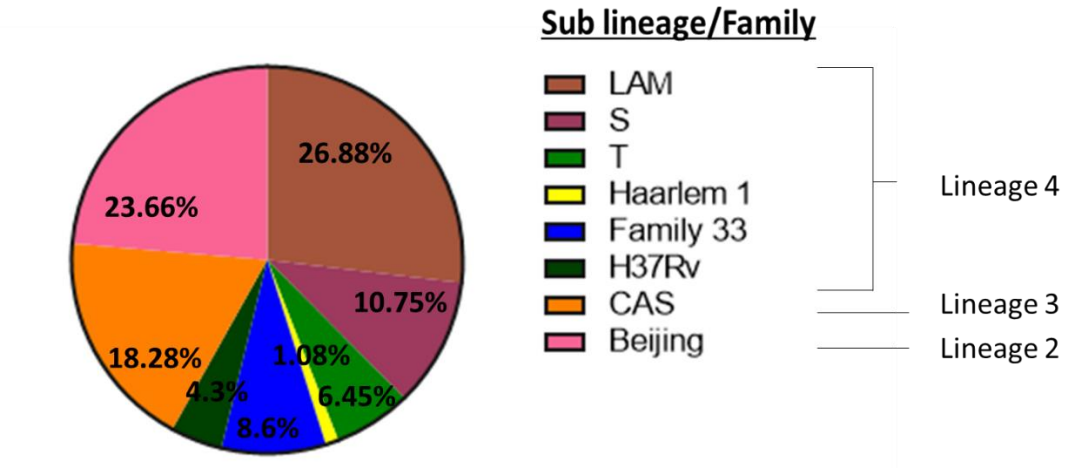


Figure 3.2A: Strain type distribution of the 93 genotyped *M. tuberculosis* samples. Pie chart demonstrates the distribution of the strains identified by spoligotyping by Lineage and sub lineage.

The spoligotyping data shown in Figure 3.2 B demonstrates the spoligotype patterns that were obtained for each sample during genotyping. The original spoligotyping blots are shown in Appendix C, Figure 2. The four samples that have been highlighted in red in Figure 3.2 B show cross contamination with the laboratory *M. tb* H37Rv spoligotype pattern. As only one of these samples was grown in the presence of CF from *M. tb* H37Rv, it is possible that the contamination observed in the other samples was caused by bacteria escaping during culture filtrate preparation. Nine out of the 10 patients' samples that were grown under the various laboratory conditions (59135, 59144, 59151, 59152, 59161, 59162, 59164, 59165, 59172) revealed spoligotype patterns that differed from the original sputum sample that did not undergo any culturing in the laboratory (Figure 3.2 B). These data suggests that the majority of the patients analysed in this cohort are infected with more than one *M. tb* strain. Only one patient was infected with a single strain (59143), whilst four were infected with two strains (59144, 59152, 59154, 59165) and 5 with three strains (59135, 59161, 59162, 59164, 59172).

Overlapping spoligotype patterns were observed in 50% (59135, 59154, 59161, 59162, 59165) of the uncultured sputum samples, whilst only 5% of the cultured samples (59144

MPN_M2, 59154 MPN_M1, 59165 MPN_M1 and 59165 MPN_RV1) displayed a mixed pattern. Of these mixed samples, 89% (8/9) were classified as belonging to the Family 33 sub lineage by spoligotype analysis. All Family 33 strains in this cohort displayed different spoligotype patterns (Figure 3.2 B). The Family 33 spoligotype pattern obtained from the uncultured sputum of patient 59135 is identical to the pattern that is displayed from an overlap of the T1, T2 and LAM 3 strains that have been identified in the patient. Additional spacers, spacer 35 and 36, were also detected in sputum. Similarly, the spoligotype pattern of the uncultured sputum of patient 59162 is identical to the pattern that is displayed from an overlap of the LAM 3 and Beijing strains that have been identified in the patient, with additional spacers detected in sputum (spacers 9-11, 20, 22-24). Hence, multiple strain infections can result in false spoligotype patterns (Sanoussi *et al., 2*017). False negative spacers in sputum (spacers present in cultured samples but not in sputum) were observed in the sputum sample of patient 59161's, which is an overlap of the Beijing and T2 strains identified in the patient. The spoligotype pattern for the cultured sputum sample of patient 59144 MPN_M2, when analysed in SPOTCLUST does not display a Beijing strain, however, the presence of both spacers 35 and 36 that distinguish the Beijing strain are present for this sample, which may suggest the presence of a Beijing strain, perhaps at a low level. The cultured sputum sample (MPN_M1) of patient 59154 confirms the presence of an additional strain in sputum that was not detected under any of the other growth conditions. The additional spacers seen in this sample are also present in the uncultured sputum sample. However, SPOTCLUST did not detect this additional strain. The mixed patterns observed in the uncultured sputum and sputum cultured at MPN_M1 and MPN_RV1 of patient 59165 suggest that the patient is co-infected with the laboratory strain H37Rv, and the Beijing strain. The reason for this observation in sputum is not completely understood since the original sputum sample was not cultured in the laboratory. However, for the cultured samples, it may be attributed to a cross-contamination event with H37Rv-CF since the MPN assay was supplemented with CF from H37Rv. An *in-silico* analysis was used to validate the strain types identified by spoligotyping.

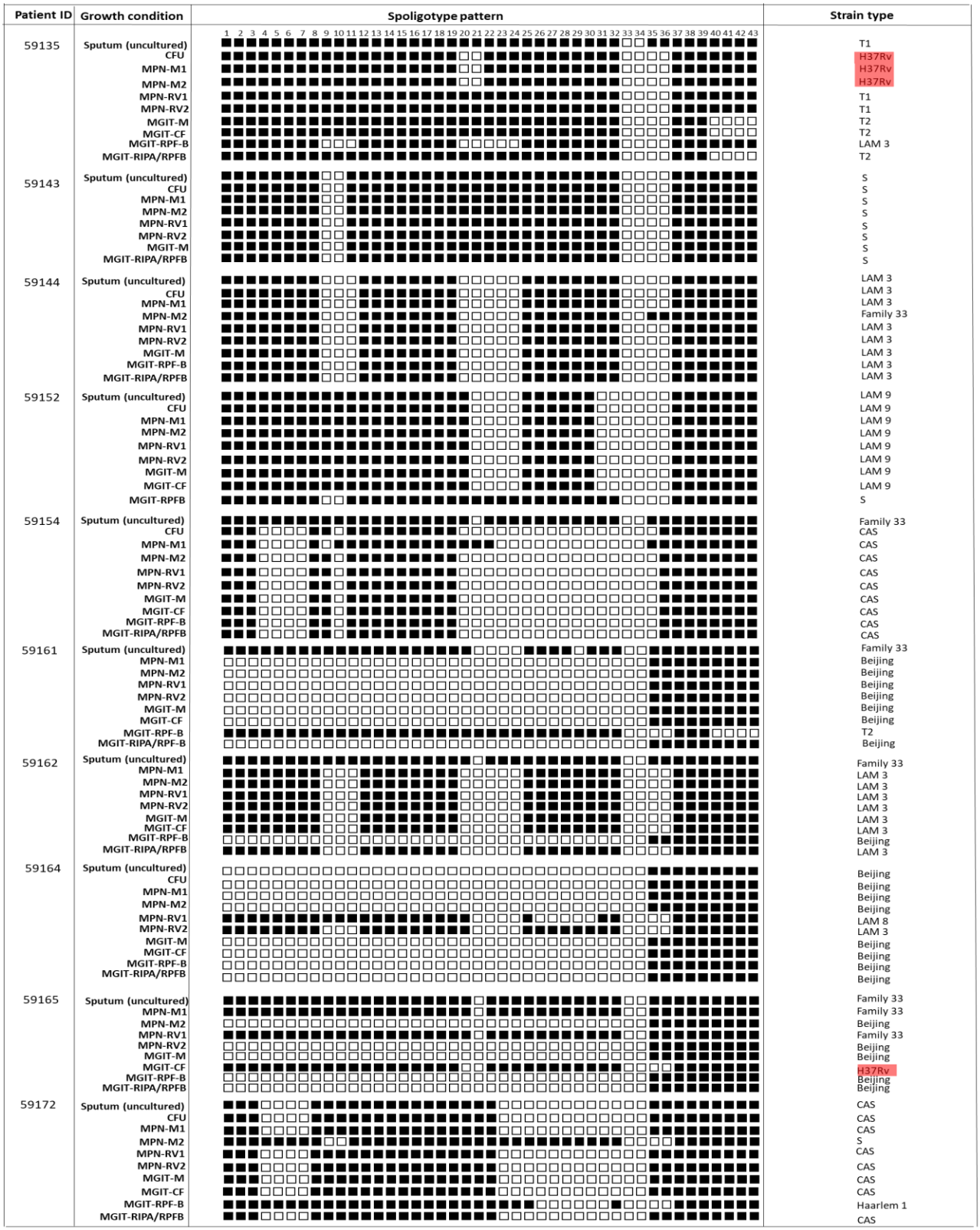


Figure 3.2B: Paired spoligotype patterns and strain types of *M. tuberculosis* clinical samples. Spoligotype patterns derived from baseline sputum and sputum cultured across various growth conditions from ten *M. tb* infected patients. Samples highlighted in red are contaminated with the *M. tb* H37Rv reference strain. The variability found in the direct repeat (DR) locus, which contains a variable number of short direct repeats, is used to classify *M. tb* strains by strain type. The presence or absence of these repeats at 43 spaces in the DR locus determines the spoligotype family. Black dots represent the presence of a spacer sequence at the specific position. Patterns were translated into lineages using the SPOTCLUST software. Spacer sequences were read from the left to right to determine the strain type.

3.4. Next generation sequencing of genomic DNA

Whole genome sequencing provides information on the entire genetic profile of an organism for analysis. The effect of sputum culture in different growth conditions on the within-sample genetic diversity was investigated in this study by a comparison of the genomic DNA sequences obtained from sputum cultured in various growth conditions.

3.4.1. *In silico* analysis of strain spoligotype

Strain spoligotype prediction accuracy was assessed by comparing experimentally predicted octal codes and strain spoligotypes to *in silico* predicted octal codes and strain spoligotypes derived from the DNA sequencing data of the *M. tb* samples. The *in silico* results are more reliable in predicting strain spoligotypes because they reduce the potential effects of experimental errors, which could result in inaccurate readings. Paired end genomic sequences of the clinical samples were obtained using the Illumina HiSeq next generation sequencing platform from Rutgers University in Newarck, Columbia. The forward DNA sequencing reads of each sample were used to generate the *in silico* octal codes using SpolPred, which were then translated into strain types by SPOTCLUST. SPOTCLUST was also used to generate experimental octal codes derived from the experimental strain spoligotype patterns in section 3.3. Table 3.1 summarises the results of the comparison. Strain spoligotypes inferred from the *in silico* analysis matched with the experimental results for 77/93 samples (83%) (Green). Four samples (4%) had different *in silico* and experimental results, (Red) whilst 12 samples (13%) were indeterminate (Blue). Detailed results are summarised in Appendix D, Table 2. Of the four non-matched samples (red blocks), sample 59162 (MGIT RIPA/RPFB) showed the presence of *Mycobacterium microti* based on whole genome sequencing (WGS) data, suggesting that this patient was most likely co-infected with another member of the MTBC. The detection of infections with mycobacteria other than tuberculosis (Mott) is usually possible in genotyping techniques that are able to detect mycobacteria at species level. There was a discrepancy between the uncultured sputum sample and the sputum cultured at MPN M1 for patient 59144, where experimental spoligotyping detected a LAM 3 and a Family 33 strain in each sample, but *in silico* analysis detected a Beijing strain and a Family 35 strain. Although the Beijing strain was not visible on the experimental spoligotype pattern of the uncultured sputum sample (Figure 3.2 B), the presence of spacers 35 and 36 which are characteristic of Lineage 2 strains in the spoligotype pattern of the sample upon culture at the MPN_M1 condition, suggest that the patient may be co-infected with a Beijing strain as identified by WGS. In the uncultured sputum of patient 59152, a

LAM 7 strain was detected rather than a LAM 9 strain. A 1-digit variation of 1 to 3 was observed in the 12th number of the 15-digit octal code for the CAS strains of patient 59172 (highlighted in pink in Appendix D, Table D2). There was no correlation in the samples that were indeterminate except for the majority of them being the uncultured sputum samples.

Table 3.1: Matched, mismatched, and indeterminate *in silico* predicted spoligotypes

	Sputum	CFU	MPN _M1	MPN _M2	MPN _RV1	MPN _RV2	MGIT_ M	MGIT_ CF	MGIT_ RPFB	MGIT_RIP A/RPFB
59135										
59143										
59144										
59152										
59154										
59161										
59162										
59164										
59165										
59172										

Green blocks indicate matched *in silico* and experimental results, red blocks represent mismatched results, and blue blocks represent undetermined *in silico* results. Blocks with diagonal lines represent samples that were unavailable for analysis.

3.4.2. Quality control

Pre-processing and quality control are critical steps in NGS data analysis to ensure that high quality data are used in the analysis. Both forward and reverse DNA sequencing reads from each sample were quality assessed using FastQC. FastQC utilizes specific metrics to provide information on the quality of the sequencing data. The Phred scale was developed in the Human Genome Project to assess the base quality of nucleotides generated by automated sequencing and is now used to represent confidence scores in other contexts of genome science (Schmutz *et al., 2*004). In the context of FastQC, Phred quality scores represent the likelihood that a base was called incorrectly during sequencing. The analysis was carried out on all sequenced samples. As example, Figure 3.3 depicts the quality scores (in Phred scale)

across all individual bases of the 150 bp reads for samples 59162 MPN-RV2 (A&B) and 59135 MPN-RV1 (C&D), in the forward and the reverse orientation, respectively. The mean value (blue horizontal line) indicates that the quality scores decreased towards the end of the reads which is expected due to strand desynchronisation during nucleotide incorporation (known as dephasing) as the number of cycles increases (Nakamura *et al.*, 2011). This is common in sequencing by synthesis DNA technologies. Despite this decrease, the average Phred quality score of all samples at each nucleotide base was > 30, which translates to a 0.001% sequencing error probability. This indicates that the sequencing data was of excellent quality.

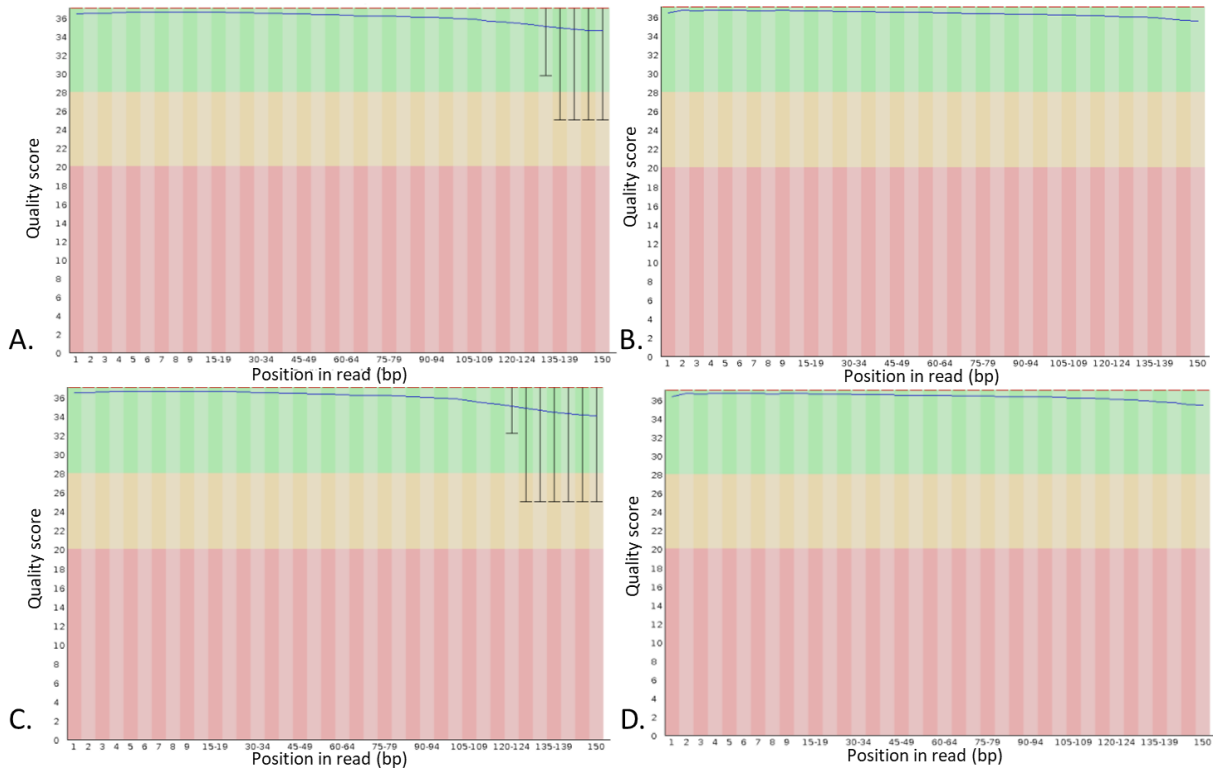


Figure 3.3: Per base sequencing quality of paired end sequencing reads. Quality score of individual nucleotides in the forward (**A&C**) and reverse (**B&D**) reads for samples 59162 MPN-RV2 (A&B) and 59135 MPN-RV1 (A&C) respectively. The graph separates the quality score of the bases as being poor (red, Phred 0-19), reasonable (orange, Phred 20-28) and of very good (green, Phred 29-40) quality. The blue line indicates the mean quality score across the individual bases within the reads. Black vertical lines indicate the 90th percentile.

3.4.3. Genome Coverage

The sequencing data was processed using a bioinformatics data analysis pipeline, as detailed in section 2.7.1.1. The coverage statistics obtained after read mapping to the *M. tuberculosis* H37Rv reference genome were analysed. In DNA sequences derived from genomic DNA

obtained from sputum, the coverage depth ranged from 0.5818x-204.53x which correlated with the proportion of mapped reads (coverage breadth) (range 0.38%-60.01%). No correlation was observed between DNA concentration and coverage depth. For example, the concentration of DNA in the sputum of patient 59172 (466.19 ng/l) was more than 200 times that of sample 59164 (2.09 ng/l) (Figure 3.1), but sample 59164 had a greater depth of reference genome coverage. For both samples, the average depth of coverage was 1.118x and 204.530x, respectively. Coverage was significantly higher in sequences obtained from DNA isolated from cultured samples under all growth conditions (129.734x - 696.525x, 95.46% - 99.82%) compared to the uncultured sputum samples ($p < 0.0001$) (Figure 3.4 A). Comparison of these cultured sequences indicated that coverage was the highest in samples that were grown in the CFU assay compared to those grown using both the MPN ($p < 0.0001$) and MGIT (0.0012) assays (Figure 3.4 B). For whole genome analysis, sequences are required to have at least a 30x depth of coverage and at least 50% of reads that map to the reference genome (Genestet *et al., 2*019a). Only 1 of the 10 sputum derived sequences met this criterion hence, for the remaining 9 sputum samples partial genome analysis was conducted. The nine samples may not have met the criteria due to low *M. tb* loads in sputum, as well as the presence of host DNA and other microorganisms in sputum. Eleven sequences from the cultured samples had a coverage breadth of < 95% based on the reference genome and therefore, these were excluded from further analysis. The high depth and breadth of coverage, combined with the high quality of the sequencing data, provided a high degree of confidence in detecting variants in the genomes of the various *M. tb* strains under study.

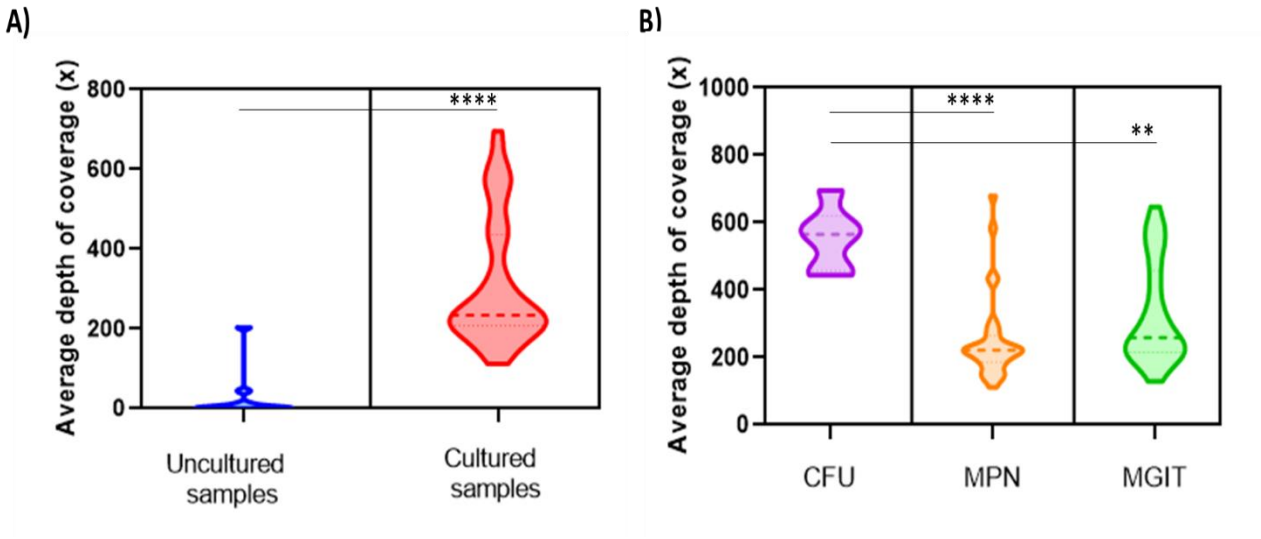


Figure 3.4: Depth of genome coverage. Violin plot demonstrating the average depth of the reference *M. tb* genome coverage of sequences obtained from sequenced *M. tb* isolated from **A)** uncultured sputum samples and sputum cultured under various growth conditions ($p < 0.0001$****) and **B)** sputum cultured using the CFU, MPN and MGIT assays ($p < 0.0001$****, $p=0.0012$**).

3.5. Metagenomic analysis of *M. tuberculosis* sputum samples

Prior to variant detection, a metagenomic analysis of the uncultured sputum samples was performed to assess the proportion of contaminating DNA sequences (from oral microflora or human DNA) that might produce false variants. The number of DNA reads in the sputum metagenome ranged from 4 460 401-6 305 552 (Table 3.2). All samples, with the exception of 59161, had a high percentage (> 80%) of taxonomically classified reads indicating that the database of reference genomes that was used was large enough to allow for read classification in our study. The majority of the reads were classified against the chordate and bacterial databases, with no reads shared with fungal or protozoan genomes. The number of reads assigned to viral genomes is negligible. Differences were observed in non-mycobacterial read contamination in each sputum sample. Of the 1,119 genera identified, *Actinomyces*, *Homo*, *Klebsiella*, *Mycobacterium*, *Prevotella*, *Rothia*, *Staphylococcus* and *Streptococcus* were the most abundant in all sputum samples (Figure 3.5). These contaminating reads are typical as these species are prevalent in the human oral flora and respiratory cavities (Dekaboruah *et al.*, 2020). As expected, human DNA contamination was observed in all sputum samples but it was prevalent in four samples (59143, 59152, 59162 and 59172), with the chordate phylum accounting for more than 55% of the sequence reads (highlighted in orange in Table.3.2). Taxonomically, *Homo sapiens* are classified under the Chordate phylum. *M. tb* was recovered in the sputum samples in different proportions and a correlation was observed between *M. tb*

read abundance and coverage breadth. These results suggest that the chemical decontamination of the sputum samples was able to reduce contamination with microorganisms while preserving mycobacterial species, but failed to reduce human DNA contamination. Hence, two filter steps were performed: 1) removal of unmapped reads, and 2) removal of reads with a mapping quality of < 30, to allow for enhanced *M. tuberculosis* genome analysis.

Table 3.2: Table comparing the proportion of contaminating classified and unclassified reads for each of the sputum samples

Patient ID	Number of raw reads	Classified reads	Unclassified reads	Chordate reads	Bacterial reads	Viral reads	Fungal reads	Protozoan reads
59135	5 104 733	87.7%	12.3%	33.6%	54%	0.0227%	0%	0%
59143	6 083 195	98.7%	1.33%	92.3%	6.26%	0.0284%	0%	0%
59144	6 306 552	84.4%	15.6%	0.12%	84.3%	0.00962%	0%	0%
59152	4 640 401	97.4%	2.58%	69.8%	27.5%	0.0196%	0%	0%
59154	5 278 108	78.2%	21.8%	12.9%	65.3%	0.0142%	0%	0%
59161	4 530 996	57.7%	42.3%	1.39%	56.1%	0.00426%	0%	0%
59162	6 540 067	91.9%	8.13%	57.2%	33.4%	1.25%	0%	0%
59164	5 649 520	93.3%	6.67%	3.4%	89.9%	0.00595%	0%	0%
59165	5 649 062	85.2%	14.8%	33.3%	51.8%	0.0474%	0%	0%
59172	5 959 199	99.6%	0.44%	96.1%	3.39%	0.00404%	0%	0%

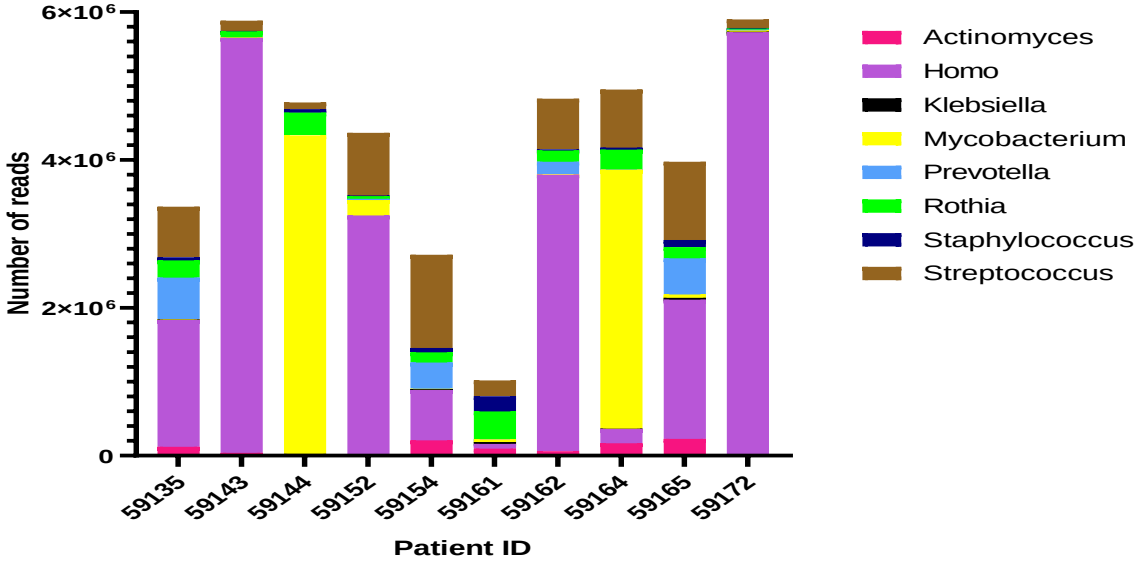


Figure 3.5: Taxonomic assignment of sequence reads in sputum samples. The proportion of contaminating reads were assessed from the raw sequence data files of the uncultured sputum samples using Kraken. Kraken assigns short DNA sequences into taxa by using the k-mer content of each sequence as a taxonomic identifier. For taxonomic classification, each species in the NCBI database has a unique taxonomic identifier that is matched with that of the DNA sequence reads. This figure depicts the presence of contaminating bacterial species in all sputum samples. The different colours represent the proportion of the various bacterial genera, respectively.

3.6. High confidence variant detection in the *M. tb* genome

The number of variants identified within a specific sample reflects the degree of genetic diversity between the genomes that are being compared. High confidence variants were called from the genomic sequences obtained from each patient's uncultured sputum and sputum cultured in various growth conditions for analysis. Heterozygous variants were additionally called from the uncultured sputum samples together with the cultured sputum samples that displayed mixed strain spoligotype patterns. The presence of heterozygous base calls at specific sites within the genome indicates the presence of alternative alleles at the site, confirming infection with multiple strains.

3.6.1 Detection of nucleotide polymorphisms in the sequenced genomes

This analysis included the filtered sputum samples (10) for partial genome analysis, as well as the cultured samples that met the selection criteria (> 95% coverage breadth). The analysis of both the sputum derived and cultured sequences yielded a total of 128 408 high quality SNPs, 2538 insertions, and 2503 deletions with respect to the *M. tb* H37Rv reference strain (Appendix D, Table 3). The majority of these mutations were missense upstream and downstream gene variants (intergenic). Variants detected from the cultured sequences ranged from 43-1333, and those identified from the raw sputum sequences ranged from 121-62 771 (Figure 3.6). A significant number of variants (62 771 bp) were identified from the sputum sample of patient 59144 (Not shown in Figure 3.6). Although this patient's sputum sample had the highest proportion of reads assigned to the genus *Mycobacterium* (Figure 3.4), only 0.95% (41008/4308442) of these reads were accounted for by *M. tb*. The majority (68%) of the reads were assigned to the species *Mycobacterium minnesotense*. As *M. tb* and *M. minnesotense* share sequence similarity, the large number of variants that were identified in this sample was most likely due to the high mapping of reads in regions of similarity, resulting in a large number of false positive variants. The data in Figure 3.5 show that sputum culture has varying effects on the within-sample genetic diversity. The number of variants detected in the uncultured sputum of patient 59135 (999 bp) and 59154 (1360 bp) exceeded the number of variants detected in any of their cultured samples across all growth conditions. In culture, patient 59143 had the highest number of base pair differences upon culture in the un-supplemented MPN assay (M1) (725 bp), whilst patients 59144 (834 bp) and 59161 (1308 bp) had the highest number of base pair differences upon culture in the CF-supplemented MPN assay (RV1). Genetic diversity increased upon culture of patient 59165's sputum sample in the un-supplemented MGIT assay (1320 bp), whereas culture in the CF-

supplemented MGIT assay increased the genetic diversity in both patient 59152 (764 bp) and 59162 (1304 bp). Both patient 59164 (1333 bp) and 59172 (1292) had the highest number of base pair differences upon culture in the CFU assay.

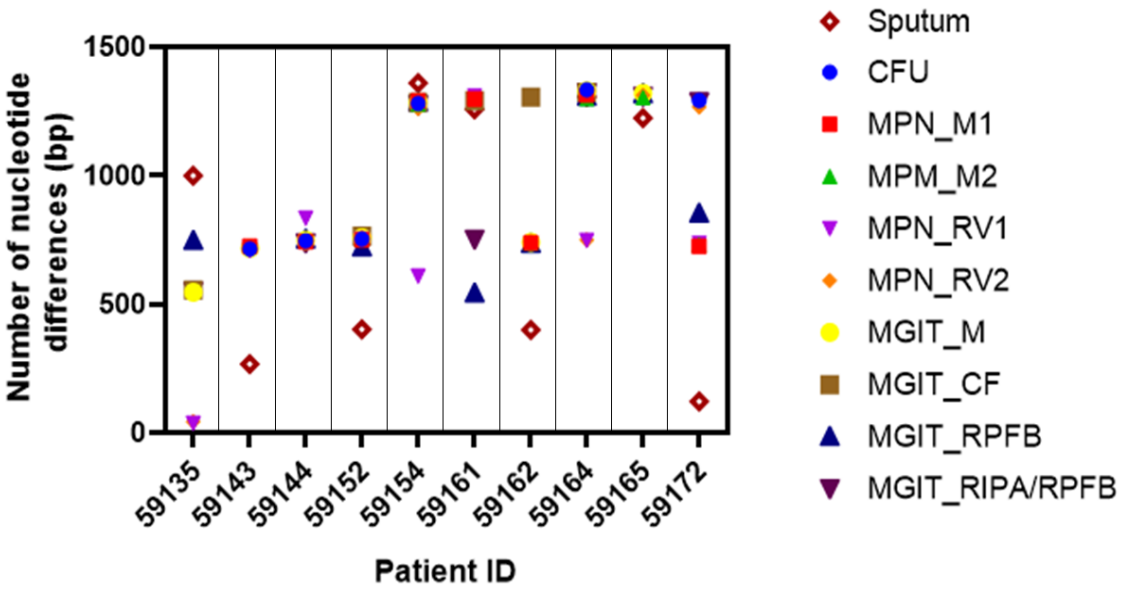


Figure 3.6: Variant selection across the various growth conditions. Depicts the number of variants detected from the comparison of the WGS data for each patients' samples under the various growth conditions to the *M. tb* H37Rv reference strain.

3.6.2. Detection of heterozygous nucleotide polymorphisms in the sequenced genomes

The presence of mixed strains in a sample was determined by the number of heterozygous variants found in the sample. Heterozygous variants were identified from all partially recovered genomic sequences derived from the uncultured sputum. The inclusion of the sputum sample of patient 59143, who had evidence of a single strain infection, was to distinguish mixed strain infections from single strain infections in sputum. Heterozygous variants were also identified in only one (59154 MPN_M1) of the four cultured samples that displayed mixed patterns from the spoligotyping data (Figure 3.2). The exclusion of the remaining three samples (59144 MPN_M2, 59165 MPN_M1, 59165 MPN_RV1) from the analysis was due to insufficient reference genome coverage breadth (< 95%) to allow for whole genome analysis. The number of heterozygous variants identified from these samples ranged from 13-381 SNPs (Table 3.3).

Using the heterozygous site method developed by Sobkowiak *et al.* (2018) which used a heterozygous SNP threshold of ≥ 20 SNP sites for the classification of mixed samples, 9/10 of the samples (shaded in blue in Table 3.3) were classified as having mixed strains. Patient

59164's uncultured sputum sample was classified as having a single infecting strain (not shaded in Table 3.3) since it did not meet the second criterion of being classified as having a mixture of strains (between 11-19 heterozygous SNP sites, and > 1.5% heterozygous to total SNP sites ratio). Patient 59143 who had evidence of a single strain infection by the same strain spoligotype pattern in sputum and in culture was classified as a mixed sample, and patient 59164 with evidence of a mixed strain infection by different strain spoligotype patterns in sputum and in culture, was classified as having a single infecting strain. The partial recovery of the genomic sequences from the uncultured sputum samples may account for this discrepancy. Patient 59154's MPN M1 cultured sputum was classified as having mixed strains, confirming that the mixed spoligotype pattern obtained from that sample was indeed the result of two infecting strains.

Table 3.3: Detection of mixed infections using the number of heterozygous SNPs and their proportion to the total number of SNPs identified in each patient's sputum sample.

Patient ID and growth condition	No of heterozygous SNPs	Total no of SNPs	Proportion of heterogenous SNPS to total SNPs
59135 Uncultured sputa	202	999	20%
59143 Uncultured sputa	83	266	32%
59152 Uncultured sputa	33	402	8%
59154 Uncultured sputa	355	1360	26 %
59161 Uncultured sputa	221	1257	18 %
59162 Uncultured sputa	106	400	27 %
59164 Uncultured sputa	13	1306	1 %
59165 Uncultured sputa	276	1223	23 %
59172 Uncultured sputa	24	121	20 %
59154 MPN_M1	86	1283	7%

Sputum samples that were identified as mixed using the heterozygous site method are shaded in blue

3.7. Phylogenetic construction of the genomic sequences

A phylogeny of the sequenced genomes was constructed using the Maximum likelihood approach to establish evolutionary relationships between the sequences according to their genetic distance (Figure 3.5). This analysis was carried out using fixed variants that were

identified in each sample. Bootstrap values infer the statistical confidence in the positioning of the branch nodes and the topology of the tree. In a single patient, strains that share a common evolutionary path and diverge into separate branches represent mixed strain *M. tuberculosis* infections which can be identified using phylogenetic tree analysis (Gan *et al., 2*016). The phylogenetic tree analysis of the sequences (Figure 3.6) correlated well with spoligotype analysis of the nine patients that displayed different strain spoligotypes across the various growth conditions. The samples in the phylogenetic tree clustered according to their respective lineages confirming the strain type distribution as identified by the spoligotyping data (section 3.3). Patient 59172's uncultured sputum and sputum cultured at MPN RV1 did not cluster according to lineage identified by spoligotyping (indicated with an asterisk in Figure 3.6), which could be due to the genetic distances between their genomes and high sequence similarity with its neighbouring sequences on the tree. Pairwise genetic distance comparisons between the genomes (measured in SNPs) indicate the evolutionary divergence of the sequences as numbers of base pair differences. The genomic sequence of the CAS strain identified in the patients' uncultured sputum differed by 1128 bps from the genomic sequence of the cultured sputum sample in the CFU assay, which is the nearest CAS strain sequence within the patient's cluster. A variant distance of 877 bps was identified between the genomic sequence of the patients' cultured sample at MPN_RV1 from the genomic sequence of the nearest CAS strain sequence within the patient's cluster (59172 CFU). The non-clustering of these two samples is also attributable to the high sequence similarity with its neighbouring sequences in the tree, where the genomic sequence between the uncultured sample and the neighbouring sequence of 59152 Sputum were separated by a variant distance of 312 bp. High sequence similarity was observed between the cultured sample and its neighbouring sample (59144 MGIT_M), whose sequences were separated by a variant distance of 6 bp deletions. The close genetic relatedness between these samples may be a result of direct transmission of these strains during hospital visits or an unlikely event of sample cross-contamination during sequencing.

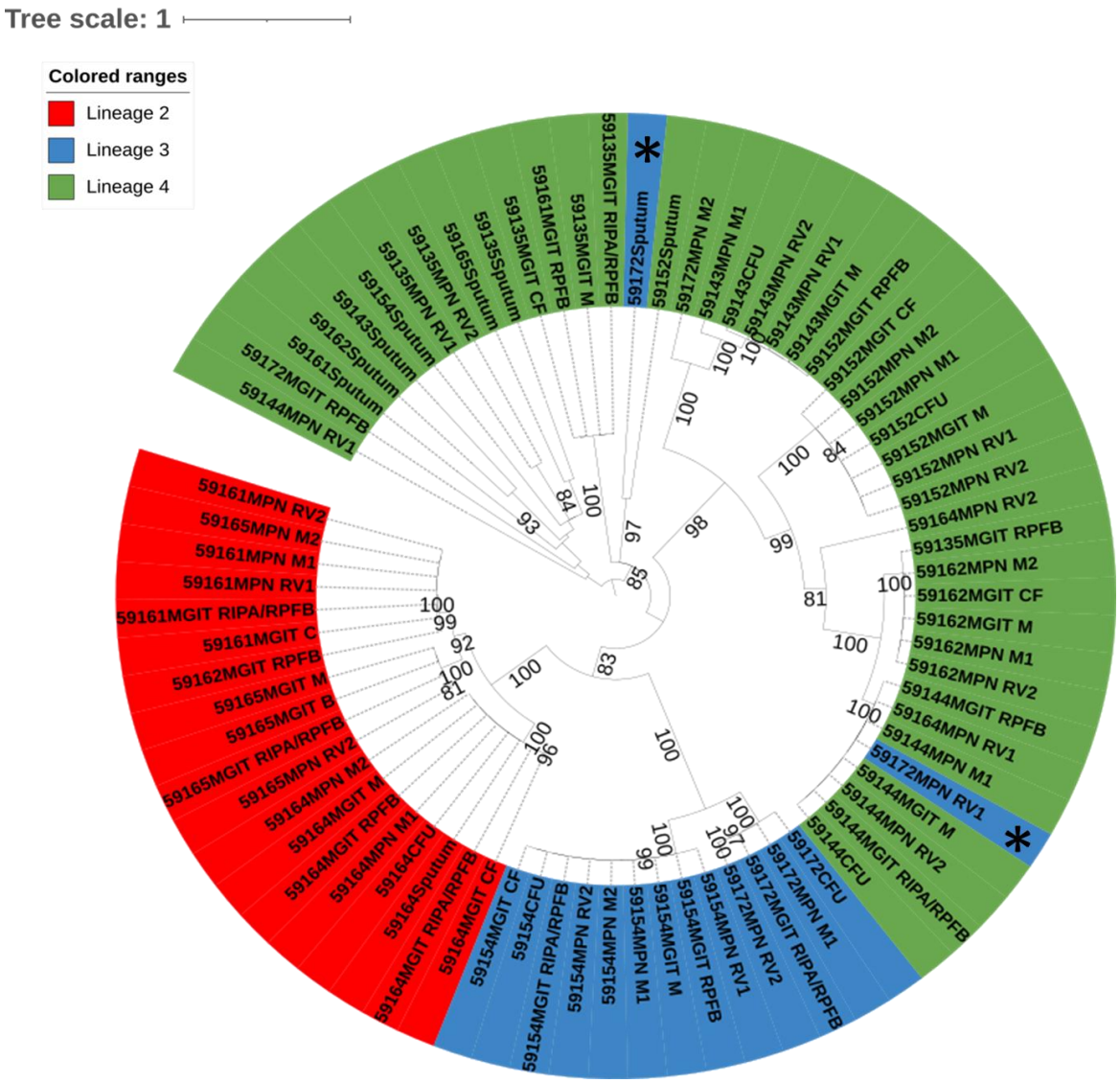


Figure 3.7: Maximum likelihood phylogenetic tree of directly sequenced and cultured *M. tb* clinical samples. Colour coding of lineages: Red- Lineage 2; Blue- Lineage 3; Green-Lineage 4. The tree is rooted with *M. tb H37Rv* as the recent common ancestor of all samples in the tree. Branch lengths indicate the rate of sequence divergence. Branch nodes are annotated midway with bootstrap values (75-100) which infers the statistical confidence of the positioning of each clade in the tree. Asterisk indicates samples that did not cluster according to their respective lineages. Scale bar represents the number of nucleotide substitutions per site.

3.8. Impact of different culture conditions on within-sample genetic diversity

The impact of culturing sputum under different growth conditions, with or without growth factor supplementation, on the within-sample genetic diversity was studied by analysing pairwise genomic SNP differences. A comparison of the detected variants in each patient's samples grown under different culture conditions revealed significant variation among the genomes (Figure 3.8). A genetic distance of 93 bp was observed between patient 59143's sputum sample cultured at MGIT_M and the cultured sputum sample at MPN_M1 (Figure

3.8 A), whilst a distance of 214 bps was observed between the sputum cultured at MGIT_M and MPN_M2 for patient 59165. (Figure 3.8 B). Patient 59152 had a 385 bp distance between sputum cultured at MGIT_RPFB and MGIT_CF (Figure 3.8 C). Large genetic distances were observed between the remaining *M. tb* genomes, with patient 59144 having a distance of 421 bps between MGIT_RIPA/RPFB and MPN_RV1 cultured sputum (Figure 3.8 D), patient 59154 having a distance of 620 bps between MPN_RV1 and MPN_M1 cultured sputum (Figure 3.9 E), and patient 59135 having a distance of 689 bps between MGIT_RPFB and MPN_RV1 cultured sputum (Figure 3.8 F). For patient 59162, a distance of 901 bp was observed between sputum cultured at MGIT_RPFB and MGIT_C (Figure 3.8 G), and a distance of 944 bp was observed between sputum cultured at MPN_RV1 and at CFU for patient 59164 (Figure 3.8 H). A similar increase in genetic diversity was seen in patient 59172, who had a genetic distance of 992 bp between sputum cultured at MGIT_RIPA/RPFB and MPN_M2 (Figure 3.8 I), and patient 59161, who had a genetic distance of 1003 bp between sputum cultured at MGIT_RPFB and MPN_RV1 (Figure 3.8 J). These data suggests that sputum culture has varying effects on the genetic diversity which became apparent under different growth culture conditions.

Interestingly, when stratifying genetic diversity by the number of nucleotide polymorphisms detected under the various culture conditions in each patient (Figure 3.6), some clustering was observed from those infected with multiple strains (59135, 59144, 59152, 59154, 59161, 59162, 59164, 59165, 59172) that was not observed in the single strain infected patient (59143). The different culture conditions (in each patient) performed differently depending on the infecting strain. Clustering of the culture conditions (in respect to the number of variants that were detected at that growth condition) was higher in patients polyclonally infected with either the Beijing or CAS strains (59154, 59161, 59162, 59164, 59165, 59172) but was substantially lower in patients infected with a single strain (59143) or were polyclonally infected with strains other than the above-mentioned strains (59135, 59144, 59152). These data suggest that culture selects for a specific genotype in sputum.

A)

	59143 CFU	59143MPN _M1	59143MPN _RV1	59143MPN _RV2	59143MGIT _M
59143CFU	0				
59143MPN_M1	91	0			
59143MPN_RV1	6	92	0		
59143MPN_RV2	6	91	4	0	
59143MGIT_M	4	93	2	6	0

B)

	59165MPN _M2	59165MPN _RV2	59165MGIT _M	59165MGIT _B	59165MGIT_RIPA/ RPFBA
59165MPN_M2	0				
59165MPN_RV2	207	0			
59165MGIT_M	214	12	0		
59165MGIT_B	204	15	26	0	
59165MGIT_RIPA/RPFBA	212	5	15	20	0

C)

	59152 CFU	59152MPN _M1	59152MPN _M2	59152MPN _RV1	59152MPN _RV2	59152MGIT _M	59152MGIT _CF	59152MGIT _RPFB
59152CFU	0							
59152MPN_M1	4	0						
59152MPN_M2	4	2	0					
59152MPN_RV1	3	5	5	0				
59152MPN_RV2	3	7	7	6	0			
59152MGIT_M	4	8	8	7	6	0		
59152MGIT_CF	5	9	9	8	7	4	0	
59152MGIT_RPFB	384	383	382	384	382	384	385	0

D)

	59144 CFU	59144MPN _M1	59144MPN _RV1	59144MPN _RV2	59144MGIT _M	59144MGIT _RPFB	59144MGIT _RIPA/RPFB
59144CFU	0						
59144MPN_M1	10	0					
59144MPN_RV1	420	419	0				
59144MPN_RV2	7	13	420	0			
59144MGIT_M	4	10	419	5	0		
59144MGIT_RPFB	11	5	418	16	13	0	
59144MGIT_RIPA/RPFB	4	10	421	5	4	13	0

E)

	59154 CFU	59154MPN _M1	59154MPN _M2	59154MPN _RV1	59154MPN _RV2	59154MGIT _M	59154MGIT _CF	59154MGIT _RPFB	59154MGIT_RIPA /RPFB
59154CFU	0								
59154MPN_M1	12	0							
59154MPN_M2	6	8	0						
59154MPN_RV1	615	620	612	0					
59154MPN_RV2	10	12	6	618	0				
59154MGIT_M	10	6	6	618	8	0			
59154MGIT_CF	6	10	2	614	8	8	0		
59154MGIT_RPFB	10	7	7	617	11	5	9	0	
59154MGIT_RIPA/RPFB	7	9	3	615	7	7	5	8	0

F)

	59135MPN_RV1	59135MPN_RV2	59135MGIT_M	59135MGIT_CF	59135MGIT_RPFB	59135MGIT_RIPA/RPFB
59135MPN_RV1	0					
59135MPN_RV2	10	0				
59135MGIT_M	495	491	0			
59135MGIT_CF	504	500	16	0		
59135MGIT_RPFB	689	686	511	511	0	
59135MGIT_RIPA/RPFB	496	492	19	17	504	0

G)

	59162MPN_M1	59162MPN_M2	59162MPN_RV2	59162MGIT_M	59162MGIT_CF	59162MGIT_RPFB
59162MPN_M1	0					
59162MPN_M2	7	0				
59162MPN_RV2	4	7	0			
59162MGIT_M	4	5	2	0		
59162MGIT_CF	8	7	6	4	0	
59162MGIT_RPFB	899	898	898	898	901	0

H)

	59164CFU	59164MPN_M1	59164MPN_M2	59164MPN_RV1	59164MPN_RV2	59164MGIT_M	59164MGIT_CF	59164MGIT_RPFB	59164MGIT_RIPA/RPFB
59164CFU	0								
59164MPN_M1	21	0							
59164MPN_M2	25	26	0						
59164MPN_RV1	944	936	930	0					
59164MPN_RV2	942	933	928	325	0				
59164MGIT_M	21	16	27	937	932	0			
59164MGIT_CF	21	22	33	943	942	25	0		
59164MGIT_RPFB	26	29	16	934	932	31	35	0	
59164MGIT_RIPA/RPFB	20	15	30	942	940	19	11	32	0

I)

	59172CFU	59172MPN_M1	59172MPN_M2	59172MPN_RV1	59172MPN_RV2	59172MGIT_RPFB	59172MGIT_RIPA/RPFB
59172CFU	0						
59172MPN_M1	521	0					
59172MPN_M2	989	543	0				
59172MPN_RV1	877	455	399	0			
59172MPN_RV2	18	518	987	876	0		
59172MGIT_RPFB	932	535	577	532	932	0	
59172MGIT_RIPA/RPFB	20	521	992	880	11	933	0

J)

	59161MPN_M1	59161MPN_RV1	59161MPN_RV2	59161MGIT_C	59161MGIT_RPFB	59161MGIT_RIPA/RPFB
59161MPN_M1	0					
59161MPN_RV1	11	0				
59161MPN_RV2	9	16	0			
59161MGIT_C	6	15	9	0		
59161MGIT_RPFB	994	1003	996	996	0	
59161MGIT_RIPA/RPFB	9	13	14	11	998	0

Figure 3.8: Pairwise SNP differences between the genomes of cultured *M. tb* sputum samples. Within sample genetic distance comparison for patients **A)** 59143, **B)** 59165, **C)** 59152, **D)** 59144, **E)** 59154, **F)** 59135, **G)** 59162, **H)** 59164, **I)** 59172, **J)**59161 across the different growth conditions. Genomes were compared individually under the different growth conditions. Genomes with the greatest genetic distance are reported. Samples highlighted in yellow represent the intersection of the same genome. Samples shaded in blue represent the samples with the largest genetic distance which was read from the bottom left to the upper matrix (across).

3.9. Effect of culture on the population structure in single strain infections

Since the effect of culture on the population structure had been determined, pairwise SNP distance comparisons between the genomes of the same infecting strain type were analysed to determine the effect of culture on the clonal composition of single strain infections in sputum. Since 90% of the patients were polyclonally infected, this analysis was carried out using only the genomes of the same strain type across the various growth conditions from each patient. The smallest variant distance between the genomes of the same infecting strain were observed in the LAM 3 strains of patient 59162 (8 bp) (Figure 3.9 A), the LAM 9 strains infecting patient 59152 (9 bp) (Figure 3.9 B), and the T1 strains infecting patient 59135 (10 bp) (Figure 3.9 C). This was followed by patient 59161 infected with the Beijing strains (16 bp) (Figure 3.9 D), patient 59135 infected with the T2 strains (19 bp) (Figure 3.9 E), patient 59164 infected with the Beijing strains (35 bp) (Figure 3.9 F), patient 59143 infected with the S strains (93 bp) (Figure 3.9 G). Large differences were observed in the Beijing strains infecting patient 59165 (214 bp) (Figure 3.9 H) and in the LAM 3 strains infecting patient 59144 (421 bp) (Figure 3.9 I). The CAS strains infecting patients 59154 and 59172 were the most heterogeneous, with up to 620 bp and 877 bp differences, respectively (Figure 3.9 I & J). The detection of these genotypes in the same strain were only possible to observe as a result of culturing the sputum in the CFU assay (9%) (1/11), MPN_M1(27%) (3/11), MPN_RV1 (18%) (2/11), MPN_RV2(9%), MGIT_M (9%), MGIT_CF (18%), and MGIT_AB (9%) conditions. These results suggest the presence of minority bacterial populations in sputum that emerge in culture.

A)

	59162MPN_M1	59162MPN_M2	59162MPN_RV2	59162MGIT_M	59162MGIT_CF
59162MPN_M1	0				
59162MPN_M2	7	0			
59162MPN_RV2	4	7	0		
59162MGIT_M	4	5	2	0	
59162MGIT_CF	8	7	6	4	0

B)

	59152CFU	59152MPN_M1	59152MPN_M2	59152MPN_RV1	59152MPN_RV2	59152MGIT_M	59152MGIT_CF
59152CFU	0						
59152MPN_M1	4	0					
59152MPN_M2	4	2	0				
59152MPN_RV1	3	5	5	0			
59152MPN_RV2	3	7	7	6	0		
59152MGIT_M	4	8	8	7	6	0	
59152MGIT_CF	5	9	9	8	7	4	0

C)

	59135MPN_RV1	59135MPN_RV2
59135MPN_RV1	0	
59135MPN_RV2	10	0

D)

	59161MPN_M1	59161MPN_RV1	59161MPN_RV2	59161MGIT_C	59161MGIT_RIPA/RPFB
59161MPN_M1	0				
59161MPN_RV1	11	0			
59161MPN_RV2	9	16	0		
59161MGIT_C	6	15	9	0	
59161MGIT_RIPA/RPFB	9	13	14	11	0

E)

	59135MGIT_M	59135MGIT_CF	59135MGIT_RIPA/RPFB
59135MGIT_M	0		
59135MGIT_CF	16	0	
59135MGIT_RIPA/RPFB	19	17	0

F)

	59164CFU	59164MPN_M1	59164MPN_M2	59164MGIT_M	59164MGIT_CF	59164MGIT_RPFB	59164MGIT_RIPA/RPFB
59164CFU	0						
59164MPN_M1	21	0					
59164MPN_M2	25	26	0				
59164MGIT_M	21	16	27	0			
59164MGIT_CF	21	22	33	25	0		
59164MGIT_RPFB	26	29	16	31	35	0	
59164MGIT_RIPA/RPFB	20	15	30	19	11	32	0

G)

	59143CFU	59143MPN_M1	59143MPN_RV1	59143MPN_RV2	59143MGIT_M
59143CFU	0				
59143MPN_M1	91	0			
59143MPN_RV1	6	92	0		
59143MPN_RV2	6	91	4	0	
59143MGIT_M	4	93	2	6	0

H)

	59165MPN_M2	59165MPN_RV2	59165MGIT_M	59165MGIT_B	59165MGIT_RIPA/RPFB
59165MPN_M2	0				
59165MPN_RV2	207	0			
59165MGIT_M	214	12	0		
59165MGIT_B	204	15	26	0	
59165MGIT_RIPA/RPFB	210	3	13	18	0

I)

	59144CFU	59144MPN_M1	59144MPN_RV1	59144MPN_RV2	59144MGIT_M	59144MGIT_RPFB	59144MGIT_RIPA/RPFB
59144CFU	0						
59144MPN_M1	10	0					
59144MPN_RV1	420	419	0				
59144MPN_RV2	7	13	420	0			
59144MGIT_M	4	10	419	5	0		
59144MGIT_RPFB	11	5	418	16	13	0	
59144MGIT_RIPA/RPFB	4	10	421	5	4	13	0

J)

	59154CFU	59154MPN_M1	59154MPN_M2	59154MPN_RV1	59154MPN_RV2	59154MGIT_M	59154MGIT_CF	59154MGIT_RPFB	59154MGIT_RIPA/RPFB
59154CFU	0								
59154MPN_M1	12	0							
59154MPN_M2	6	8	0						
59154MPN_RV1	615	620	612	0					
59154MPN_RV2	10	12	6	618	0				
59154MGIT_M	10	6	6	618	8	0			
59154MGIT_CF	6	10	2	614	8	8	0		
59154MGIT_RPFB	10	7	7	617	11	5	9	0	
59154MGIT_RIPA/RPFB	7	9	3	615	7	7	5	8	0

K)

	59172CFU	59172MPN_M1	59172MPN_RV1	59172MPN_RV2	59172MGIT_RIPA/RPFB
59172CFU	0				
59172MPN_M1	521	0			
59172MPN_RV1	877	455	0		
59172MPN_RV2	18	518	876	0	
59172MGIT_RIPA/RPFB	20	521	880	11	0

Figure 3.9: Pairwise SNP differences between the genomes of the same strain genotype in cultured *M. tb* sputum. Within sample genetic distance comparison for patients **A)** 59162 (LAM 3 strains), **B)** 59152 (LAM 9 strains), **C)** 59135 (T1 strains), **D)** 59161 (Beijing strains), **E)** 59135 (T2 strains), **F)** 59164 (Beijing strains) **G)** 59143 (S strains), **H)** 59165 (Beijing strains), **I)** 59144 (LAM 3 strains), **J)** 59154 (CAS strains) **K)** 59172 (CAS strains) across the different growth conditions. Samples are arranged by ascending order in respect to the genetic distances between the genomes. Samples shaded in green represent the intersection of the same genome. Samples shaded in grey represent the samples with the largest genetic distance which was read from the bottom left to the upper matrix (across).

3.10. Investigation and comparative analysis of *M. tb* drug target and housekeeping genes

As the required genomic sequence depth for most (90%) of the raw sputum samples was not obtained, the intended objective, which was to conduct a comparative analysis of this sequence data to the sequence data obtained from samples that were cultured under different growth conditions, could not be met. As a result, focus was shifted towards observing genetic polymorphisms in known drug targets of the four first-line drugs, namely Isoniazid (*INH*) (Rv1484), Ethambutol (*EMBA*, *EMBB*, *EMBC*) (Rv3794, Rv3795, Rv3793), Rifampicin (*RIF*)(Rv0667) and Pyrazinamide (*PZA*) (Rv2043c), as well as in genes that regulate the expression of housekeeping genes, namely *sigA* (Rv2703) and *sigB* (Rv2710). Gene annotations were obtained from Mycobrowser (https://mycobrowser.epfl.ch/releases). The *M. tb sigA* and *sigB* genes code for two sigma factors of RNA polymerase which are vital for the initiation of transcription. Genomic coverage of the respective genes from the sputum derived sequences was manually assessed prior to the analysis. Results are summarised in Table 3.4. One sputum sample had sufficient coverage for all genes, whereas the other samples had partial gene coverage which still allowed for analysis. The sputum sample of patient 59144 was excluded from further analysis since it was suspected of having a high number of false positive variants.

Table 3.4: Coverage of the four first-line *M. tb* drug-targets and regulatory genes from each patient's sputum derived sequence.

	INH	***EMBA***	***EMBB***	***EMBC***	***PZA***	***RIF***	***SigA***	***SigB***
59135		×	×	×	×			
59143		×	×	×	×			×
59144					×			
59152						×		
59154	×	×		×	×			
59161	×		×	×				
59162	×	×		×	×			
59164								
59165	×	×			×	×		
59172	×	×						

Samples that had partial gene coverage are shaded in grey and those that had full gene coverage are shaded in blue. ×= Insufficient/No gene coverage for analysis.

No drug resistance conferring mutations were observed from the sputum derived and cultured sequences. The data confirmed that all strains used in this study were drug susceptible to the first four-line anti TB drugs and did not acquire mutations during culture under the various growth conditions. However, missense mutations in the *sigA* gene were found in the sputum derived sequences of patients 59135, 59154, 59161, 59162, and 59165 (Table 3.5).

Table 3.5: Mutations in the *sigA* gene and the resulting change in amino acid

Gene	Patient ID	Mutation	Effect
sigA (Rv2703)	59135 Sputum	946 G » C 947 C » A	Ala 316 Pro Ala 316 Glu
	59154 Sputum	946 G » C 947 C » T	Ala 316 Pro Ala 316 Val
	59161 Sputum	946 G » C 947 C » T 1480 C » A	Ala 316 Pro Ala 316 Val Gln 494 Lys
	59162 Sputum	946 G » C 947 C » T 1022 A » T	Ala 316 Pro Ala 316 Val Tyr 341 Phe
	59165 Sputum	946 G » C 947 C » T 1139 G » C 1144 A » G 1153 G » C 1156 C » A 1480 C » A	Ala 316 Pro Ala 316 Val Gly 380 Ala Ile 382 Val Glu 385 Gln Leu 386 Met Gln 494 Lys

The number of missense variants identified in the *M. tb sigA* gene varied between the patients, with two double mutants (59135, 59154), two triple mutants (59161, 59162), and one patient with a total of seven missense variants (59165) identified in uncultured sputum. The G » C mutation at position 3018 780 (946) was common in all uncultured sputum samples, followed by a C » T mutation at position 3018781 (947) in most of the patients uncultured sputum sample. These mutations were manually inspected from the alignment file

using Integrative genome browser (Appendix C, Figure 3). None of these mutations were detected in any of the cultured sequences. Thirteen amino acid sequences (sigA-M) have been annotated as putative sigma factors which all belong to the sigma-70 family (Campbell *et al.*, 2002). The mutations in the sigA gene fall within the sigma70-region 2 (pfam04542) domain (uniProtKB accession number: P9WGI1). There were no mutations in the *sigB* gene.

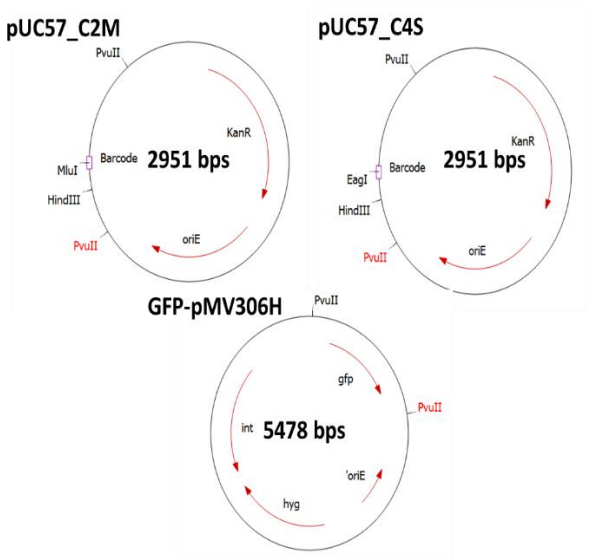
3.11. Artificial mixed strain *M. tuberculosis* infection modelling

Infections with mixed strains of *M. tb* result in poor treatment outcomes (Shin *et al.,* 2018). A modelling framework for these infections provides insight into the mechanism of strain competition during mixed strain *M. tb* infections, allowing for a better understanding of disease dynamics at the individual level.

3.11.1. Confirmation of the pUC57-C2M and pUC57_C4S vector by restriction digest analysis

Plasmids pUC57_C2M and pUC57_C4S were used for the generation of the barcoded LAM and Beijing strains (Figure 3.10 A). A positive control plasmid (pMV306H) was used to assess the transformation efficiency. The plasmid DNA was extracted and purified from the *Escherichia coli* strain using the NucleoBond Xtra Midi/Maxi kit. Before cloning the respective barcodes, the plasmids were profiled with various restriction enzymes to ensure that they had not undergone any genetic rearrangement (Figure 3.10 B). The PvuII and HindIII enzymes in each case digest outside of the barcode containing multiple cloning site (MCS). Identical band sizes were obtained from this digestion since both plasmids differed only in the multiple cloning site where their respective barcodes were inserted (Figure 3.10 C). The third enzyme was unique to each barcode and digested once within the MCS to differentiate the plasmids from each other. All enzymes produced the correct fragments for both pUC57-C2M and pUC57_C4S plasmids as well as the positive control plasmid (Figure 3.9 C). In addition, each plasmid was digested with the respective enzyme that gave the unique specific bands (indicated with a red asterisk in Figure 3.9 D) to confirm that the pUC57_C2M and pUC57_C4S plasmids are different from each other. Both plasmids were uncut from this digestion as seen by the presence of a supercoiled, linear, and nicked band which confirmed that the experimental plasmids were different from each other. As both plasmids were confirmed to be correct, generation of the barcode labelled *M. tb* strains was pursued.

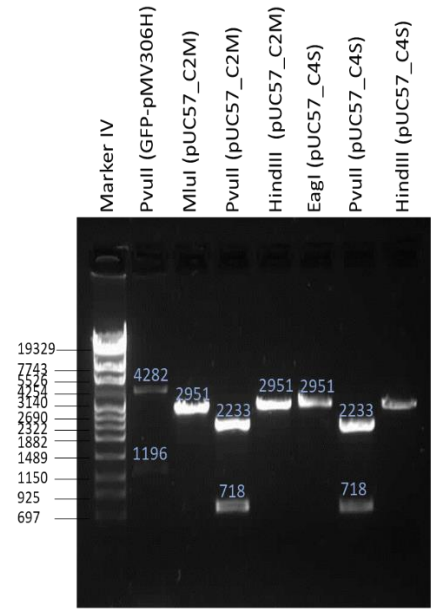
A)



B)

Enzymes	Plasmid	Expected band size
PvuII	GFP-pMV306H	4282, 1196
MluI	pUC57_C2M	2951
PvuII	pUC57_C2M	2233, 718
HindIII	pUC57_C2M	2951
EagI	pUC57_C4S	2951
PvuII	pUC57_C4S	2233, 718
HindIII	pUC57_C4S	2951

C)



D)

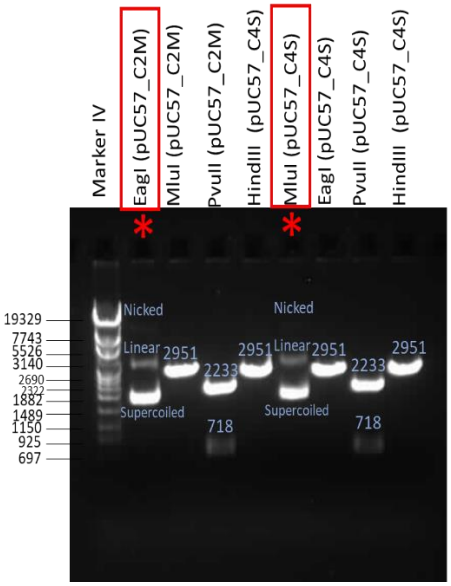


Figure 3.10: Restriction profile of the pUC57_C2M, pUC57_C4S and GFP-pMV306H plasmids. **A)** The vector maps for all three plasmids were generated using the clone manager software (Scientific and Educational Software, USA). Plasmids containing the respective barcodes (pUC57_C2M and pUC57_C4S) were digested with 3 enzymes and the control plasmid (pMV306H) was digested with a single enzyme. **B)** Expected band sizes for each of the enzymes used to digest the plasmid DNA. **C)** 1% agarose gel showing the expected DNA band sizes, lane 1(Marker IV), lane 2 (PvuII digest for the GFP-pMV306H plasmid), lane 3 (MluI digest for the pUC57_C2M plasmid), lane 4 (PvuII digest for the pUC57_C2M plasmid), lane 5 (HindIII digest for the pUC57_C2M plasmid), lane 6 (EagI digest for the pUC57_C4S plasmid), lane 7 (PvuII digest for the pUC57_C4S plasmid), lane 8 (HindIII digest for the pUC57_C4S plasmid). **D)** 1% agarose gel indicating the DNA band sizes obtained for confirmation of the two plasmids where the asterisks indicate the enzymes which were not specific for the selected plasmid. Lane 1(Marker IV), lane 2 (EagI digest for the pUC57_C2M plasmid), lane 3 (MluI digest for the pUC57_C2M plasmid), lane 4 (PvuII digest for the pUC57_C2M plasmid), lane 5 (HindIII digest for the pUC57_C2M plasmid), lane 6 (MluI digest for the pUC57_C4S plasmid), lane 7 (EagI digest for the pUC57_C4S plasmid), lane 8 (PvuII digest for the pUC57_C4S plasmid), lane 9 (HindII digest for the pUC57_C4S plasmid).

3.11.2. Generation of barcoded *M. tuberculosis* clinical strains using electroporation

The genomic integration of the DNA barcodes into the LAM and Beijing clinical strains was performed using electroporation to allow for strain identification within the mixture. Two LAM strains from samples 59144_CFU and 59152_CFU, and two Beijing strains from sample 59161 MPN_M1 and 59164 CFU, were selected from the spoligotyping data (section 3.3) and used for this experiment. One µg/ml of the plasmid DNA was transformed into electrocompetent LAM and Beijing strains respectively. The pUC57_C2M plasmid was used for the LAM strains and the pUC57_C4S plasmid was used for the Beijing strains. To calculate the transformation efficiency, electrocompetent LAM and Beijing strains were transformed with the pMV306H plasmid and as a control, electrocompetent *M. tb* H37Rv cells were electroporated with both the pUC57_C2M and pUC57_C4S plasmids individually. After recovery and phenotypic expression, different dilutions (from 10^{-0} to 10^{-3}) of the cells were spread onto Middlebrook 7H11 agar supplemented with 50 µg/ml kanamycin for the transformed pUC57-C2M and pUC57_C4S plasmids, and 50 µg/ml hygromycin for the pMV306H control plasmid to obtain the transformed cells. The plates from the electroporation are shown in Figure 3.11.

Bacterial growth was observed on the plates for the H37Rv strain, as well as the LAM and Beijing clinical strains (Figure 3.11 A-F). Bacterial growth was minimal for the clinical strains, as no growth was observed from the 10^3-dilution. The transformation efficiencies ranged from $1.25x10^3$ - $4.58x10^4$ transformants/ug DNA (Table 3.6). One of the clinical samples transformed with the pMV306H (control) plasmid had the highest transformation efficiency, while all clinical samples transformed with the pUC57 plasmids had the lowest transformation efficiency. The transformation efficiency of the second sample that was transformed with the positive control plasmid (highlighted in grey in Table 3.6) is not accurate as the image of the plate that is represented was taken earlier (after approximately 2 weeks of growth) due to contamination while, all the other plates were grown for 4 weeks. Although the transformation efficiency of the H37Rv strains was higher than that of the clinical strains, it was still not in the expected range of 10^5-10^6 transformants/ug of DNA. The low number of transformants for the clinical strains and the low transformation efficiency for these samples suggest that the genomic integration of the plasmids may have been unsuccessful and that the colonies that are seen on the plates are spontaneous mutants that developed resistance to kanamycin. To confirm the genomic integration, the colonies

were screened by PCR using primers listed in Table 2.4 that flank the barcode regions to confirm if the vector was successfully integrated into the *M. tb* chromosome.

Table 3.6: Transformation efficiencies for the pUC57_C2M, pUC57_C4S and the pMv306H plasmids in various *M. tb* strains

Sample	**Transformation efficiency**
H37Rv-pUC57_C2M	$2.43x10^4$
H37Rv_ pUC57_C4S	$2.69x10^4$
59144_ pUC57_C2M	$2.0x10^3$
59144_ pUC57_pMV306H	$4.58x10^4$
59152_ pUC57_C2M	$2.1x10^3$
59161_ pUC57_C4S	$1.58x10^3$
59161_ pUC57_pMV306H	$9.32x10^2$
59164_ pUC57_C4S	$1.25x10^3$

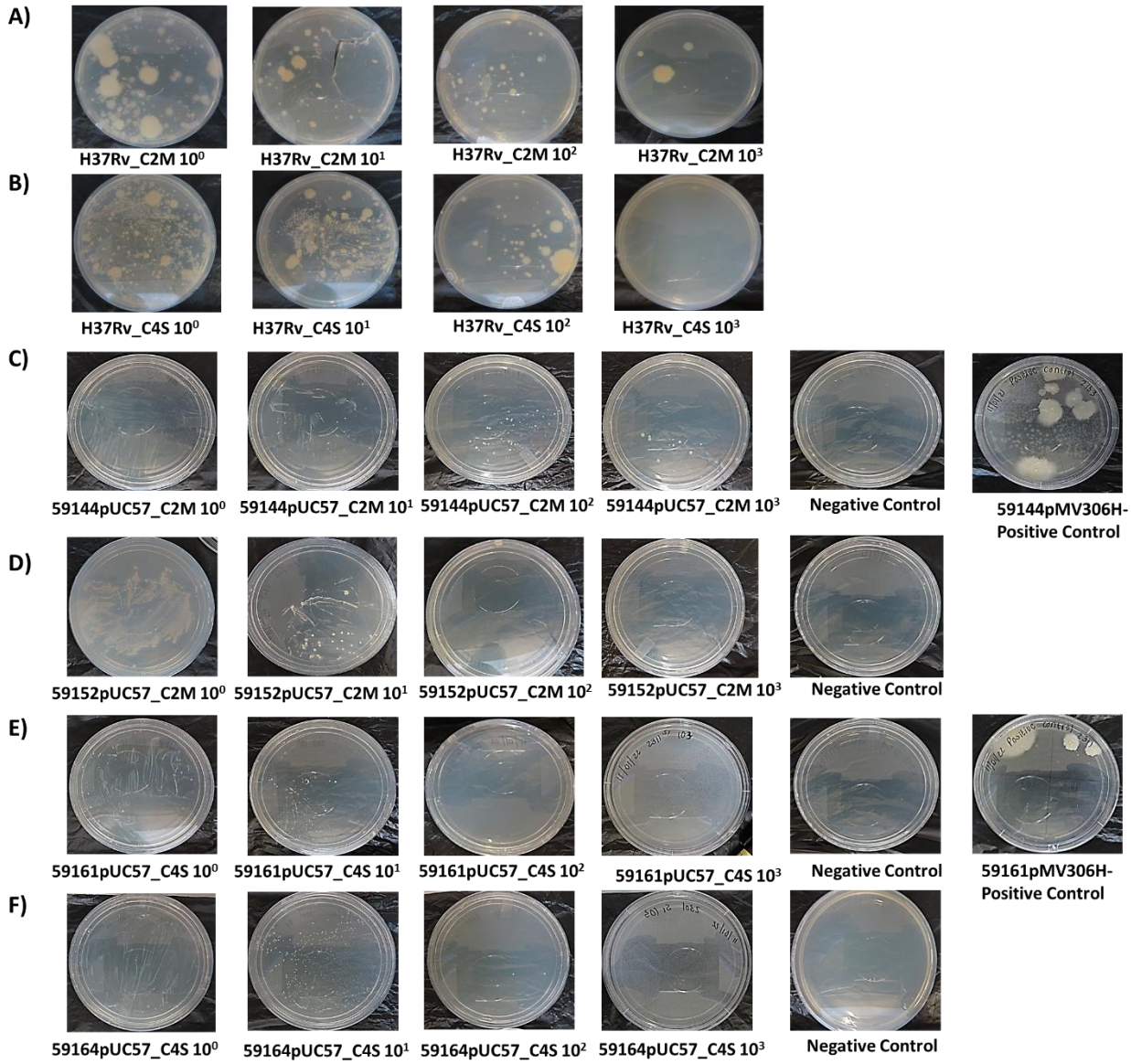


Figure 3.11: Plates showing the generation of the barcoded *M. tb* strains carried out by a site-specific recombination event. Electrocompetent *M. tb* H37Rv, LAM and Beijing cells were transformed with the pMV306H, pUC57_C2M and pUC57_C4S plasmids and were selected on 7H11 pla tes supplemented with 50 µg/ ml kanamycin for the pUC57 plasmids and 50 µg/ ml hygromycin for the pMV306H plasmids. **A)** Electrocompetent H37Rv cells were transformed with the pUC57_C2M plasmid. **B)** Electrocompetent H37Rv cells were transformed with the pUC57_C4S plasmid. **C)** Electrocompetent LAM strain cells from patient 59144 were transformed with the pUC57_C2M plasmid. Negative control contains cells only, positive cells contain the pMV306H plasmid. **D)** Electrocompetent LAM strain cells from patient 59152 were transformed with the pUC57_C2M plasmid. Negative control contains cells only. **E)** Electrocompetent Beijing strain cells from patient 59161 were transformed with the pUC57_C4S plasmid. Negative control contains cells only, positive cells contain the pMV306H plasmid. **F)** Electrocompetent Beijing strain cells from patient 59164 were transformed with the pUC57_C4S plasmid. Negative control contains cells only.

3.11.3. Confirmation of newly generated barcoded *M. tb* strains by PCR screening

Colonies from the electroporation plates were screened using PCR to verify the presence of the genetic barcode in the *M. tb* chromosomal genome (Figure 3.12). A non-template control was included to rule out contamination. The absence of DNA bands on the non-template controls confirms the absence of contaminants in the reaction mixture (Figure 3.12). No DNA bands were obtained from any of the screened colonies which indicates that the generation of the barcoded strains was unsuccessful. The failed integration of the DNA barcode could be due to various reasons. Cell damage may have occurred during pulsing, causing the pores of the bacteria to remain open, or the plasmid concentration may have been insufficient in comparison to the bacterial cell concentration. However, due to time constraints, the experiment could not be optimised.

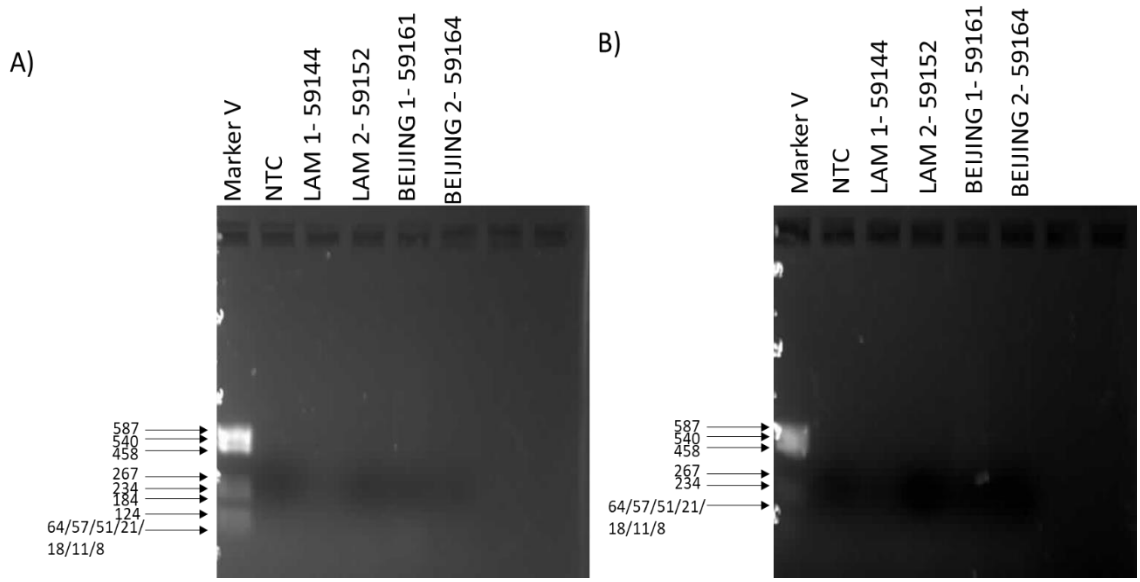


Figure 3.11: PCR screening of the generated barcoded *M. tb* strains. The gels display the samples that were run at an annealing temperature of **A)** 64°C and **B)** 62°C. Three µL of each sample was electrophoresed on a 4% agarose gel at 90V for 60 minutes. Lane 1 (Marker V), lane 2 (LAM strain of patient 59144), lane 3 (LAM strain of patient 59152), lane 4 (Beijing strain of patient 59161), lane 5 (Beijing strain of patient 59162).

4. DISCUSSION

Despite the availability of a reliable and effective drug regimen, as well as accurate diagnostic tests, tuberculosis remains a global epidemic that affects nearly one-third of the world's population (Getahun *et al.,* 2015). The therapeutic management of this disease is challenged by the presence of heterogeneous bacterial populations within the host that result either due to the microevolution of a single strain or by polyclonal infection with distinct *M. tb* strains (Liu *et al.,* 2015). Conventional culture methods for diagnosis often include the inoculation of the sputum sample into various types of growth media to allow for genotypic or phenotypic analysis. Since bacteria have different growth requirements, culturing sputum derived *M. tb* bacteria is limited by the selective nature of artificial media, which favours the enumeration of the bacterial population or strain type that is best suited for growth in the selected media (Dusthackeer *et al.,* 2019). Growth factor supplementation in liquid media has been shown to improve the recovery to around 80 – 99.9% of viable *M. tuberculosis* in sputa that was undetected by a standard agar-based assay (Mukamolova *et al., 2*010). The increased bacterial yield during culture under favourable conditions is a possible source of unexplored genomic diversity which may provide insight into understanding the influence of different growth conditions on within-patient mycobacterial genetic diversity. In this study, we aimed to determine whether different axenic media with or without growth factor supplementation impacted within-sample genetic heterogeneity of the mycobacteria in the sputum samples. In addition, we also aimed to understand the mechanism of strain behaviour in polyclonal infections.

Spoligotyping has a limited sensitivity in the detection of mixed strains in sputum

In this study, spoligotyping was performed on the ten uncultured sputum samples and their respective cultured sputum samples grown in the CFU, MPN and MGIT assays, with or without growth factor supplementation, for the classification of strain types. Spoligotype patterns were obtained from all uncultured sputum samples (100%), which is higher than the 98.5% obtained in a previous study that aimed to characterise the potential bias introduced by sputum culture on the prevalence of MTBC lineages by comparing spoligotype patterns obtained from direct sputum to those obtained from sputum culture (Sanoussi *et al., 2*017). Differences may be attributed to different sputum decontamination and DNA extraction methods. Using SPOTCLUST, 50% (5/10) of the uncultured sputum samples (59135, 59154, 59161, 59162, 59165) and 4% (3/83) of the cultured sputum samples (59144 MPN_M2, 59165 MPN_M1, 59165 MPN_RV1) with evidence of mixed spoligotype patterns by the

presence of overlapping spacers, were classified as Family 33 strains. Family 33 strains were characterised by the absence of spacers 33-34 in the spoligotype pattern (Vitol *et al., 2*006). The Family 33 strains in this study displayed spoligotype patterns that differ from this description by Vitol et. al (2006) and also displayed spoligotype patterns that differ from each other. In addition to the 33rd and 34th spacers, all Family 33 strains in this study lacked spacers in other locations within the DR locus. The resulting spoligotype patterns from these were an overlap of the spacers from the infecting strains, which in all cases, lacked spacers 33-34. As a result, these mixed patterns were classified as belonging to the Family 33 sub lineage because they met the criteria for this classification due to the absence of spacers 33-34, but they are not a true Family 33 strain. Due to the overlap of the hybridization patterns from multiple strains in a single sample which appear as a single pattern, spoligotyping can be said to have limited sensitivity in the detection of mixed strain infections in sputum. Similar results were observed in a study by Kamakoli *et al.*, (2018). Spoligotyping has been used with additional typing methods to improve the detection of mixed strain infections in sputum (Huyen *et al.*, 2012, Pandey *et al.*, 2020). Although direct sputum spoligotyping is possible, it is advantageous to perform it on cultured sputum samples because it allows the analysis to be carried out on multiple single colonies in the event that they contain different strains and because cultured sputum samples allow the analysis to be carried out on pure DNA without the presence of potential contaminants that could interfere with the hybridisation pattern.

WGS provides a greater resolution than spoligotyping in the detection of additional strains in sputum and in lineage assignment

Whole genome sequencing was used to confirm the infecting strain types that were obtained by the experimental spoligotyping. *In silico* predicted spoligotypes were in concordance with 83% of the experimentally derived spoligotypes. WGS detected the presence of a Beijing strain in patient 59144's uncultured sputum which was not visible on the experimental spoligotype pattern. The presence of spacers 35 and 36, which are characteristic of Lineage 2 strains, demonstrated the increased sensitivity of WGS in detecting the additional strain type in sputum as the Beijing strain emerged after culture in the second well of the un-supplemented MPN assay (M2). Hence, a discrepancy was noted between the experimental and *in-silico* results for this patient. The classification of the patients' cultured sample as a Family 35 strain by WGS is not understood. WGS has previously been used to identify mixed infections with non-tuberculosis species in sputum (Advani *et al., 2*019, Khieu *et al., 2*021).

In this study, WGS was also able detect an *M. tb* co-infection with a mycobacterium other than tuberculosis (MOTT) (*M. microti*) in patient 59162 at the MGIT_RIPA/RPFB culture condition which was not detected by experimental spoligotyping. Infections caused by MOTT are frequently missed due to clinical similarities with *M. tb*, including the need for specialized growth media for detection. The MGIT assay promotes the growth of various mycobacteria (Donohue, 2018) and therefore, is useful in detecting co-infection with mycobacteria other than *M. tb*. Instead of a LAM 9 strain, WGS identified a LAM 7 strain in patient 59152's sputum. The reason for this discrepancy is not understood due to the limited information on LAM 7 strains. The discrepancy in the octal code of patient 59172's CAS strain highlights the accurate classification of strains to their respective sub lineages by WGS. Although the spoligotype patterns of patient 59154's CAS strain (deletion of spacers 4-7, 10, 20-34) and patient 59172's CAS strain (deletion of spacers 4-7, 23-34) are different, the spoligotype strain assignment tool (SPOTCLUST) (Vitol *et al.,* 2006) assigned the same octal code for both strains. WGS was able to distinguish the two strains from each other by means of a different octal code and by the clustering pattern of both strains on the phylogenetic tree. This was further confirmed by previous studies, where strains with the octal code of patient 59154 were assigned to the CAS1-kili sub-lineage and strains with the octal code of patient 59172 were assigned to the CAS1-delhi sub lineage (Bulane, 2012, Taye *et al., 2*021). These data suggests that WGS has a greater resolution than spoligotyping in detection of additional mycobacterial infecting strains in sputum and in lineage assignment.

WGS was further used to detect mixed infections in sputum and in culture using the heterozygous sites method by Sobkowiak *et al.* (2018). Heterozygous variants were identified in uncultured sputum and in cultured sputum samples that displayed mixed patterns during spoligotyping analysis. Due to the emergence of different strain spoligotype patterns in uncultured sputum and sputum cultured in the various growth conditions, the experimental spoligotyping data suggested mixed infections in 90% (9/10) of the patients. The heterozygous site method also identified mixed infections in 90% of the patients. However, there was a mismatch between the classification from the spoligotyping data and the classification from the heterozygous site method between two patients where one was classified as having a single infecting strain by spoligotyping but was classified as having mixed strains by the heterozygous site method (patient 59143) and another one who was classified as having a mixed strain infection by spoligotyping (patient 59164) but was classified as having a single strain by the heterozygous site method. The mismatch of patient

59143 may be due to the sample's low coverage depth, which could have missed genomic regions that would have correctly classified the sample. Patient 59143's uncultured sputum sample had the lowest depth of coverage (0.582x) of all the uncultured sputum samples. For patient 59164, the mismatch may be due to the detection of a single infecting strain in sputum (as evidenced by the sample's spoligotype pattern). However, the additional strain should still be detected in sputum even if it is present in a small proportion. These results further highlight the sensitivity of WGS in detecting additional strain types in sputum, as 8/9 patients identified as polyclonally infected by spoligotyping were classified as mixed by WGS using only the sequence data from the uncultured sputum sample whereas spoligotyping required spoligotype patterns from cultured sputum that differed from the spoligotype pattern of the uncultured sputum for the classification of mixed infections.

Direct sputum sequencing without target enrichment challenges the recovery of *M. tb* DNA in sputum

DNA isolated from the uncultured sputum of each patient and its respective cultured samples under the various growth conditions was sequenced. The stored uncultured sputum samples were purified using the N-acetyl-l-cysteine-Sodium hydroxide method, and the stored cultured bacterial samples were purified using an antibiotic mixture (PANTA) prior to WGS sequencing. Reference genome coverage depth (0.581x-204.53x) and breadth (0.38%-60.01%) were significantly lower in the uncultured sputum samples compared to the cultured sputum samples (129.734x - 696.525x and 95.51%-99.42%) and across the various growth conditions. The low coverage depth of the uncultured sputum samples in this study was greater than that obtained in a study by Doughty *et al.* (2014), which sought to investigate the potential of direct sputum sequencing without culture or DNA target-specific amplification to characterize smear-positive MTBC strains from West Africa. The coverage depth of the sputum-derived sequences in this study ranged from 0.002x to 0.7x. (Doughty *et al., 2*014). Using various hybridisation protocols, studies have shown that target enrichment of DNA derived from uncultured sputum using RNA and DNA capture baits significantly improves *M. tb* recovery in sputum. Nimmo *et al.* (2019) compared three different hybridisation protocols (SureSelect standard, SureSelect Low Input and MolYsis + SureSelect Input) for direct sputum sequencing and discovered that the MolYsis SureSelect Low Input kit improved reference genome coverage the most (68.42x-135.36x). This strategy may be employed in future research for the analysis of *M. tb* genomes in uncultured sputum samples.

Genetic diversity is both genotypic and culture dependent

To determine if sputum culture affects within-sample genetic diversity, variants were identified in the genomic sequences of the various *M. tb* samples in respect to the H37Rv reference strain and were compared to each other. Since mixed strain infections were identified in most of the patients, within-patient genetic diversity was expected due to the different genotypes of the infecting strains. The different strain types in the polyclonally infected patients emerged during sputum culture across the various growth conditions, as demonstrated by the spoligotyping data and the pairwise genetic distances between the genomic sequences of the cultured samples, which is of consequence in the diagnosis of patients with multiple strain infections in diagnostic settings that rely on only one diagnostic technique or those that are not suitable to promote the growth of the additional strain types in sputum. The type of effect that culture has on the sample genetic diversity is demonstrated by the results that were obtained from the identification of the variants in the *M. tb* samples in respect to the reference H37Rv strain. Since the results were reported based on the number of variants detected at each culture condition, the clustering of the majority of the culture conditions in the polyclonally infected patients at a similar region suggests that culture selects for a specific genotype in sputum. When the genetic distances between the genomes of each patient were compared, the genetic distance increased when a different strain type emerged in culture. These findings are consistent with those obtained by Moreno-Molina *et al.* (2021), who used WGS to investigate the prevalence of polyclonal infections by analysing and comparing the genomic sequences derived from sputum and in surgical resections from tuberculosis patients. The genetic distance in polyclonally infected patients was the largest when comparing a Beijing or CAS strain to the other strain types in the polyclonal infection, which is similar to what Moreno-Molina *et al.* (2021) found from a comparison of lineage 2 strains (includes Beijing) to the other strains that were identified in patients with different genotypes in sputum and from surgical resections.

Simultaneous infections with multiple subpopulations

A comparison of the sequences from the same infecting strain genotype in each patient was used to investigate the effect of sputum culture on the clonal composition of sputum. The genetic distance between these strains ranged from 8 to 877 bp. Although a universal SNP-threshold to delineate the microevolution of strains has not been determined in literature, the genetic distances observed between these strains may suggest the presence of clonally heterogenous bacterial populations in sputum. These findings are comparable with those of

Liu *et al.* (2015), who found 10-14 bp differences in serial sputum samples while investigating clonal heterogeneity in sputum, and those of Faksri *et al.* (2016), who observed a 2-61 bp difference between serial isolates from patients using WGS to investigate genetic markers that distinguish infections and reinfections in MDR and XDR-TB patients. Factors such as selection pressures, sample sizes, the genotype of the different isolates and time differences between the sequenced isolates may contribute to this difference. These genetically distinct genotypes were detected in the various growth assays, specifically the (M1) media and CF-supplemented (RV1) MPN-assay, as well as the media supplemented MGIT assay, CF-supplemented MGIT assay and the RIPA/RPFB supplemented MGIT assay. For one patient (59172), this was observed upon culture in the CFU assay. These results highlight the differential culturability of bacteria in sputum. This is consistent with those of Chengalroyen et .al (2016) who was able to detect the presence of differentially culturable bacteria in sputum. These results, along with those from the previous analysis, support the study's hypothesis that different culture methods reveal differentially culturable populations, resulting in an increase in the observed genetic diversity within a single sputum sample.

Mutations in the *M. tb sigA* gene alter gene expression

The partial recovery of the genomic sequences from uncultured sputum samples allowed for the detection of mutations, if any, in the four first-line anti-TB drug targets as well as in two genes that are involved in the regulation of housekeeping genes, namely *sigA* and *sigB*. In all of the samples, no mutations were found in any of the drug resistance genes or the *sigB* gene. However, mutations in the *sigA* gene were found in the uncultured sputum of 5 patients. Sigma factors play an important role in transcription initiation by directing the RNA polymerase to bind to the -35 and -10 promoter regions and ensuring the stability of the RNA polymerase-promoter complex. (Tare *et al.*, 2013). The sig70 family consists of four regions, but the second region is the most highly conserved due to the presence of the -10-promoter recognition helix and the primary core RNA polymerase binding site (Campbell et al., 2002). As a result, the missense mutations that have been identified in the *sigA* gene alter the structure of the protein, which may affect the polymerase-promoter interaction, altering the transcription of important housekeeping genes such as virulence genes (Wu *et al.*, 2004). None of the mutations that were identified in sputum were identified in culture. The reason for this is not understood as the same mutations that are found in sputum should be present in culture.

Generation of barcode labelled strains

In this study, the goal of developing an artificial mix strain infection of *M. tb* to gain an understanding of strain competition within the host was not met. Individual *M. tb* strains that were selected for this model were transformed with a plasmid containing a 50 bp barcode to be used for the differentiation of the different strains within the artificial mixture. The transformation of the *M. tb* strains was successful however the clones did not contain the integrated plasmid. Exposing cells to high electrical voltages during electroporation increases the permeability of the membrane. The increased permeability is only temporary, which allows for the integration of the plasmid DNA into the chromosome. The failure of the plasmid to integrate could be attributed to the cells' inability to repair their membrane after pulsing, which could have occurred as a result of cell death caused by high electrical voltage exposure. The bacterial cells could also have been overly concentrated, making the plasmid concentration insufficient for transformation.

5. CONCLUDING REMARKS

The aims and objectives of this study in evaluating the effect of sputum culture on the within sample genetic diversity in different growth conditions was achieved. It was noted that spoligotyping is less sensitive in identifying mixed strain infections in sputum owing to the overlap of spoligotype patterns from the individual infecting strains and WGS was shown to have the ability to overcome this limitation. The different growth requirements of bacteria were demonstrated by the emergence of different infecting strain types in sputum under various growth conditions which may have an impact on the diagnosis of multiple *M. tb* strain infections in sputum. Furthermore, culture was shown to be selective of a specific genotype in sputum, favouring strains from the CAS and Beijing sub lineages. The detection of genetically heterogeneous *M. tb* bacterial populations in the sputum of patients who are already infected with multiple *M. tb* strains suggests that mixed strain *M. tb* infections may have complex infection patterns.

6. FUTURE WORK

Although the full impact of sputum culture on genetic diversity could not be determined due to partial recovery of genomic sequences, target enrichment of DNA derived from uncultured sputum may overcome this limitation and allow for comparison of sequence data from uncultured sputum samples to sequence data from cultured samples. Different software (such as QuantTB) may be used for the detection of mixed strain infections in future studies as they

provide different functionalities that can provide insightful information using WGS data. The implemented genetic barcoding system has the potential to provide insight into the precise mechanism of strain competitions within the host by providing an understanding of how different strains survive in mixed environments.

7. REFERENCES

Abubakar, I., Pimpin, L., Ariti, C., Beynon, R., Mangtani, P., Sterne, J., Fine, P., Smith, P., Lipman, M. & Elliman, D. (2013). Systematic review and meta-analysis of the current evidence on the duration of protection by bacillus Calmette-Guérin vaccination against tuberculosis. *Health technology assessment,* **17,** 1.

Acharya, B., Acharya, A., Gautam, S., Ghimire, S. P., Mishra, G., Parajuli, N. & Sapkota, B. (2020). Advances in diagnosis of Tuberculosis: an update into molecular diagnosis of *Mycobacterium tuberculosis*. *Molecular Biology Reports: An International Journal on Molecular and Cellular Biology,* **47,** 4065-4075.

Advani, J., Verma, R., Chatterjee, O., Pachouri, P. K., Upadhyay, P., Singh, R., Yadav, J., Naaz, F., Ravikumar, R. & Buggi, S. (2019). Whole genome sequencing of *Mycobacterium tuberculosis* clinical isolates from India reveals genetic heterogeneity and region-specific variations that might affect drug susceptibility. *Frontiers in microbiology,* **10,** 309.

Aguirre-Blanco, A. M., Lukey, P. T., Cliff, J. M. & Dockrell, H. M. (2007). Strain-dependent variation in *Mycobacterium bovis* BCG-induced human T-cell activation and gamma interferon production in vitro. *Infection and immunity,* **75,** 3197-3201.

Alonso-Rodríguez, N., Martínez-Lirola, M., Herránz, M., Sanchez-Benitez, M., Barroso, P., Bouza, E. & De Viedma, D. G. (2008). Evaluation of the new advanced 15-loci MIRU-VNTR genotyping tool in *Mycobacterium tuberculosis* molecular epidemiology studies. *BMC microbiology,* **8,** 1-9.

Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data. Available online at: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Ardito, F., Posteraro, B., Sanguinetti, M., Zanetti, S. & Fadda, G. (2001). Evaluation of BACTEC Mycobacteria Growth Indicator Tube (MGIT 960) automated system for drug susceptibility testing of *Mycobacterium tuberculosis. Journal of clinical microbiology,* **39,** 4440-4444.

Ayrapetyan, M., Williams, T. C. & Oliver, J. D. (2015). Bridging the gap between viable but non-culturable and antibiotic persistent bacteria. *Trends in microbiology,* **23,** 7-13.

Bainomugisa, A., Lavu, E., Hiashiri, S., Majumdar, S., Honjepari, A., Moke, R., Dakulala, P., Hill-Cawthorne, G. A., Pandey, S., Marais, B. J., Coulter, C. & Coin, L. (2018). Multi-clonal evolution of multi-drug-resistant/extensively drug-resistant *Mycobacterium tuberculosis* in a high-prevalence setting of Papua New Guinea for over three decades. *Microbial genomics,* **4,** e000147.

Barreto, M., Cunha, S., Pereira, S., Genser, B., Hijjar, M., Yury Ichihara, M., De Brito, S., Dourado, I., Cruz, A. & Santa'ana, C. (2005). Neonatal BCG protection against tuberculosis lasts for 20 years in Brazil. *The international journal of tuberculosis and lung disease,* **9,** 1171-1173.

Barzon, L., Lavezzo, E., Militello, V., Toppo, S. & Palù, G. (2011). Applications of next-generation sequencing technologies to diagnostic virology. *International Journal of Molecular Sciences,* **12,** 7861-7884.

Bates, J. H., Stead, W. & Rado, T. A. (1976). Phage type of tubercle bacilli isolated from patients with two or more sites of organ involvement. *American Review of Respiratory Disease,* **114,** 353-358.

Bates, M. N., Khalakdina, A., Pai, M., Chang, L., Lessa, F. & Smith, K. R. (2007). Risk of tuberculosis from exposure to tobacco smoke: a systematic review and meta-analysis. *Archives of internal medicine,* **167,** 335-342.

Beye, M., Fahsi, N., Raoult, D. & Fournier, P.-E. (2018). Careful use of 16S rRNA gene sequence similarity values for the identification of *Mycobacterium* species. *New microbes and new infections,* **22,** 24-29.

Bilen, M., Dufour, J.-C., Lagier, J.-C., Cadoret, F., Daoud, Z., Dubourg, G. & Raoult, D. (2018). The contribution of culturomics to the repertoire of isolated human bacterial and archaeal species. *Microbiome,* **6,** 94.

Black, P., De Vos, M., Louw, G., Van Der Merwe, R., Dippenaar, A., Streicher, E. M., Abdallah, A., Sampson, S., Victor, T. & Dolby, T. (2015). Whole genome sequencing reveals genomic heterogeneity and antibiotic purification in *Mycobacterium tuberculosis* isolates. *BMC genomics,* **16,** 1-14.

Böddinghaus, B., Rogall, T., Flohr, T., Blöcker, H. & Böttger, E. C. (1990). Detection and identification of mycobacteria by amplification of rRNA. *Journal of clinical microbiology,* **28,** 1751-1759.

Bradner, L., Robbe-Austerman, S., Beitz, D. & Stabel, J. (2013). Chemical decontamination with N-acetyl-l-cysteine–sodium hydroxide improves recovery of viable *Mycobacterium avium subsp. paratuberculosis* organisms from cultured milk. *Journal of clinical microbiology,* **51,** 2139-2146.

Brazier, B., McShane, H. (2020). Towards new TB vaccines. *Seminars in Immunopathology*, **42,** 315-331.

Bruchfeld, J., Correia-Neves, M. & Källenius, G. (2015). Tuberculosis and HIV coinfection. *Cold Spring Harbor Perspective in Medicine,* **5,** a017871.

Brudey, K., Driscoll, J. R., Rigouts, L., Prodinger, W. M., Gori, A., Al-Hajoj, S. A., Allix, C., Aristimuño, L., Arora, J. & Baumanis, V. (2006). *Mycobacterium tuberculosis* complex genetic diversity: mining the fourth international spoligotyping database (SpolDB4) for classification, population genetics and epidemiology. *BMC microbiology,* **6,** 1-17.

Bulane, A., (2012). Genetic relatedness of *Mycobacterium tuberculosis* strains obtained from Kalafong Hospital in Pretoria using Spoligotyping and MIRU-VNTR typing (Doctoral dissertation, University of Pretoria).

Buregyeya, E., Kulane, A., Colebunders, R., Wajja, A., Kiguli, J., Mayanja, H., Musoke, P., Pariyo, G. & Mitchell, E. (2011). Tuberculosis knowledge, attitudes, and health-seeking behaviour in rural Uganda. *The International Journal of Tuberculosis and Lung Disease,* **15,** 938-942.

Campbell, E. A., Muzzin, O., Chlenov, M., Sun, J. L., Olson, C. A., Weinman, O., Trester-Zedlitz, M. L. & Darst, S. A. (2002). Structure of the bacterial RNA polymerase promoter specificity σ subunit. *Molecular cell,* **9,** 527-539.

Camus, J.-C., Pryor, M. J., Médigue, C. & Cole, S. T. (2002). Re-annotation of the genome sequence of *Mycobacterium tuberculosis* H37Rv. *Microbiology,* **148,** 2967-2973.

Canetti, D., Riccardi, N., Martini, M., Villa, S., Di Biagio, A., Codecasa, L., Castagna, A., Barberis, I., Gazzaniga, V. & Besozzi, G. (2020). HIV and tuberculosis: The paradox of dual illnesses and the challenges of their fighting in the history. *Tuberculosis,* **122,** 1921.

Castañeda-Hernández, D. M. & Rodriguez-Morales, A. J. (2013). Epidemiological burden of tuberculosis in developing countries. *Current Topics in Public Health,* **2,** 317-340.

Cave, M. D., Eisenach, K. D., Mcdermott, P. F., Bates, J. H. & Crawford, J. T. (1991). IS6110: conservation of sequence in the *Mycobacterium tuberculosis* complex and its utilization in DNA fingerprinting. *Molecular and cellular probes,* **5,** 73-80.

Caws, M., Marais, B., Heemskerk, D. & Farrar, J. (2015). *Tuberculosis in adults and children*, Springer Nature.

Cerrone, M., Wang, X., Neary, M., Weaver, C., Fedele, S., Day-Weber, I., Owen, A., Hill, A., Mcclure, M. & Boffito, M. (2019). Pharmacokinetics of efavirenz 400 mg once daily coadministered with isoniazid and rifampicin in human immunodeficiency virus–infected individuals. *Clinical Infectious Diseases,* **68,** 446-452.

Chang, A. Y., Chau, V., Landas, J. A. & Pang, Y. (2017). Preparation of calcium competent *Escherichia coli* and heat-shock transformation. *JEMI methods,* **1,** 22-25.

Chaturvedi, S., Frame, P. & Newman, S. L. (1995). Macrophages from human immunodeficiency virus-positive persons are defective in host defense against *Histoplasma capsulatum*. *Journal of Infectious Diseases,* **171,** 320-327.

Chengalroyen, M. D., Beukes, G. M., Gordhan, B. G., Streicher, E. M., Churchyard, G., Hafner, R., Warren, R., Otwombe, K., Martinson, N. & Kana, B. D. (2016). Detection and quantification of differentially culturable tubercle bacteria in sputum from patients with tuberculosis. *American journal of respiratory and critical care medicine,* **194,** 1532-1540.

Cheo, D. L., Titus, S. A., Byrd, D. R., Hartley, J. L., Temple, G. F. & Brasch, M. A. (2004). Concerted assembly and cloning of multiple DNA segments using in vitro site-specific recombination: functional analysis of multi-segment expression clones. *Genome research,* **14,** 2111-2120.

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., Land, S. J., Lu, X. & Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly,* **6,** 80-92.

Cole, S., Brosch, R., Parkhill, J., Garnier, T., Churcher, C., Harris, D., Gordon, S., Eiglmeier, K., Gas, S. & Barry, C. R. (1998). Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature,* **396,** 190-190.

Coll, F., Mallard, K., Preston, M. D., Bentley, S., Parkhill, J., Mcnerney, R., Martin, N. & Clark, T. G. (2012). SpolPred: rapid and accurate prediction of *Mycobacterium tuberculosis* spoligotypes from short genomic sequences. *Bioinformatics,* **28,** 2991-2993.

Coll, F., Phelan, J., Hill-Cawthorne, G. A., Nair, M. B., Mallard, K., Ali, S., Abdallah, A. M., Alghamdi, S., Alsomali, M., Ahmed, A. O., Portelli, S., Oppong, Y., Alves, A., Bessa, T. B., Campino, S., Caws, M., Chatterjee, A., Crampin, A. C., Dheda, K., Furnham, N., Glynn, J. R., Grandjean, L., Ha, D. M., Hasan, R., Hasan, Z., Hibberd, M. L., Joloba, M., Jones-López, E. C., Matsumoto, T., Miranda, A., Moore, D. J., Mocillo, N., Panaiotov, S., Parkhill, J., Penha, C., Perdigão, J., Portugal, I., Rchiad, Z., Robledo, J., Sheen, P., Shesha, N. T., Sirgel, F. A., Sola, C., Sousa, E. O., Streicher, E. M., Van Helden, P., Viveiros, M., Warren, R. M., Mcnerney, R., Pain, A. & Clark, T. G. (2018). Genome-wide analysis of multi- and extensively drug-resistant *Mycobacterium tuberculosis*. *Nature Genetics,* **50,** 307-2,316A-316B.

Comas, I., Coscolla, M., Luo, T., Borrell, S., Holt, K. E., Kato-Maeda, M., Parkhill, J., Malla, B., Berg, S., Thwaites, G., Yeboah-Manu, D., Bothamley, G., Mei, J., Wei, L., Bentley, S., Harris, S. R., Niemann, S., Diel, R., Aseffa, A., Gao, Q., Young, D. & Gagneux, S. (2013). Out-of-Africa migration and neolithic coexpansion of *Mycobacterium tuberculosis* with modern humans. *Nature genetics,* **45,** 1176-1182.

Cruciani, M., Scarparo, C., Malena, M., Bosco, O., Serpelloni, G. & Mengoli, C. (2004). Meta-analysis of BACTEC MGIT 960 and BACTEC 460 TB, with or without solid media, for detection of mycobacteria. *Journal of clinical microbiology,* **42,** 2321-5.

Danecek, P., Schiffels, S. & Durbin, R. (2014). Multiallelic calling model in bcftools (-m). Available online at: http://samtools.github.io/bcftools/.

Datta, B., Hazarika, A., Shewade, H. D., Ayyagari, K. & Kumar, A. M. V. (2017). Digital chest X-ray through a mobile van: public private partnership to detect sputum negative pulmonary TB. *BMC Research Notes,* **10,** 96.

De Jong, B. C., Hill, P. C., Aiken, A., Awine, T., Martin, A., Adetifa, I. M., Jackson-Sillah, D. J., Fox, A., Kathryn, D. & Gagneux, S. (2008). Progression to active tuberculosis, but not transmission, varies by *Mycobacterium tuberculosis* lineage in The Gambia. *The Journal of infectious diseases,* **198,** 1037-1043.

Dekaboruah, E., Suryavanshi, M. V., Chettri, D. & Verma, A. K. (2020). Human microbiome: an academic update on human body site specific surveillance and its possible role. *Archives of microbiology,* **202,** 2147-2167.

Dheda, K., Barry, C. E., 3rd & Maartens, G. (2016). Tuberculosis. *Lancet,* **387,** 1211-26.

Do Thu, T., Kumar, A., Ramaswamy, G., Htun, T., Le Van, H., Nguyen Quang Vo, L., Thi Thu Dong, T., Codlin, A., Forse, R. & Crewsell, J. (2020). An innovative public–private mix model for improving tuberculosis care in Vietnam: how well are we doing? *Tropical medicine and infectious disease,* **5,** 26.

Donohue, M. J. (2018). Increasing nontuberculous mycobacteria reporting rates and species diversity identified in clinical laboratory reports. *BMC Infectious Diseases,* **18,** 163.

Dookie, N., Padayatchi, N., Lessells, R. J., Naicker, C. L., Chotoo, S. & Naidoo, K. (2020). Individualized treatment of multidrug-resistant tuberculosis using whole-genome sequencing and expanded drug-susceptibility testing. *Clinical Infectious Diseases,* **71,** 2981-2985.

Dörr, T., Lewis, K. & Vulić, M. (2009). SOS response induces persistence to fluoroquinolones in *Escherichia coli*. *PLoS genetics,* **5,** e1000760-e1000760.

Doughty, E. L., Sergeant, M. J., Adetifa, I., Antonio, M. & Pallen, M. J. (2014). Culture-independent detection and characterisation of *Mycobacterium tuberculosis* and *M. africanum* in sputum samples using shotgun metagenomics on a benchtop sequencer. *PeerJ,* **2,** e585.

Driscoll, J. R. (2009). Spoligotyping for molecular epidemiology of the *Mycobacterium tuberculosis* complex. *Methods in Molecular Biology,* **551,** 117-28.

Du Bruyn, E., Ruzive, S., Arlehamn, C. S. L., Sette, A., Sher, A., Barber, D. L., Wilkinson, R. J. & Riou, C. (2021). *Mycobacterium tuberculosis* specific CD4 T cells expressing CD153 inversely associate with bacterial load and disease severity in human tuberculosis. *Mucosal immunology,* **14,** 491-499.

Duarte, R., Lönnroth, K., Carvalho, C., Lima, F., Carvalho, A., Muñoz-Torrico, M. & Centis, R. (2018). Tuberculosis, social determinants, and co-morbidities (including HIV). *Pulmonology,* **24,** 115-119.

Dusthackeer, A., Balasubramanian, M., Shanmugam, G., Priya, S., Nirmal, C. R., Sam Ebenezer, R., Balasubramanian, A., Mondal, R. K., Thiruvenkadam, K. & Hemanth Kumar, A. (2019). Differential culturability of *Mycobacterium tuberculosis* in culture-negative sputum of patients with pulmonary tuberculosis and in a simulated model of dormancy. *Frontiers in microbiology,* **10,** 2381.

Dyavar, S. R., Mykris, T. M., Winchester, L. C., Scarsi, K. K., Fletcher, C. V. & Podany, A. T. (2020). Hepatocytic transcriptional signatures predict comparative drug interaction potential of rifamycin antibiotics. *Scientific reports,* **10,** 1-12.

Eldholm, V. & Balloux, F. 2016. Antimicrobial resistance in *Mycobacterium tuberculosis*: the odd one out. *Trends in microbiology,* **24,** 637-648.

Faksri, K., Tan, J. H., Disratthakit, A., Xia, E., Prammananan, T., Suriyaphol, P., Khor, C. C., Teo, Y. Y., Ong, R. T. H. & Chaiprasert, A. (2016). Whole-genome sequencing analysis of serially isolated multi-drug and extensively drug resistant *Mycobacterium tuberculosis* from Thai patients. *PLOS ONE,* **11,** e0160992.

Falzon, D., Timimi, H., Kurosinski, P., Migliori, G. B., Van Gemert, W., Denkinger, C., Isaacs, C., Story, A., Garfein, R. S. & Do Valle Bastos, L. G. (2016). Digital health for the End TB Strategy: developing priority products and making them work. *European Respiratory Journal,* **48,** 29-45.

Fang, Z., Morrison, N., Watt, B., Doig, C. & Forbes, K. J. (1998). IS6110 transposition and evolutionary scenario of the direct repeat locus in a group of closely related *Mycobacterium tuberculosis* strains. *Journal of Bacteriology,* **180,** 2102-9.

Farhat, M. R., Sultana, R., Iartchouk, O., Bozeman, S., Galagan, J., Sisk, P., Stolte, C., Nebenzahl-Guimaraes, H., Jacobson, K. & Sloutsky, A. (2016). Genetic determinants of drug resistance in *Mycobacterium tuberculosis* and their diagnostic value. *American journal of respiratory and critical care medicine,* **194,** 621-630.

Fatima, S., Kumari, A., Das, G. & Dwivedi, V. P. (2020). Tuberculosis vaccine: A journey from BCG to present. *Life sciences,* **252,** 117594.

Fiedler, S. & Wirth, R. (1988). Transformation of bacteria with plasmid DNA by electroporation. *Analytical biochemistry,* **170,** 38-44.

Filliol, I., Sola, C. & Rastogi, N. (2000). Detection of a previously unamplified spacer within the DR locus of *Mycobacterium tuberculosis*: epidemiological implications. *Journal of Clinical Microbiology,* **38,** 1231-4.

Fine, P. E. (1995). Variation in protection by BCG: implications of and for heterologous immunity. *The Lancet,* **346,** 1339-1345.

Folkvardsen, D. B., Norman, A., Andersen, Å. B., Rasmussen, E. M., Lillebaek, T. & Jelsbak, L. (2018). A major *Mycobacterium tuberculosis* outbreak caused by one specific genotype in a low-incidence country: exploring gene profile virulence explanations. *Scientific reports,* **8,** 1-8.

Forbes, B. A. (2017). Mycobacterial taxonomy. *Journal of clinical microbiology,* **55,** 380-383.

Ford, C., Yusim, K., Ioerger, T., Feng, S., Chase, M., Greene, M., Korber, B. & Fortune, S. (2012). *Mycobacterium tuberculosis*–heterogeneity revealed through whole genome sequencing. *Tuberculosis,* **92,** 194-201.

Forrellad, M. A., Klepp, L. I., Gioffré, A., Sabio Y Garcia, J., Morbidoni, H. R., Santangelo, M. D. L. P., Cataldi, A. A. & Bigi, F. (2013). Virulence factors of the *Mycobacterium tuberculosis* complex. *Virulence,* **4,** 3-66.

Franz, E., Gras, L. M. & Dallman, T. (2016). Significance of whole genome sequencing for surveillance, source attribution and microbial risk assessment of foodborne pathogens. *Current Opinion in Food Science,* **8,** 74-79.

Frick, M. & Lessem, E. (2020). Tuberculosis Research Funding Trends, 2005–2019. *New York, NY: Treatment Action Group*.

Gagneux, S. (2012). Host–pathogen coevolution in human tuberculosis. *Philosophical Transactions of the Royal Society B: Biological Sciences,* **367,** 850-859.

Gagneux, S. (2018). Ecology and evolution of *Mycobacterium tuberculosis*. *Nature Reviews Microbiology,* **16,** 202-213.

Gan, M., Liu, Q., Yang, C., Gao, Q. & Luo, T. (2016). Deep whole-genome sequencing to detect mixed infection of *Mycobacterium tuberculosis*. *PloS one,* **11,** e0159029.

García De Viedma, D., Marín, M., Ruiz Serrano, M. J., Alcalá, L. & Bouza, E. (2003). Polyclonal and compartmentalized infection by *Mycobacterium tuberculosis* in patients with both respiratory and extrarespiratory involvement. *The Journal of infectious diseases,* **187,** 695-699.

Gardy, J. L., Johnston, J. C., Sui, S. J. H., Cook, V. J., Shah, L., Brodkin, E., Rempel, S., Moore, R., Zhao, Y. & Holt, R. (2011). Whole-genome sequencing and social-

network analysis of a tuberculosis outbreak. *New England Journal of Medicine,* **364,** 730-739.

Gazzinelli, R. T., Bala, S., Stevens, R., Baseler, M., Wahl, L., Kovacs, J. & Sher, A. (1995). HIV infection suppresses type 1 lymphokine and IL-12 responses to Toxoplasma gondii but fails to inhibit the synthesis of other parasite-induced monokines. *The Journal of Immunology,* **155,** 1565-1574.

Genestet, C., Hodille, E., Westeel, E., Ginevra, C., Ader, F., Venner, S., Lina, G., Bryant, J. E., Berland, J.-L. & Dumitrescu, O. (2019a). Subcultured *Mycobacterium tuberculosis* isolates on different growth media are fully representative of bacteria within clinical samples. *Tuberculosis,* **116,** 61-66.

Genestet, C., Tatai, C., Berland, J. L., Claude, J. B., Westeel, E., Hodille, E., Fredenucci, I., Rasigade, J. P., Ponsoda, M., Jacomo, V., Vachée, A., Gaudart, A., Gaillard, J. L., Roux, A. L., Ader, F., Tararbit, K., Terpant, G., Bryant, J. E., Lina, G. & Dumitrescu, O. (2019b). Prospective whole-genome sequencing in tuberculosis outbreak investigation, France, 2017-2018. *Emerging Infectious Diseases,* **25,** 589-592.

Gengenbacher, M. & Kaufmann, S. H. E. (2012). *Mycobacterium tuberculosis*: success through dormancy. *FEMS Microbiology Reviews,* **36,** 514-532.

Getahun, H., Matteelli, A., Chaisson, R. E. & Raviglione, M. (2015). Latent *Mycobacterium tuberculosis* infection. *New England Journal of Medicine,* **372,** 2127-2135.

Gillespie, S. H. (2002). Evolution of drug resistance in *Mycobacterium tuberculosis:* clinical and molecular perspective. *Antimicrobial agents and chemotherapy,* **46,** 267-274.

Groenen, P. M., Bunschoten, A. E., Van Soolingen, D. & Van Embden, J. D. (1993). Nature of DNA polymorphism in the direct repeat cluster of *Mycobacterium tuberculosis*; application for strain differentiation by a novel typing method. *Molecular Microbiology,* **10,** 1057-65.

Guirado, E. & Schlesinger, L. (2013). Modeling the *Mycobacterium tuberculosis* granuloma–the critical battlefield in host immunity and disease. *Frontiers in immunology,* **4,** 98.

Guo, C., Ashrafian, H., Ghafur, S., Fontana, G., Gardner, C. & Prime, M. 2020. Challenges for the evaluation of digital health solutions—A call for innovative evidence generation approaches. *NPJ Digital Medicine,* **3,** 1-14.

Gutti, G., Arya, K. & Singh, S. K. (2019). Latent tuberculosis infection (LTBI) and its potential targets: An investigation into dormant phase pathogens. *Mini reviews in medicinal chemistry,* **19,** 1627-1642.

Hakamata, M., Takihara, H., Iwamoto, T., Tamaru, A., Hashimoto, A., Tanaka, T., Kaboso, S. A., Gebretsadik, G., Ilinov, A. & Yokoyama, A. (2020). Higher genome mutation rates of Beijing lineage of *Mycobacterium tuberculosis* during human infection. *Scientific reports,* **10,** 1-11.

Hancock, J. M. (1995). The contribution of slippage-like processes to genome evolution. *Journal of Molecular Evolution,* **41,** 1038-47.

Harris, R. L., Modayil, P., Adam, J., Sharland, M., Heath, P., Planche, T. & Daya, H. (2009). Cervicofacial nontuberculous mycobacterium lymphadenitis in children: Is surgery always necessary? *International Journal of Pediatric Otorhinolaryngology,* **73,** 1297-1301.

Heemskerk, D., Caws, M., Marais, B. & Farrar, J. (2015). Clinical manifestations. *Tuberculosis in Adults and Children.* Springer.

Hempelmann, J. A., Lockwood, C. M., Konnick, E. Q., Schweizer, M. T., Antonarakis, E. S., Lotan, T. L., Montgomery, B., Nelson, P. S., Klemfuss, N. & Salipante, S. J. (2018). Microsatellite instability in prostate cancer by PCR or next-generation sequencing. *Journal for immunotherapy of cancer,* **6,** 1-7.

Hett, E. C., Chao, M. C., Steyn, A. J., Fortune, S. M., Deng, L. L. & Rubin, E. J. (2007). A partner for the resuscitation-promoting factors of *Mycobacterium tuberculosis*. *Molecular Microbiology,* **66,** 658-668.

Hirsh, A. E., Tsolaki, A. G., Deriemer, K., Feldman, M. W. & Small, P. M. (2004). Stable association between strains of *Mycobacterium tuberculosis* and their human host populations. *Proceedings of the National Academy of Sciences,* **101,** 4871-4876.

Holley, R. W., Apgar, J., Everett, G. A., Madison, J. T., Marquisee, M., Merrill, S. H., Penswick, J. R. & Zamir, A. (1965). Structure of a ribonucleic acid. *Science***, 147,** 1462-1465.

Hope, J. C. & Villarreal-Ramos, B. (2008). Bovine TB and the development of new vaccines. *Comparative Immunology, Microbiology, and Infectious Diseases,* **31,** 77-100.

Horsburgh Jr, C. R. (1996). Epidemiology of disease caused by nontuberculous mycobacteria. *Seminars in respiratory infections,* **11**, 244-251.

Hosek, J., Svastova, P., Moravkova, M., Pavlik, I. & Bartos, M. (2006). Methods of mycobacterial DNA isolation from different biological material: a review. *Veterinarni medicina,* **51,** 180-192.

Houben, R. M. G. J. & Dodd, P. J. (2016). The global burden of latent tuberculosis infection: A re-estimation using mathematical modelling. *PLoS medicine,* **13,** e1002152-e1002152.

Hoyt, K. J., Sarkar, S., White, L., Joseph, N. M., Salgame, P., Lakshminarayanan, S., Muthaiah, M., Vinod Kumar, S., Ellner, J. J. & Roy, G. (2019). Effect of malnutrition on radiographic findings and mycobacterial burden in pulmonary tuberculosis. *PloS one,* **14,** e0214011.

Hsiao, K.-Y., Sun, H. S. & Tsai, S.-J. (2017). Circular RNA–new member of noncoding RNA with novel functions. *Experimental Biology and Medicine,* **242,** 1136-1141.

Hunter, R. L., Actor, J. K., Hwang, S.-A., Karev, V. & Jagannath, C. (2014). Pathogenesis of post primary tuberculosis: immunity and hypersensitivity in the development of cavities. *Annals of Clinical & Laboratory Science,* **44,** 365-387.

Huyen, M. N. T., Kremer, K., Lan, N. T. N., Cobelens, F. G. J., Buu, T. N., Dung, N. H., Caws, M., Tiemersma, E. W. & Van Soolingen, D. (2012). Mixed tuberculosis infections in rural South Vietnam. *Journal of clinical microbiology,* **50,** 1586-1592.

Iwamoto, T., Yoshida, S., Suzuki, K., Tomita, M., Fujiyama, R., Tanaka, N., Kawakami, Y. & Ito, M. (2007). Hypervariable loci that enhance the discriminatory ability of newly proposed 15-loci and 24-loci variable-number tandem repeat typing method on *Mycobacterium tuberculosis* strains predominated by the Beijing family. *FEMS microbiology letters,* **270,** 67-74.

Jagielski, T., Van Ingen, J., Rastogi, N., Dziadek, J., Mazur, P. K. & Bielecki, J. (2014). Current methods in the molecular typing of *Mycobacterium tuberculosis* and other mycobacteria. *BioMed Research International,* **2014,** 645802.

Jang, J., Becq, J., Gicquel, B., Deschavanne, P. & Neyrolles, O. (2008). Horizontally acquired genomic islands in the tubercle bacilli. *Trends in microbiology,* **16,** 303-308.

Kamerbeek, J., Schouls, L., Kolk, A., Van Agterveld, M., Van Soolingen, D., Kuijper, S., Bunschoten, A., Molhuizen, H., Shaw, R. & Goyal, M. (1997). Simultaneous detection and strain differentiation of *Mycobacterium tuberculosis* for diagnosis and epidemiology. *Journal of clinical microbiology,* **35,** 907-914.

Kana, B. D., Gordhan, B. G., Downing, K. J., Sung, N., Vostroktunova, G., Machowski, E. E., Tsenova, L., Young, M., Kaprelyants, A. & Kaplan, G. (2008). The resuscitation-promoting factors of *Mycobacterium tuberculosis* are required for virulence and resuscitation from dormancy but are collectively dispensable for growth in vitro. *Molecular microbiology,* **67,** 672-684.

Kaplan, G., Post, F. A., Moreira, A. L., Wainwright, H., Kreiswirth, B. N., Tanverdi, M., Mathema, B., Ramaswamy, S. V., Walther, G. & Steyn, L. M. (2003). *Mycobacterium tuberculosis* growth at the cavity surface: A microenvironment with failed immunity. *Infection and immunity,* **71,** 7099-7108.

Kato-Maeda, M., Rhee, J. T., Gingeras, T. R., Salamon, H., Drenkow, J., Smittipat, N. & Small, P. M. (2001). Comparing genomes within the species *Mycobacterium tuberculosis*. *Genome research,* **11,** 547-554.

Katoh, K., Misawa, K., Kuma, K. I. & Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic acids research,* **30,** 3059-3066.

Kchouk, M., Gibrat, J.-F. & Elloumi, M. (2017). Generations of sequencing technologies: from first to next generation. *Biology and Medicine,* **9**, 3.

Kedzierska, K., Mak, J., Mijch, A., Cooke, I., Rainbird, M., Roberts, S., Paukovics, G., Jolley, D., Lopez, A. & Crowe, S. M. (2000). Granulocyte-macrophage colony-stimulating factor augments phagocytosis of *Mycobacterium avium* complex by human immunodeficiency virus type 1-infected monocytes/macrophages in vitro and in vivo. *The Journal of infectious diseases,* **181,** 390-394.

Kendall, M. A., Lalloo, U., Fletcher, C. V., Wu, X., Podany, A. T., Cardoso, S. W., Ive, P. & Benson, C. A. (2021). Safety and pharmacokinetics of double-dose Lopinavir/Ritonavir+ Rifampin versus Lopinavir/Ritonavir+ daily Rifabutin for treatment of human immunodeficiency virus–tuberculosis coinfection. *Clinical Infectious Diseases,* **73,** 706–715.

Khieu, V., Ananta, P., Kaewprasert, O., Laohaviroj, M., Namwat, W. & Faksri, K. (2021). Whole-genome sequencing analysis to identify infection with multiple species of nontuberculous mycobacteria. *Pathogens,* **10,** 879.

Kirby, T. (2021). Global tuberculosis progress reversed by COVID-19 pandemic. *The Lancet Respiratory Medicine,* **9,** 118-119.

Kofler, R., Orozco-Terwengel, P., De Maio, N., Pandey, R. V., Nolte, V., Futschik, A., Kosiol, C. & Schlötterer, C. (2011). PoPoolation: a toolbox for population genetic analysis of next generation sequencing data from pooled individuals. *PloS one,* **6,** 15925.

Kozak, R. A., Alexander, D. C., Liao, R., Sherman, D. R., Behr, M. A. & Flynn, J. L. (2011). Region of difference 2 contributes to virulence of *Mycobacterium tuberculosis*. *Infection and Immunity*, **79,** 59-66.

Kulski, J. K. (2016). Next-generation sequencing—an overview of the history, tools, and "Omic" applications. *Next generation sequencing-advances, applications, and challenges,* **10,** 61964.

Lakoh, S., Jiba, D. F., Baldeh, M., Adekanmbi, O., Barrie, U., Seisay, A. L., Deen, G. F., Salata, R. A. & Yendewa, G. A. (2021). Impact of COVID-19 on tuberculosis case detection and treatment outcomes in Sierra Leone. *Tropical Medicine and Infectious Disease,* **6,** 154.

Langemeyer, L., Fröhlich, F. & Ungermann, C. (2018). Rab GTPase function in endosome and lysosome biogenesis. *Trends in cell biology,* **28,** 957-970.

Lazzari, T. K., Forte, G. C. & Silva, D. R. (2018). Nutrition Status Among HIV-Positive and HIV-Negative Inpatients with Pulmonary Tuberculosis. *Nutrition in Clinical Practice,* **33,** 858-864.

Lee, Y., Raviglione, M. C. & Flahault, A. (2020). Use of Digital Technology to Enhance Tuberculosis Control: Scoping Review. *Journal of medical Internet research,* **22,** e15727.

Leung, C. C., Lange, C. & Zhang, Y. (2013). Tuberculosis: current state of knowledge: an epilogue. *Respirology,* **18,** 1047-1055.

Li, H. & Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics,* **25,** 1754-1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. & Durbin, R. (2009). 1000 Genome project data processing subgroup. 2009. The sequence alignment/map format and samtools. *Bioinformatics,* **25,** 2078-2079.

Li, S., Liu, B., Peng, M., Chen, M., Yin, W., Tang, H., Luo, Y., Hu, P. & Ren, H. (2017). Diagnostic accuracy of Xpert MTB/RIF for tuberculosis detection in different regions with different endemic burden: A systematic review and meta-analysis. *PloS one,* **12,** e0180725.

Li, T.-L., Chan, T.-H., Wang, C.-H., Jou, R., Yu, M.-C., Putri, D. U., Lee, C.-H. & Lin, Y.-H. (2021). Acquired resistance to isoniazid during Isoniazid monotherapy in a subject with latent infection following household Rifampicin-resistant tuberculosis contact: A case report. *Infection and Drug Resistance,* **14,** 1505.

Liu, C. H., Liu, H. & Ge, B. (2017). Innate immunity in tuberculosis: host defense vs pathogen evasion. *Cellular & molecular immunology,* **14,** 963-975.

Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D., Lu, L. & Law, M. (2012). Comparison of next-generation sequencing systems. *Journal of Biomedicine and Biotechnology,* **2012**, 1-11.

Liu, Q., Via, L. E., Luo, T., Liang, L., Liu, X., Wu, S., Shen, Q., Wei, W., Ruan, X., Yuan, X., Zhang, G., Barry, C. E. & Gao, Q. (2015). Within patient microevolution of *Mycobacterium tuberculosis* correlates with heterogeneous responses to treatment. *Scientific Reports*, **5**, 17507.

Lombardi, G., Di Gregori, V., Girometti, N., Tadolini, M., Bisognin, F. & Dal Monte, P. (2017). Diagnosis of smear-negative tuberculosis is greatly improved by Xpert MTB/RIF. *PloS one,* **12,** e0176186.

Lönnroth, K., Jaramillo, E., Williams, B. G., Dye, C. & Raviglione, M. (2009). Drivers of tuberculosis epidemics: the role of risk factors and social determinants. *Social science & medicine,* **68,** 2240-2246.

López De Armentia, M. M., Amaya, C. & Colombo, M. I. (2016). Rab GTPases and the Autophagy Pathway: Bacterial targets for a suitable biogenesis and trafficking of their own vacuoles. *Cells,* **5,** 11.

Ludwig, W., Euzéby, J., Schumann, P., Busse, H.-J., Trujillo, M. E., Kämpfer, P. & Whitman, W. B. (2012). Road map of the phylum *Actinobacteria*. *Bergey's manual of systematic bacteriology.* **3,** 1-28.

Maguga-Phasha, N., Munyai, N., Mashinya, F., Makgatho, M. & Mbajiorgu, E. (2017). Genetic diversity and distribution of *Mycobacterium tuberculosis* genotypes in Limpopo, South Africa. *BMC infectious diseases,* **17,** 1-8.

Mandal, S., Njikan, S., Kumar, A., Early, J. V. & Parish, T. (2019). The relevance of persisters in tuberculosis drug discovery. *Microbiology,* **165,** 492-499.

Mankiewicz, E. & Liivak, M. (1975). Phage types of *Mycobacterium tuberculosis* in cultures isolated from Eskimo patients. *American Review of Respiratory Disease,* **111,** 307-312.

Martín, A., Herranz, M., Ruiz Serrano, M. J., Bouza, E. & García De Viedma, D. (2010). The clonal composition of *Mycobacterium tuberculosis* in clinical specimens could be modified by culture. *Tuberculosis,* **90,** 201-207.

Mascarello, A., Mori, M., Chiaradia-Delatorre, L. D., Menegatti, A. C. O., Monache, F. D., Ferrari, F., Yunes, R. A., Nunes, R. J., Terenzi, H., Botta, B. & Botta, M. (2013).

Discovery of *Mycobacterium tuberculosis* protein tyrosine phosphatase B (PtpB) inhibitors from natural products. *PLOS ONE,* **8,** e77081.

Mathema, B., Kurepina, N. E., Bifani, P. J. & Kreiswirth, B. N. (2006). Molecular epidemiology of tuberculosis: current insights. *Clinical microbiology reviews,* **19,** 658-685.

Mathuria, J. P., Sharma, P., Prakash, P., Samaria, J. K., Katoch, V. M. & Anupurba, S. (2008). Role of spoligotyping and IS6110-RFLP in assessing genetic diversity of *Mycobacterium tuberculosis* in India. *Infection, Genetics, and Evolution,* **8,** 346-51.

Mchenry, M. L., Bartlett, J., Igo, R. P., Jr., Wampande, E. M., Benchek, P., Mayanja-Kizza, H., Fluegge, K., Hall, N. B., Gagneux, S., Tishkoff, S. A., Wejse, C., Sirugo, G., Boom, W. H., Joloba, M., Williams, S. M. & Stein, C. M. (2020). Interaction between host genes and *Mycobacterium tuberculosis* lineage can affect tuberculosis severity: Evidence for coevolution? *PLOS Genetics,* **16,** e1008728.

Mechal, Y., Benaissa, E., Benlahlou, Y., Bssaibis, F., Zegmout, A., Chadli, M., Malik, Y. S., Touil, N., Abid, A. & Maleb, A. (2019). Evaluation of GeneXpert MTB/RIF system performances in the diagnosis of extrapulmonary tuberculosis. *BMC infectious Diseases,* **19,** 1-8.

Migliori, G. B., Tiberi, S., Zumla, A., Petersen, E., Chakaya, J. M., Wejse, C., Torrico, M. M., Duarte, R., Alffenaar, J. W. & Schaaf, H. S. (2020). MDR/XDR-TB management of patients and contacts: Challenges facing the new decade. The 2020 clinical update by the Global Tuberculosis Network. *International Journal of Infectious Diseases,* **92,** 15-25.

Miller, P. B., Zalwango, S., Galiwango, R., Kakaire, R., Sekandi, J., Steinbaum, L., Drake, J. M., Whalen, C. C. & Kiwanuka, N. (2021). Association between tuberculosis in men and social network structure in Kampala, Uganda. *BMC Infectious Diseases,* **21,** 1023.

Millet, J.-P., Moreno, A., Fina, L., Del Baño, L., Orcau, A., De Olalla, P. G. & Cayla, J. A. (2013). Factors that influence current tuberculosis epidemiology. *European Spine Journal,* **22,** 539-548.

Mohammadipanah F & Dehhaghi M. (2017). Classification and Taxonomy of Actinobacteria. *Biology and Biotechnology of Actinobacteria.* 51-77. Springer.

Moreno-Mendieta, S., Barrera-Rosales, A., Mata-Espinosa, D., Barrios-Payán, J., Sánchez, S., Hernández-Pando, R. & Rodríguez-Sanoja, R. (2019). Raw starch microparticles as BCG adjuvant: Their efficacy depends on the virulence of the infection strains. *Vaccine,* **37,** 5731-5737.

Moreno-Molina, M., Shubladze, N., Khurtsilava, I., Avaliani, Z., Bablishvili, N., Torres-Puente, M., Villamayor, L., Gabrielian, A., Rosenthal, A., Vilaplana, C., Gagneux, S., Kempker, R. R., Vashakidze, S. & Comas, I. (2021). Genomic analyses of *Mycobacterium tuberculosis* from human lung resections reveal a high frequency of polyclonal infections. *Nature Communications,* **12,** 2716.

Mukamolova, G. V., Kaprelyants, A. S., Young, D. I., Young, M. & Kell, D. B. (1998). A bacterial cytokine. *Proceedings of the National Academy of Sciences,* **95,** 8916-8921.

Mukamolova, G. V., Turapov, O., Malkin, J., Woltmann, G. & Barer, M. R. (2010). Resuscitation-promoting factors reveal an occult population of tubercle bacilli in sputum. *American journal of respiratory and critical care medicine,* **181,** 174-180.

Nabisere, R., Musaazi, J., Denti, P., Aber, F., Lamorde, M., Dooley, K. E., Aarnoutse, R., Sloan, D. J. & Sekaggya-Wiltshire, C. (2020). Pharmacokinetics, SAfety/tolerability, and EFficacy of high-dose RIFampicin in tuberculosis-HIV co-infected patients on efavirenz-or dolutegravir-based antiretroviral therapy: study protocol for an open-label, phase II clinical trial (SAEFRIF). *Trials,* **21,** 1-9.

Nahid, P., Dorman, S. E., Alipanah, N., Barry, P. M., Brozek, J. L., Cattamanchi, A., Chaisson, L. H., Chaisson, R. E., Daley, C. L., Grzemska, M., Higashi, J. M., Ho, C. S., Hopewell, P. C., Keshavjee, S. A., Lienhardt, C., Menzies, R., Merrifield, C., Narita, M., O'brien, R., Peloquin, C. A., Raftery, A., Saukkonen, J., Schaaf, H. S., Sotgiu, G., Starke, J. R., Migliori, G. B. & Vernon, A. (2016). Official American thoracic society/centers for disease control and prevention/infectious diseases society of America clinical practice guidelines: Treatment of drug-Susceptible tuberculosis. *Clinical Infectious Diseases,* **63,** e147-e195.

Nakamura, K., Oshima, T., Morimoto, T., Ikeda, S., Yoshikawa, H., Shiwa, Y., Ishikawa, S., Linak, M. C., Hirai, A. & Takahashi, H. (2011). Sequence-specific error profile of Illumina sequencers. *Nucleic acids research,* **39,** 90.

Nambot, S., Thevenon, J., Kuentz, P., Duffourd, Y., Tisserant, E., Bruel, A.-L., Mosca-Boidron, A.-L., Masurel-Paulet, A., Lehalle, D. & Jean-Marçais, N. (2018). Clinical whole-exome sequencing for the diagnosis of rare disorders with congenital anomalies and/or intellectual disability: substantial interest of prospective annual reanalysis. *Genetics in Medicine,* **20,** 645-654.

Namouchi, A., Didelot, X., Schöck, U., Gicquel, B. & Rocha, E. P. (2012). After the bottleneck: Genome-wide diversification of the *Mycobacterium tuberculosis* complex by mutation, recombination, and natural selection. *Genome research,* **22,** 721-734.

Narasimhan, P., Wood, J., Macintyre, C. R. & Mathai, D. (2013). Risk factors for tuberculosis. *Pulmonary Medicine,* **2013,** 828939.

Nathavitharana, R. R. & Friedland, J. S. (2015). A tale of two global emergencies: tuberculosis control efforts can learn from the Ebola outbreak. *European Respiratory Society,* ***46,*** 293-296.

Nguipdop-Djomo, P., Heldal, E., Rodrigues, L. C., Abubakar, I. & Mangtani, P. (2016). Duration of BCG protection against tuberculosis and change in effectiveness with time since vaccination in Norway: a retrospective population-based cohort study. *The Lancet infectious diseases,* **16,** 219-226.

Nhamoyebonde, S. & Leslie, A. (2014). Biological differences between the sexes and susceptibility to tuberculosis. *The Journal of infectious diseases,* **209,** 100-106.

Nieuwenhuizen, N. E. & Kaufmann, S. H. (2018). Next-generation vaccines based on bacille Calmette–Guérin. *Frontiers in immunology,* **9,** 121.

Nikitushkin, V., Demina, G. & Kaprelyants, A. (2016). Rpf proteins are the factors of reactivation of the dormant forms of actinobacteria. *Biochemistry,* **81,** 1719-1734.

Nikolayevskyy, V., Trovato, A., Broda, A., Borroni, E., Cirillo, D. & Drobniewski, F. (2016). MIRU-VNTR Genotyping of *Mycobacterium tuberculosis* strains using QIAxcel technology: A multicentre evaluation study. *PLOS ONE,* **11,** e0149435.

Nimmo, C., Shaw, L. P., Doyle, R., Williams, R., Brien, K., Burgess, C., Breuer, J., Balloux, F. & Pym, A. S. (2019). Whole genome sequencing *Mycobacterium tuberculosis* directly from sputum identifies more genetic diversity than sequencing from culture. *BMC genomics,* **20,** 1-9.

Oelofse, S., Esmail, A., Diacon, A., Conradie, F., Olayanju, O., Ngubane, N., Howell, P., Everitt, D., Crook, A. & Mendel, C. (2021). Pretomanid with bedaquiline and linezolid for drug-resistant TB: a comparison of prospective cohorts. *The International Journal of Tuberculosis and Lung Disease,* **25,** 453-460.

Okonechnikov, K., Conesa, A. & García-Alcalde, F. (2016). Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics,* **32,** 292-294.

Osman, M., Du Preez, K., Seddon, J. A., Claassens, M. M., Dunbar, R., Dlamini, S. S., Welte, A., Naidoo, P. & Hesseling, A. C. (2021). Mortality in South African children and adolescents routinely treated for tuberculosis. *Pediatrics,* **147,** e032490.

Padayatchi, N., Bionghi, N., Osman, F., Naidu, N., Ndjeka, N., Master, I., Brust, J., Naidoo, K. & Ramjee, A. (2020). Treatment outcomes in patients with drug-resistant TB-HIV co-infection treated with bedaquiline and linezolid. *The International Journal of Tuberculosis and Lung Disease,* **24,** 1024-1031.

Pai, M., Behr, M. A., Dowdy, D., Dheda, K., Divangahi, M., Boehme, C. C., Ginsberg, A., Swaminathan, S., Spigelman, M., Getahun, H., Menzies, D. & Raviglione, M. (2016). Tuberculosis. *Nature Reviews Disease Primers,* **2,** 16076.

Palomino, J. C., Leão, S. C. & Ritacco, V. (2007). *Tuberculosis (2007); from basic science to patient care.* Tuberculosistextbook. com, Brazil.

Pandey, P., Bhatnagar, A. K., Mohan, A., Sachdeva, K. S., Samantaray, J. C., Guleria, R. & Singh, U. B. (2020). *Mycobacterium tuberculosis* polyclonal infections through treatment and recurrence. *PLOS ONE,* **15,** e0237345.

Papakonstantinou, D., Dunn, S. J., Draper, S. J., Cunningham, A. F., O'shea, M. K. & Mcnally, A. (2021). Mapping gene-by-gene single-nucleotide variation in 8,535 *Mycobacterium tuberculosis* genomes: A resource to support potential vaccine and drug development. *Msphere,* **6,** e01224-20.

Penn-Nicholson, A., Georghiou, S. B., Ciobanu, N., Kazi, M., Bhalla, M., David, A., Conradie, F., Ruhwald, M., Crudu, V., Rodrigues, C., Myneedu, V. P., Scott, L., Denkinger, C. M. & Schumacher, S. G. (2022). Detection of isoniazid, fluoroquinolone, ethionamide, amikacin, kanamycin, and capreomycin resistance by the Xpert MTB/XDR assay: a cross-sectional multicentre diagnostic accuracy study. *Lancet Infectectious Diseases,* **22**, 242-249.

Pérez-Lago, L., Comas, I., Navarro, Y., González-Candelas, F., Herranz, M., Bouza, E. & García-De-Viedma, D. (2014). Whole genome sequencing analysis of intrapatient microevolution in *Mycobacterium tuberculosis*: potential impact on the inference of tuberculosis transmission. *The Journal of infectious diseases,* **209,** 98-108.

Petruševska, V., Lasić, K. & Mornar, A. (2020). Compatibility investigation for a new antituberculotic fixed dose combination with an adequate drug delivery. *Drug Development and Industrial Pharmacy,* **46,** 1298-1307.

Picard, T. 2018. Broad Institute, GitHub repository. Available online at: https://broadinstitute.github.io/picard/.

Podlesek, Z. & Žgur Bertok, D. (2020). The DNA damage inducible SOS response is a key player in the generation of bacterial persister cells and population wide tolerance. *Frontiers in Microbiology,* **11,** 1785.

Portevin, D., Gagneux, S., Comas, I. & Young, D. (2011). Human macrophage responses to clinical isolates from the *Mycobacterium tuberculosis* complex discriminate between ancient and modern lineages. *PLoS pathogens,* **7,** e1001307.

Prezzemolo, T., Guggino, G., La Manna, M. P., Di Liberto, D., Dieli, F. & Caccamo, N. (2014). Functional signatures of human CD4 and CD8 T cell responses to *Mycobacterium tuberculosis*. *Frontiers in immunology,* **5,** 180.

Quail, M. A., Swerdlow, H. & Turner, D. J. (2009). Improved protocols for the illumina genome analyzer sequencing system. *Current protocols in human genetics,* **62,** 18.2.1-18.2.27.

Quinlan, A. R. & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics,* **26,** 841-842.

Rasool, G., Khan, A. M., Mohy-Ud-Din, R. & Riaz, M. (2019). Detection of *Mycobacterium tuberculosis* in AFB smear-negative sputum specimens through MTB culture and

GeneXpert® MTB/RIF assay. *International journal of immunopathology and pharmacology,* **33,** 174.
Rehm, J., Samokhvalov, A. V., Neuman, M. G., Room, R., Parry, C., Lönnroth, K., Patra, J., Poznyak, V. & Popova, S. (2009). The association between alcohol use, alcohol use disorders and tuberculosis (TB). A systematic review. *BMC public health,* **9,** 1-12.
Ren, S.-X., Fu, G., Jiang, X.-G., Zeng, R., Miao, Y.-G., Xu, H., Zhang, Y.-X., Xiong, H., Lu, G. & Lu, L.-F. (2003). Unique physiological and pathogenic features of Leptospira interrogans revealed by whole-genome sequencing. *Nature,* **422,** 888-893.
Reuter, A., Hughes, J. & Furin, J. (2019). Challenges and controversies in childhood tuberculosis. *The Lancet,* **394,** 967-978.
Rich, M. & Jaramillo, E. (2008). *Guidelines for the programmatic management of drug-resistant tuberculosis,* World Health Organization.
Riethmueller, J., Vonthein, R., Borth-Bruhns, T., Grassmé, H., Eyrich, M., Schilbach, K., Stern, M. & Gulbins, E. (2008). DNA quantification and fragmentation in sputum after inhalation of recombinant human deoxyribonuclease. *Cellular Physiology and Biochemistry,* **22,** 347-352.
Ritz, N., Dutta, B., Donath, S., Casalaz, D., Connell, T. G., Tebruegge, M., Robins-Browne, R., Hanekom, W. A., Britton, W. J. & Curtis, N. (2012). The influence of bacille Calmette-Guerin vaccine strain on the immune response against tuberculosis: a randomized trial. *American journal of respiratory and critical care medicine,* **185,** 213-222.
Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G. & Mesirov, J. P. (2011). Integrative genomics viewer. *Nature biotechnology,* **29,** 24-26.
Rodrigues, L. C., Kerr-Pontes, L. R. S., Frietas, M. V. C. & Barreto, M. L. (2007). Long lasting BCG protection against leprosy. *Vaccine,* **25,** 6842-6844.
Roetzer, A., Diel, R., Kohl, T. A., Rückert, C., Nübel, U., Blom, J., Wirth, T., Jaenicke, S., Schuback, S., Rüsch-Gerdes, S., Supply, P., Kalinowski, J. & Niemann, S. (2013). Whole-genome sequencing versus traditional genotyping for investigation of a *Mycobacterium tuberculosis* Outbreak: A Longitudinal Molecular Epidemiological Study. *PLOS Medicine,* **10,** e1001387.
Rolandelli, A., Pellegrini, J. M., Hernández Del Pino, R. E., Tateosian, N. L., Amiano, N. O., Morelli, M. P., Castello, F. A., Casco, N., Levi, A. & Palmero, D. J. (2019). The non-synonymous rs763780 single-nucleotide polymorphism in IL17f gene is associated with susceptibility to tuberculosis and advanced disease severity in argentina. *Frontiers in immunology,* **10,** 2248.
Ru, H., Liu, X., Lin, C., Yang, J., Chen, F., Sun, R., Zhang, L. & Liu, J. (2017). The impact of genome region of difference 4 (RD4) on mycobacterial virulence and BCG efficacy. *Frontiers in Cellular and Infection Microbiology,* **7**, 239.
Ryu, Y. J. (2015). Diagnosis of pulmonary tuberculosis: recent advances and diagnostic algorithms. *Tuberculosis and respiratory diseases,* **78,** 64-71.
Sanger, F., Brownlee, G. G. & Barrell, B. G. (1965). A two-dimensional fractionation procedure for radioactive nucleotides. *Journal of molecular biology,* **13,** 373.
Sann, W. W. M., Namwat, W., Faksri, K., Swe, T. L., Swe, K. K., Thwin, T. & Sangka, A. (2020). Genetic diversity of *Mycobacterium tuberculosis* using 24-locus MIRU-VNTR typing and Spoligotyping in Upper Myanmar. *The Journal of Infection in Developing Countries,* **14,** 1296-1305.
Sanoussi, C. N. D., Affolabi, D., Rigouts, L., Anagonou, S. & De Jong, B. (2017). Genotypic characterization directly applied to sputum improves the detection of *Mycobacterium africanum* West African 1, under-represented in positive cultures. *PLoS neglected tropical diseases,* **11,** e0005900.

Schmutz, J., Wheeler, J., Grimwood, J., Dickson, M., Yang, J., Caoile, C., Bajorek, E., Black, S., Chan, Y. M. & Denys, M. (2004). Quality assessment of the human genome sequence. *Nature,* **429,** 365-368.

Shamputa, I. C., Jugheli, L., Sadradze, N., Willery, E., Portaels, F., Supply, P. & Rigouts, L. (2006). Mixed infection and clonal representativeness of a single sputum sample in tuberculosis patients from a penitentiary hospital in Georgia. *Respiratory research,* **7,** 1-10.

Shelby, P. W., Lia, M. P. & Israel, A. (2017). Collaborative public–private initiatives targeting multidrug-resistant tuberculosis (MDR-TB) supported by the Lilly MDR-TB Partnership: experiences in 2012–2016. *Journal of healthcare leadership,* **9,** 47.

Shin, S. S., Modongo, C., Baik, Y., Allender, C., Lemmer, D., Colman, R. E., Engelthaler, D. M., Warren, R. M. & Zetola, N. M. (2018). Mixed *Mycobacterium tuberculosis*-Strain Infections Are Associated With Poor Treatment Outcomes Among Patients With Newly Diagnosed Tuberculosis, Independent of Pretreatment Heteroresistance. *The Journal of infectious diseases,* **218,** 1974-1982.

Shleeva, M. O., Bagramyan, K., Telkov, M. V., Mukamolova, G. V., Young, M., Kell, D. B. & Kaprelyants, A. S. (2002). Formation and resuscitation of "non-culturable" cells of *Rhodococcus rhodochrous* and *Mycobacterium tuberculosis* in prolonged stationary phase. *Microbiology,* **148,** 1581-1591.

Shockey, A. C., Dabney, J. & Pepperell, C. S. (2019). Effects of host, sample, and in vitro culture on genomic diversity of pathogenic mycobacteria. *Frontiers in genetics,* **10,** 477.

Sinshaw, W., Kebede, A., Bitew, A., Tesfaye, E., Tadesse, M., Mehamed, Z., Yenew, B., Amare, M., Dagne, B. & Diriba, G. (2019). Prevalence of tuberculosis, multidrug resistant tuberculosis, and associated risk factors among smear negative presumptive pulmonary tuberculosis patients in Addis Ababa, Ethiopia. *BMC infectious diseases,* **19,** 1-15.

Snow, K. J., Cruz, A. T., Seddon, J. A., Ferrand, R. A., Chiang, S. S., Hughes, J. A., Kampmann, B., Graham, S. M., Dodd, P. J. & Houben, R. M. (2020). Adolescent tuberculosis. *The Lancet Child & Adolescent Health,* **4,** 68-79.

Sobkowiak, B., Glynn, J. R., Houben, R. M., Mallard, K., Phelan, J. E., Guerra-Assunção, J. A., Banda, L., Mzembe, T., Viveiros, M. & Mcnerney, R. (2018). Identifying mixed *Mycobacterium tuberculosis* infections from whole genome sequence data. *BMC genomics,* **19,** 1-15.

Sohn, H., Aero, A. D., Menzies, D., Behr, M., Schwartzman, K., Alvarez, G. G., Dan, A., Mcintosh, F., Pai, M. & Denkinger, C. M. (2014). Xpert MTB/RIF testing in a low tuberculosis incidence, high-resource setting: limitations in accuracy and clinical impact. *Clinical infectious diseases,* **58,** 970-976.

Sonnenberg, M. G. & Belisle, J. T. (1997). Definition of *Mycobacterium tuberculosis* culture filtrate proteins by two-dimensional polyacrylamide gel electrophoresis, N-terminal amino acid sequencing, and electrospray mass spectrometry. *Infection and immunity*, **65**, 4515-4524.

Soundararajan, L., Kambli, P., Priyadarshini, S., Let, B., Murugan, S., Iravatham, C., Tornheim, J. A., Rodrigues, C., Gupta, R. & Ramprasad, V. L. (2020). Whole genome enrichment approach for rapid detection of *Mycobacterium tuberculosis* and drug resistance-associated mutations from direct sputum sequencing. *Tuberculosis,* **121,** 101915.

Sreevatsan, S., Pan, X., Stockbauer, K. E., Connell, N. D., Kreiswirth, B. N., Whittam, T. S. & Musser, J. M. (1997). Restricted structural gene polymorphism in the

Mycobacterium tuberculosis complex indicates evolutionarily recent global dissemination. *Proceedings of the National Academy of Sciences of the United States of America,* **94,** 9869-9874.

Srinidhi Desikan, S. N. (2015). Genetic markers, genotyping methods & next generation sequencing in *Mycobacterium tuberculosis*. *The Indian journal of medical research,* **141,** 761.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics,* **30,** 1312-1313.

Sterling, T. R., Pham, P. A. & Chaisson, R. E. (2010). HIV infection—related tuberculosis: clinical manifestations and treatment. *Clinical Infectious Diseases,* **50,** 223-230.

Stoler, N. & Nekrutenko, A. (2021). Sequencing error profiles of Illumina sequencing instruments. *NAR Genomics and Bioinformatics,* **3**, 19.

Sykes, J. E. (2013). *Canine and feline infectious diseases*, Elsevier Health Sciences.

Tagliani, E., Anthony, R., Kohl, T. A., De Neeling, A., Nikolayevskyy, V., Ködmön, C., Maurer, F. P., Niemann, S., Van Soolingen, D. & Van Der Werf, M. J. (2021). Use of a whole genome sequencing-based approach for *Mycobacterium tuberculosis* surveillance in Europe in 2017–2019: An ECDC pilot study. *European Respiratory Journal,* **57**, 272.

Tait, D. R., Hatherill, M., Van Der Meeren, O., Ginsberg, A. M., Van Brakel, E., Salaun, B., Scriba, T. J., Akite, E. J., Ayles, H. M. & Bollaerts, A. (2019). Final analysis of a trial of M72/AS01E vaccine to prevent tuberculosis. *New England Journal of Medicine,* **381,** 2429-2439.

Tameris, M. D., Hatherill, M., Landry, B. S., Scriba, T. J., Snowden, M. A., Lockhart, S., Shea, J. E., Mcclain, J. B., Hussey, G. D. & Hanekom, W. A. (2013). Safety and efficacy of MVA85A, a new tuberculosis vaccine, in infants previously vaccinated with BCG: a randomised, placebo-controlled phase 2b trial. *The Lancet,* **381,** 1021-1028.

Tare, P., Mallick, B. & Nagaraja, V. (2013). Co-evolution of specific amino acid in sigma 1.2 region and nucleotide base in the discriminator to act as sensors of small molecule effectors of transcription initiation in mycobacteria. *Molecular Microbiology,* **90,** 569-583.

Taye, H., Alemu, K., Mihret, A., Ayalew, S., Hailu, E., Wood, J. L. N., Shkedy, Z., Berg, S., Aseffa, A. & The, E. C. (2021). Epidemiology of *Mycobacterium tuberculosis* lineages and strain clustering within urban and peri-urban settings in Ethiopia. *PLOS ONE,* **16,** e0253480.

Teppawar, R. N., Chaudhari, S. P., Moon, S. L., Shinde, S. V., Khan, W. A. & Patil, A. R. (2018). *Zoonotic tuberculosis: a concern and strategies to combat*. Intechopen.

Tweed, C. D., Dawson, R., Burger, D. A., Conradie, A., Crook, A. M., Mendel, C. M., Conradie, F., Diacon, A. H., Ntinginya, N. E., Everitt, D. E., Haraka, F., Li, M., Van Niekerk, C. H., Okwera, A., Rassool, M. S., Reither, K., Sebe, M. A., Staples, S., Variava, E. & Spigelman, M. (2019). Bedaquiline, moxifloxacin, pretomanid, and pyrazinamide during the first 8 weeks of treatment of patients with drug-susceptible or drug-resistant pulmonary tuberculosis: a multicentre, open-label, partially randomised, phase 2b trial. *Lancet Respiratory Medicine,* **7,** 1048-1058.

Uren, C., Hoal, E. G. & Möller, M. (2021). *Mycobacterium tuberculosis* complex and human coadaptation: a two-way street complicating host susceptibility to TB. *Human Molecular Genetics,* **30,** 146-153.

Van vliet, S. (2015). Bacterial dormancy: how to decide when to wake up. *Current Biology,* **25,** 753-755.

Victor, T. C., De Haas, P. E., Jordaan, A. M., Van Der Spuy, G. D., Richardson, M., Van Soolingen, D., Van Helden, P. D. & Warren, R. (2004). Molecular characteristics and global spread of *Mycobacterium tuberculosis* with a western cape F11 genotype. *Journal of clinical microbiology,* **42,** 769-772.

Vitol, I., Driscoll, J., Kreiswirth, B., Kurepina, N. & Bennett, K. P. (2006). Identifying *Mycobacterium tuberculosis* complex strain families using spoligotypes. *Infection, Genetics and Evolution,* **6,** 491-504.

Walker, T. M., Kohl, T. A., Omar, S. V., Hedge, J., Del Ojo Elias, C., Bradley, P., Iqbal, Z., Feuerriegel, S., Niehaus, K. E., Wilson, D. J., Clifton, D. A., Kapatai, G., Ip, C. L. C., Bowden, R., Drobniewski, F. A., Allix-Béguec, C., Gaudin, C., Parkhill, J., Diel, R., Supply, P., Crook, D. W., Smith, E. G., Walker, A. S., Ismail, N., Niemann, S. & Peto, T. E. A. (2015). Whole-genome sequencing for prediction of *Mycobacterium tuberculosis* drug susceptibility and resistance: a retrospective cohort study. *Lancet Infectious Diseases,* **15,** 1193-1202.

Warren, R. M., Victor, T. C., Streicher, E. M., Richardson, M., Beyers, N., Van Pittius, N. C. G. & Van Helden, P. D. (2004). Patients with active tuberculosis often have different strains in the same sputum specimen. *American journal of respiratory and critical care medicine,* **169,** 610-614.

Watson, J. D. & Crick, F. H. (1953). Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. *Nature,* **171,** 737-738.

Welsh, K. J., Risin, S. A., Actor, J. K. & Hunter, R. L. (2011). Immunopathology of postprimary tuberculosis: increased T-regulatory cells and DEC-205-positive foamy macrophages in cavitary lesions. *Clinical and Developmental Immunology,* **2011**, 307631.

Wiens, K. E., Woyczynski, L. P., Ledesma, J. R., Ross, J. M., Zenteno-Cuevas, R., Goodridge, A., Ullah, I., Mathema, B., Djoba Siawaya, J. F., Biehl, M. H., Ray, S. E., Bhattacharjee, N. V., Henry, N. J., Reiner, R. C., Kyu, H. H., Murray, C. J. L. & Hay, S. I. (2018). Global variation in bacterial strains that cause tuberculosis disease: a systematic review and meta-analysis. *BMC Medicine,* **16,** 196.

Wilfinger, W. W., Mackey, K. & Chomczynski, P. (1997). Effect of pH and ionic strength on the spectrophotometric assessment of nucleic acid purity. *Biotechniques,* **22,** 474-481.

Wong, D., Bach, H., Sun, J., Hmama, Z. & Av-Gay, Y. 2011. *Mycobacterium tuberculosis* protein tyrosine phosphatase (PtpA) excludes host vacuolar-H+–ATPase to inhibit phagosome acidification. *Proceedings of the National Academy of Sciences,* **108,** 19371-19376.

Wood, D. E. & Salzberg, S. L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome biology,* **15,** 1-12.

World Health Organization (2013). *Automated real-time nucleic acid amplification technology for rapid and simultaneous detection of tuberculosis and rifampicin resistance: Xpert MTB.* World Health Organization.

World Health Organization (2015). *Digital health for the End TB Strategy: an agenda for action*. World Health Organization.

World Health Organization (2020). *Global tuberculosis report 2020*.

World Health Organization (2021). *Global tuberculosis report 2021*.

Wu, S., Howard, S. T., Lakey, D. L., Kipnis, A., Samten, B., Safi, H., Gruppo, V., Wizel, B., Shams, H., Basaraba, R. J., Orme, I. M. & Barnes, P. F. (2004). The principal sigma factor sigA mediates enhanced growth of *Mycobacterium tuberculosis* in vivo. *Molecular Microbiology,* **51,** 1551-62.

Yang, S.-J., Chen, Y.-Y., Hsu, C.-H., Hsu, C.-W., Chang, C.-Y., Chang, J.-R. & Dou, H.-Y. (2020). Activation of M1 macrophages in response to recombinant TB vaccines with enhanced antimycobacterial activity. *Frontiers in Immunology*, **11,** 1298.

Yang, Z., Yang, D., Kong, Y., Zhang, L., Marrs, C. F., Foxman, B., Bates, J. H., Wilson, F. & Cave, M. D. (2005). Clinical relevance of *Mycobacterium tuberculosis* plcD gene mutations. *American journal of respiratory and critical care medicine*, **171,** 1436-1442.

8. APPENDICES

8.1. Appendix A- Materials

1. Culture media preparation

BBL™ Middlebrook 7H11 media

For 1 L of media, 20.5g of BBL™ Middlebrook 7H11 media was added to a 1L autoclaved bottle. Add 900 mL of distilled water and 5 mL of glycerol. Mix solution until dissolved then autoclave. Once autoclaving is complete, cool media to 55°C before the addition of 100 mL of OADC.

BD BBL™ Middlebrook 7H9 media

Weigh 4.7g of BD BBL™ Middlebrook powder and add to a 1L autoclaved bottle. Add 900 mL of distilled water and 2 mL of glycerol. Mix solution until dissolved then autoclave. Once autoclaved, cool the media, and add 2 mL of Tween®80 and 100 mL of OADC. Filter sterilise media using a stericup.

BD BBL™ Middlebrook 7H9 media with PANTA

Reconstitute a bottle of lyophilized BBL™ PANTA antibiotic mixture in 15 mL of BD BBL™ Middlebrook 7H9 liquid media. Add 2 mL of the reconstituted PANTA antibiotic mixture to 50 mL of filter sterilised BD BBL™ Middlebrook 7H9 media.

Luria-Bertani Agar (LA)

Weigh 31 g of LA powder and add to a 1L autoclaved bottle. Add 900 mL of distilled water and dissolve by stirring on a stirring pad with heat. Autoclave, allow to cool before supplementation with the necessary antibiotics.

Luria- Bertani Broth (LB)

Weigh 22.5 g of LB powder and add to a 1 L autoclaved bottle. Dissolve in 900 mL of distilled water. Autoclave. Once autoclaved, cool the media and supplement with the necessary antibiotics.

2X TY

Weigh 9 g of NaCl, 9 g of yeast extract and 18 g of tryptone. Add to a 1L autoclaved bottle and dissolve in 900 mL of distilled water. Autoclave.

2. Supplementation of Media

Table A1: Media supplements

Media supplements	Components	Description
Glycerol (2%)	Glycerol	Commercial (Sigma-Aldrich)
Tween (20%)	Tween80, distilled water	10 mL of Tween-80 dissolved in 40 mL dH_2O
OADC	Oleic Acid Albumin Dextrose Catalase	Commercial (BD BBL™)
PANTA	Polymyxin B, Amphotericin B, Nalidixic acid, Trimethoprim and Azlocillin	Commercial (BD BBL™)
Kanamycin (50 mg/ml)	Kanamycin, distilled water	2.5 mg of Kanamycin dissolved in 50 ml dH_20

Table A2: Chemical decontamination reagents

Reagent	Components	Description
4% NaOH-NaLC solution	Sodium hydroxide, N-acetyly-L-cysteine	Commercial (Diagnostic Media products)
Phosphate buffered saline (pH 7-7.4)	Phosphate buffered saline	Commercial (Diagnostic Media products)
2mm Silica glass beads	Glass beads	Commercial (Sigma-Aldrich)

Table A3: DNA extraction reagents

Reagent	Components	Description
TE Buffer (10x)	10mM Tris-HCL (pH 8), 10mM EDTA	31.5 g Tris-HCL and 0.58g EDTA dissolved in 200 ml of distilled water and autoclaved
10% SDS	20g sodium dodecyl sulfate	20 g of SDS powder dissolved in 200 ml of distilled water and autoclaved
NaCl (5M)	NaCl	14.61 g of NaCl dissolved in 50 ml of distilled water and autoclaved
CTAB/NaCL	4.1% NaCl, 10% N-cetl-N, N, N-trimetyl ammonium bromide	10 g of CTAB and 4.1 g of NaCL dissolved in 100 ml of distilled water and filter sterilised
Chloroform:isoamyl alcohol	100% Chloroform and 100% isoamyl alcohol.	24 ml chloroform and 1ml isoamyl alcohol
Isopropanol	100% iso-propyl-alcohol	Used as is

Table A4: Agarose gel electrophoresis solutions

SYBR safe stain	Commercial (Invitrogen)
50XTAE stock	242 g Tris base, 57.1 ml glacial acetic, and 100 ml EDTA (pH 8) made up to 1L with sdH2O
1xTAE working solution	40 ml of 50xTAE dissolved in 1600 mL of distilled water

Table A5: Spoligotyping reagents

Stock solutions

Reagent	Components	Description
10X SSPE	13.7 g Sodium Dihydrogen Phosphate, 105.19 g Sodium Chloride, 3.36 g EDTA.	Components were dissolved in 1L of distilled water and the pH was adjusted to 7.4.
10% SDS	50 g Sodium dodecyl sulphate	Dissolved in 500 mL of distilled water.
0.5M Ethylenediaminetetraacetic acid (EDTA)	93 g EDTA	Dissolved in 500 ml of distilled water

Working solutions

Reagent	Components	Description
2x SSPE/0.1%SDS	200 ml of 10x SSPE, 10ml of 10% SDS	Components were dissolved in 1L of distilled water.
2x SSPE/0.5%SDS	200 ml of 10x SSPE, 50ml of 10% SDS	Components were dissolved in 1L of distilled water.
1% SDS	20 ml of 10% SDS	Dissolved in 200 mL of distilled water
20 mM EDTA	20 ml of 0.5M EDTA	Dissolved in 500 mL of distilled water

Table A5: Spoligotyping materials

Products	Supplier
DRa & DRb primers	Mapmygemone (Hyderabad, India)
BCG and H3Rv controls	Mapmygemone (Hyderabad, India)
Spoligotyping membranes	Mapmygemone (Hyderabad, India)
Streptavidin-peroxidase conjugate	GE Healthcare (South Africa)
ECL detection kit	GE Healthcare (South Africa)
X-ray films	Thermofischer Scientific (South Africa)
Developer solution	Africa X-ray Industrial and Medical (South Africa)
Fixer solution	Africa X-ray Industrial and Medical (South Africa)

Molecular marker IV and V was used in this study

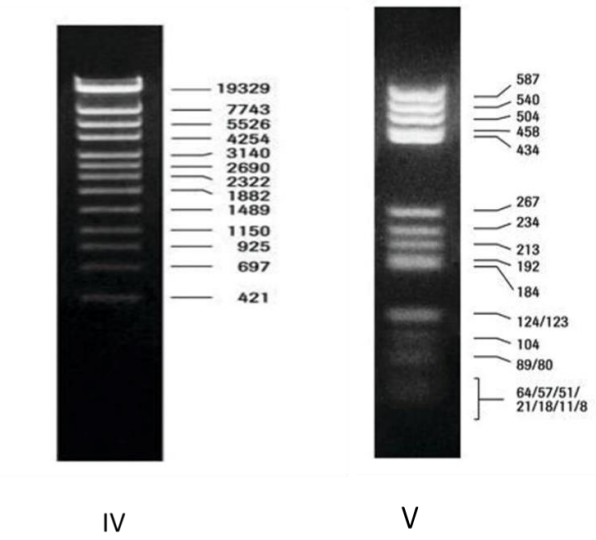


Figure A1: DNA molecular weight markers, iv and v used in this study.

8.2. Appendix B – Bioinformatics tools used for data analysis

Table B1: Software used for analysis

Program	Description
FAST QC	FAST QC is a command line and interactive software that assesses the quality of raw sequence data. This software was used to evaluate the quality of the paired end reads (available at: https://github.com/s-andrews/FastQC).
Burrows-Wheeler aligner	Burrows-wheeler aligner is a command line tool that maps raw sequence data against a known reference genome. This software was used to map the raw sequence data of the clinical samples against the *Mycobacterium tuberculosis* H37Rv reference genome (available at: https://github.com/lh3/bwa).
Samtools	Samtools is a multi-functional open-source command line tool that manipulates files with mapped reads. This tool was used for the conversion of the mapped files from the sequence alignment map format (SAM) to binary alignment map (BAM) format, sorting of aligned reads by chromosomal co-ordinates, indexing of the BAM file for visualisation and obtaining basic alignment statistics (available at: https://github.com/samtools/).
QualiMap	QualiMap is a command line and interactive software that facilitates the quality control of the aligned sequencing data. This software was used to determine the genome coverage of the aligned sequencing reads (available at: https://github.com/EagleGenomics-cookbooks/QualiMap)

Interactive genome browser (IGV)	Interactive genome browser is a graphical user interface that visualises genomic data. This software was used to visually inspect genomic variants (available at: https://software.broadinstitute.org/software/igv/).
Picard tools	Picard tools is a multi-functional command line tool that manipulates sequencing data. This tool was used to mark duplicate reads in the BAM files (available at: https://github.com/broadinstitute/picard).
BCFtools	BCFtools is a multi-functional command line tool that calls variants and manipulates variant call files. This tool was used for variant calling and filtering according to set quality thresholds (available at: https://github.com/samtools/bcftools).
BEDtools	Bedtools is a multi-functional command line tool that manipulates sequencing data. This tool was used for filtering variants that occur in known repeat regions of the *Mycobacterium tuberculosis* genome (available at: https://github.com/arq5x/bedtools).
SnpEff	SnpEff is an open-source command line tool that annotates variants and predicts their functional effects. This tool was used for the annotation and prediction of the functional consequences of high-quality variants (available at: https://github.com/pcingola/SnpEff).
Snp-dists	Snp-dists is a command line program that is used in the analysis of DNA sequences. This software was used to calculate the genetic distances between the genomes (available at: https://github.com/tseemann/snp-dists).
RaxML (Random accelerated maximum likelihood)	RaxML is a command line tool used to infer evolutionary relationships between DNA sequences. This tool was used to infer the maximum likelihood phylogeny of the sequencing data (available at: https://github.com/stamatak/standard-RAxML).
iTOL (interactive tree of life)	iTOL This tool was used to visualise the phylogeny of the sequence data (available at https://itol.embl.de/upload.cgi).
Kraken	Kraken is a command line and interactive tool that taxonomically classifies DNA sequences in metagenomic analyses. This software was used to assess the level of contaminating DNA sequences in the *M. tuberculosis* clinical samples (available at: https://github.com/DerrickWood/kraken).
Pavian	Pavian is an interactive tool that evaluates taxonomically classified metagenomic data. This tool was used to compare the level of contaminants in each of the clinical samples (available at: https://fbreitwieser.shinyapps.io/pavian/).
SPOTCLUST TB	SPOTCLUST is an interactive software used for Tuberculosis epidemiology and evolution mathematical modelling which is used for genotyping and obtaining epidemiological data on infectious diseases. This tool was used for determining the strain types of the clinical samples (available at: https://tbinsight.cs.rpi.edu/run_spotclust.html).

Appendix C: Supplementary figures

Figure C1- Illustration of the cloning strategy of the pUC57-kan vectors into the *M. tb* genome

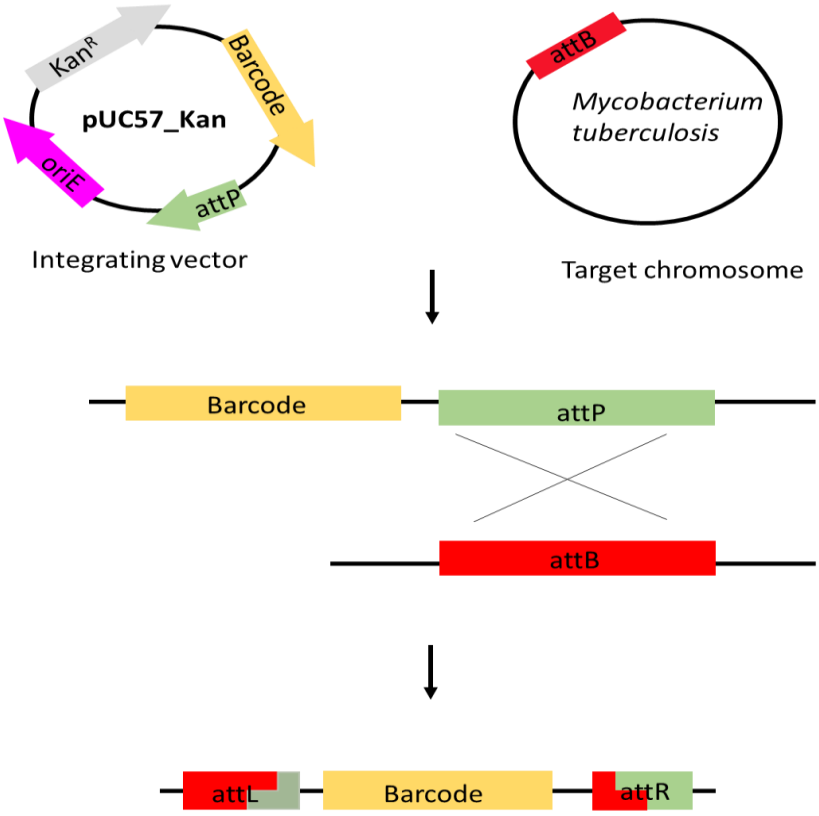


Figure 1: Illustration of the genomic integration of the DNA barcodes by a site-specific recombination event. An integrating vector containing the genetic barcode can integrate into the target chromosome at the bacterial attachment site (*attB*). A recombination reaction between *attP* (phage attachment site) and *attB* results in the formation of the *att*L and *attR* sites which contain the left and right arms of *attP* site after ligation takes place.

Figure C2: Hybridization blots from the Spoligotyping analysis

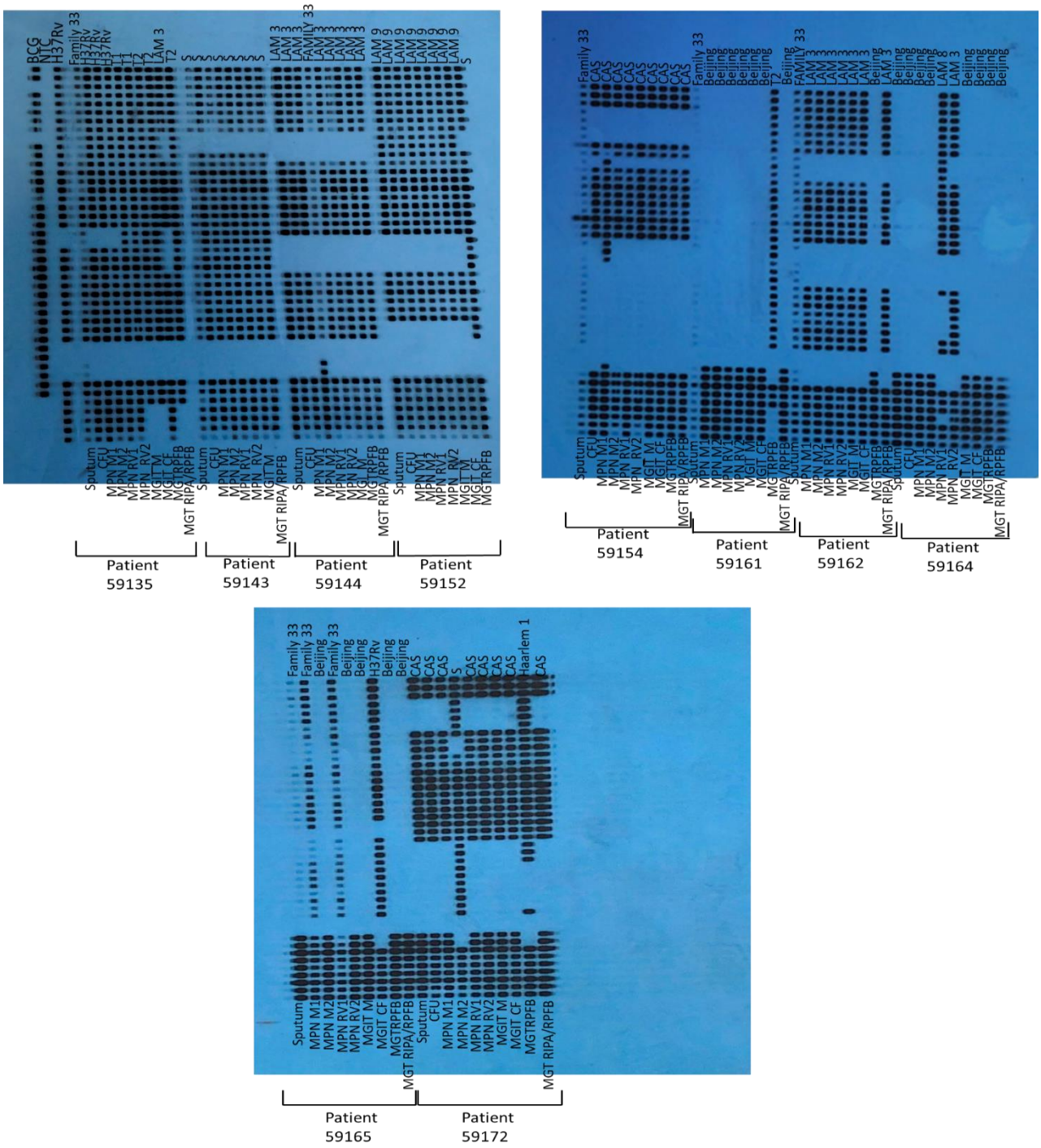


Figure 2: Hybridisation blots of the spoligotyping analysis. Depicts the spoligotype pattern and predicted strain type (by SpotClust) of each *M. tb* sample. Spoligotype patterns were obtained for each patients uncultured sputum sample and sputum cultured under the various growth conditions. Black dots indicate the presence of a spacer sequence at the specific location in the DR locus. The pattern is read from top to bottom.

Figure C3 – Visualization of the two mutations that were common in the *sigA* gene of the uncultured sputum samples

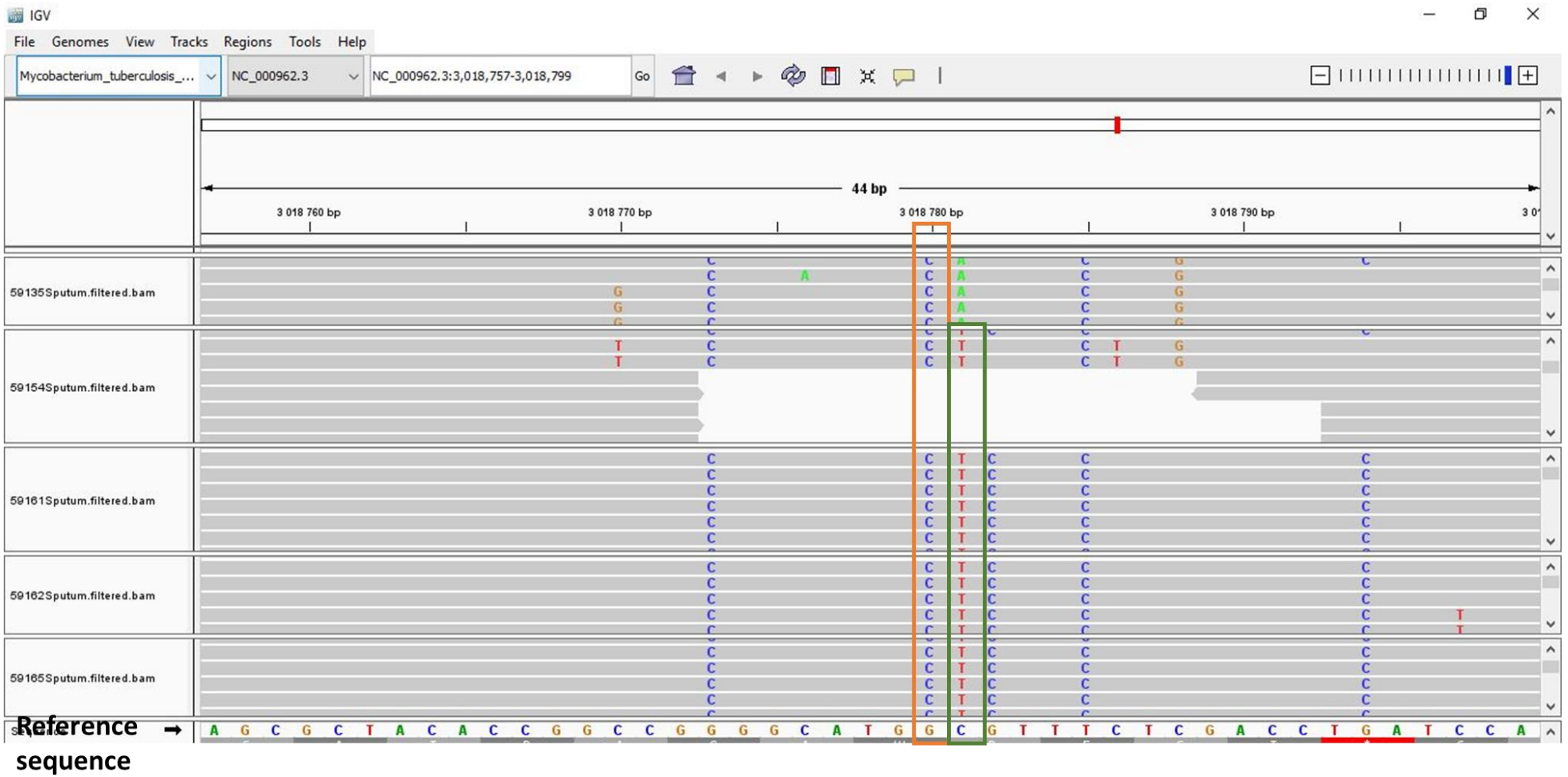


Figure 3: Integrative genome viewer (IGV) of sequencing reads from the uncultured sputum samples with mutations in the *sigA* gene. The reference sequence is located at the bottom. These mutations occurring at position 3018 780 and 3018 781 in the *M. tb* genome were common amongst the uncultured sputum samples. The G » C mutation is selected in the orange box and the C » T mutation is selected in the orange box.

Appendix D: Supplementary Tables

Table D1- Spectrophotometric assessment of extracted DNA from each *M. tuberculosis* sample

Patient ID	A260	A280	A260/280	Concentration (ng/µL)
59135 Sputum	0.229	0.122	1.87	11.47
59143 Sputum	1.240	0.662	1.87	62.01
59144 Sputum	1.402	0.739	1.90	70.11
59152 Sputum	1.520	0.803	1.89	76.00
59154 Sputum	1.223	0.659	1.47	61.15
59161 Sputum	0.195	0.123	1.59	9.77
59162 Sputum	0.961	0.509	1.89	48.04
59164 Sputum	0.058	0.035	1.67	2.90
59165 Sputum	0.066	0.035	1.86	3.28
59172 Sputum	9.324	4.864	1.92	466.19
59135 CFU	0.980	0.557	1.76	48.99
59143 CFU	5.682	2.874	1.98	284.12
59144 CFU	2.240	1.126	1.99	111.98
59152 CFU	5.463	2.743	1.99	273.17
59154 CFU	1.710	0.865	1.99	85.50
59164 CFU	1.448	0.724	2.00	72.41
59172 CFU	1.193	0.601	1.98	59.66
59135 MPN-M1	5.186	2.618	1.98	259.30
59143 MPN-M1	9.284	4.723	1.97	464.18
59144 MPN-M1	5.263	2.698	1.95	263.14
59152 MPN-M1	6.963	3.570	1.95	348.15
59154 MPN-M1	2.230	1.162	1.92	111.50
59161 MPN-M1	1.716	0.891	1.93	85.79
59162 MPN-M1	2.389	1.091	2.19	119.44
59164 MPN-M1	0.958	0.520	1.84	47.91
59165 MPN-M1	2.188	1.155	1.89	109.39
59172 MPN-M1	5.301	2.687	1.97	265.03
59135 MPN-M2	5.405	2.789	1.94	270.24
59143 MPN-M2	1.471	0.751	1.96	73.56
59144 MPN-M2	4.896	2.555	1.92	244.81
59152 MPN-M2	3.300	2.559	1.29	164.99
59154 MPN-M2	3.930	2.186	1.80	196.50
59161 MPN-M2	0.503	0.299	1.68	25.14
59162 MPN-M2	1.379	0.829	1.66	68.95
59164 MPN-M2	5.779	3.005	1.92	288.96
59165 MPN-M2	3.342	1.795	1.86	167.09
59172 MPN-M2	2.236	1.155	1.94	111.79
59135 MPN-RV1	1.571	0.839	1.87	78.56
59143 MPN-RV1	2.894	1.718	1.68	144.70
59144 MPN-RV1	3.704	2.363	1.57	185.18
59152 MPN-RV1	2.684	1.465	1.83	134.99
59154 MPN-RV1	1.509	0.828	1.82	75.42
59161 MPN-RV1	2.022	1.120	1.81	101.08
59162 MPN-RV1	7.248	3.776	1.92	362.42

59164 MPN-RV1	1.604	1.053	1.52	80.21
59165 MPN-RV1	21.470	12.487	1.82	173.48
59172 MPN-RV1	3.392	1.732	1.96	169.58
59135 MPN-RV2	7.632	6.076	1.26	381.62
59143 MPN-RV2	0.999	0.666	1.50	49.96
59144 MPN-RV2	3.354	1.679	2.00	167.69
59152 MPN-RV2	7.611	3.954	1.92	380.55
59154 MPN-RV2	0.088	0.086	1.02	4.42
59161 MPN-RV2	2.889	1.572	1.84	144.44
59162 MPN-RV2	5.191	3.136	1.66	259.53
59164 MPN-RV2	5.208	3.165	1.65	260.40
59165 MPN-RV2	5.006	3.049	1.64	250.32
59172 MPN-RV2	0.814	0.486	1.68	40.71
59135 MGIT-M	5.209	2.697	1.93	260.45
59143 MGIT-M	0.645	0.368	1.75	32.26
59144 MGIT-M	0.900	0.508	1.77	45.02
59152 MGIT-M	0.913	0.429	1.85	45.63
59154 MGIT-M	0.709	0.423	1.67	35.45
59161 MGIT-M	2.960	1.878	1.58	147.99
59162 MGIT-M	1.166	0.681	1.71	58.29
59164 MGIT-M	1.908	1.019	1.87	95.41
59165 MGIT-M	1.289	0.726	1.88	64.47
59172 MGIT-M	0.425	0.259	1.64	21.25
59135 MGIT-CF	2.976	1.581	1.88	148.80
59143 MGIT-CF	13.020	6.557	1.99	651.01
59144 MGIT-CF	1.758	0.973	1.81	87.90
59152 MGIT-CF	0.215	0.126	1.70	10.74
59154 MGIT-CF	0.317	0.177	1.79	15.83
59161 MGIT-CF	2.877	1.680	1.71	143.87
59162 MGIT-CF	0.918	0.564	1.63	45.88
59164 MGIT-CF	2.288	1.199	1.91	114.41
59165 MGIT-CF	2.355	1263	1.86	117.76
59172 MGIT-CF	0.327	0.174	1.88	16.34
59135 MGIT-RPFB	1.082	0.594	1.82	54.09
59143 MGIT- RPFB	7.406	3.920	1.89	370.30
59144 MGIT- RPFB	0.469	0.259	1.81	23.47
59152 MGIT- RPFB	1.658	0.886	1.87	82.89
59154 MGIT- RPFB	0.501	0.282	1.78	25.07
59161 MGIT- RPFB	6.503	3.385	1.92	325.16
59162 MGIT- RPFB	2.542	1.409	1.80	127.08
59164 MGIT- RPFB	2.233	1.199	1.86	111.63
59165 MGIT-RPFB	2.049	1.116	1.84	102.44
59172 MGIT- RPFB	0.385	0.219	1.79	19.25
59135 MGIT-RPFB/RIPA	0.611	0.317	1.93	30.53
59143 MGIT-RPFB/RIPA	0.650	0.340	1.91	32.52
59144 MGIT-RPFB/RIPA	0.471	0.240	1.96	23.53

59154 MGIT-RPFB/RIPA	0.446	0.253	1.76	22.28
59161 MGIT-RPFB/RIPA	2.810	1.543	1.82	140.48
59162 MGIT-RPFB/RIPA	1.411	0.783	1.80	70.54
59164 MGIT-RPFB/RIPA	0.316	0.159	1.99	15.81
59165 MGIT-RPFB/RIPA	0.518	0.301	1.93	29.07
59172 MGIT-RPFB/RIPA	0.162	0.084	1.92	8.11

Abbreviations: M1= sputum grown in media from the first well of the MPN assay, M2 = sputum grown in media from the second well of the MPN assay, RV= sputum grown with culture filtrate, RV1= Culture from the first well of the MPN assay supplement with culture filtrate, RV2 = Culture from the second well of the MPN assay supplemented with culture filtrate, RPFB= sputum grown in media supplemented with resuscitation promoting factor B, RPFB/RIPA= sputum grown in media supplemented with the recombinant resuscitation promoting factor B/ Resuscitation *interacting protein*. A260 measures the absorbance of DNA at 260 nm. A280cmeasures the absorbance of the DNA at 280 nm. A260/A280 is a ratio of the absorption of DNA at 260nm and at 280 nm for the assessment of DNA purity.

Table D2- Experimental and SpolPred predicted spoligotypes for 93 *M. tuberculosis* clinical samples Spoligotype mismatches between silico predicted and experimental determined spoligotypes are highlighted in red. Samples shaded in pink represent a one-digit discrepancy between the in silico predicted and experimentally predicted octal code.

	Experimental results		SpolPred results	
Patient ID	**Octal code**	**Spoligotype**	**Octal code**	**Spoligotype**
59135 Sputum	777777777760771	T1	Not determined	
59135 CFU	777777477760771	H37Rv	777777477760771	H37Rv
59135 MPN M1	777777477760771	H37Rv	777777477760771	H37Rv
59135 MPN M2	777777477760771	H37Rv	777777477760771	H37Rv
59135 MPN RV1	777777777760771	T1	777777777760771	T1
59135 MPN RV2	777777777760771	T1	777777777760771	T1
59135 MGIT M	777777777760700	T2	777777777760771	T1
59135 MGIT CF	777777777760700	T2	777777777760700	T2
59135 MGIT RPFB	776177407760771	LAM 3	776177407760771	LAM 3
59135 MGIT RIPA/RPFB	777777777760700	T2	777777777760700	T2
59143 Sputum	776377777760771	S	Not determined	
59143 CFU	776377777760771	S	776377777760771	S
59143 MPN M1	776377777760771	S	776377777760771	S
59143 MPN M2	776377777760771	S	Not determined	
59143 MPN RV1	776377777760771	S	776377777760771	S
59143 MPN RV2	776377777760771	S	776377777760771	S
59143 MGIT M	776377777760771	S	776377777760771	S
59143 MGIT RIPA/RPFB	776377777760771	S	776377777760771	S
59144 Sputum	776177607760771	LAM 3	000000000000011	Beijing

59144 CFU	776177607760771	LAM 3	776177607760771	LAM 3
59144 MPN M1	776177607760771	LAM 3	776177607760771	LAM 3
59144 MPN M2	776177607763771	Family 33	200000000000020	Family 35
59144 MPN RV1	776177607760771	LAM 3	776177607760771	LAM 3
59144 MPN RV2	776177607760771	LAM 3	776177607760771	LAM 3
59144 MGIT M	776177607760771	LAM 3	776177607760771	LAM 3
59144 MGIT RPFB	776177607760771	LAM 3	776177607760771	LAM 3
59144 MGIT RIPA/RPFB	776177607760771	LAM 3	776177607760771	LAM 3
59152 Sputum	777777607700771	LAM 9	100001002000040	LAM 7
59152 CFU	777777607700771	LAM 9	777777607700771	LAM 9
59152 MPN M1	777777607700771	LAM 9	777777607700771	LAM 9
59152 MPN M2	777777607700771	LAM 9	777777607700771	LAM 9
59152 MPN RV1	777777607700771	LAM 9	777777607700771	LAM 9
59152 MPN RV2	777777607700771	LAM 9	777777607700771	LAM 9
59152 MGIT M	777777607700771	LAM 9	777777607700771	LAM 9
59152 MGIT CF	777777607700771	LAM 9	777777607700771	LAM 9
59152 MGIT RPFB	776377777760771	S	776377777760771	S
59154 Sputum	777777677763771	Family 33	Not determined	
59154 CFU	703377400001771	CAS	703377400001771	CAS
59154 MPN M1	702777740003771	CAS	702777740003771	CAS
59154 MPN M2	703377400001771	CAS	703377400001771	CAS
59154 MPN RV1	703377400001771	CAS	703377400001771	CAS
59154 MPN RV2	703377400001771	CAS	703377400001771	CAS
59154 MGIT M	703377400001771	CAS	703377400001771	CAS
59154 MGIT CF	703377400001771	CAS	703377400001771	CAS
59154 MGIT RPFB	703377400001771	CAS	703377400001771	CAS
59154 MGIT RIPA/RPFB	703377400001771	CAS	703377400001771	CAS
59161 Sputum	777777607503771	Family 33	Not determined	
59161 MPN M1	000000000003771	Beijing	000000000003771	Beijing
59161 MPN M2	000000000003771	Beijing	Not determined	
59161 MPN RV1	000000000003771	Beijing	000000000003771	Beijing
59161 MPN RV2	000000000003771	Beijing	000000000003771	Beijing
59161 MGIT M	000000000003771	Beijing	000000000003771	Beijing
59161 MGIT CF	000000000003771	Beijing	000000000003771	Beijing
59161 MGIT RPFB	777777777760700	T2	777777777760700	T2
59161 MGIT RIPA/RPFB	000000000003771	Beijing	000000000003771	Beijing
59162 Sputum	777777677763771	Family 33	Not determined	
59162 MPN M1	703377400001771	LAM 3	703377400001771	LAM 3
59162 MPN M2	703377400001771	LAM 3	703377400001771	LAM 3
59162 MPN RV1	703377400001771	LAM 3	703377400001771	LAM 3
59162 MPN RV2	703377400001771	LAM 3	703377400001771	LAM 3
59162 MGIT M	703377400001771	LAM 3	703377400001771	LAM 3
59162 MGIT CF	703377400001771	LAM 3	703377400001771	LAM 3
59162 MGIT RPFB	000000000003771	Beijing	000000000003771	Beijing

59162 MGIT RIPA/RPFB	000000000003771	LAM 3	000000001000000	*M. microti*
59164 Sputum	000000000003771	Beijing	000000000003771	Beijing
59164 CFU	000000000003771	Beijing	000000000003771	Beijing
59164 MPN M1	000000000003771	Beijing	000000000003771	Beijing
59164 MPN M2	777777604063731	Beijing	777777604063731	LAM 8
59164 MPN RV1	776177607761771	Lam	776177607761771	LAM 3
59164 MPN RV2	000000000003771	LAM	000000000003771	Beijing
59164 MGIT M	000000000003771	Beijing	000000000003771	Beijing
59164 MGIT CF	000000000003771	Beijing	000000000003771	Beijing
59164 MGIT RPFB	000000000003771	Beijing	000000000003771	Beijing
59164 MGIT RIPA/RPFB	000000000003771	Beijing	000000000003771	Beijing
59165 Sputum	776177607763771	Family 33	Not determined	
59165 MPN M1	776177607763771	Family 33	Not determined	
59165 MPN M2	000000000003771	Beijing	000000000003771	Beijing
59165 MPN RV1	776177607763771	Family 33	Not determined	
59165 MPN RV2	000000000003771	Beijing	000000000003771	Beijing
59165 MGIT M	000000000003771	Beijing	000000000003771	Beijing
59165 MGIT CF	777777477760771	H37Rv	777777477760771	H37Rv
59165 MGIT RPFB	000000000003771	Beijing	000000000003771	Beijing
59165 MGIT RIPA/RPFB	000000000003771	Beijing	000000000003771	Beijing
59172 Sputum	703377400001771	CAS	Not determined	
59172 CFU	703377400001771	CAS	7037777400003771	CAS
59172 MPN M1	703377400001771	CAS	7037777400003771	CAS
59172 MPN M2	776377777760771	S	776377777760771	S
59172 MPN RV1	703377400001771	CAS	7037777400003771	CAS
59172 MPN RV2	703377400001771	CAS	7033777400003771	CAS
59172 MGIT M	703377400001771	CAS	7033777400003771	CAS
59172 MGIT CF	703377400001771	CAS	7033777400003771	CAS
59172 MGIT RPFB	777777774020771	Haarlem 1	777777774020771	Haarlem 1
59172 MGIT RIPA/RPFB	703377400001771	CAS	7037777400003771	CAS

Table D3- The number of SNPs, insertions and deletions identified from the WGS data of each *M. tb* sample

Patient ID	SNPs	INSERTIONS	DELETIONS
59135 Sputum	998	1	0
59135 MPN_RV1	24	6	6
59135 MPN_RV2	29	8	6
59135 MGIT_M	501	22	23
59135 MGIT_CF	510	20	25
59135 MGIT_RPFB	692	28	27
59135 MGIT_RIPA/RPFB	503	24	24
59143 Sputum	265	1	0
59143 CFU	665	25	24
59143 MPN_M1	678	28	19
59143 MPN_RV1	662	27	20
59143 MPN_RV2	661	29	21
59143 MGIT_M	664	32	22
59144 Sputum	62 731	20	20
59144 CFU	692	28	26
59144 MPN_M1	695	23	24
59144 MPN_RV1	775	29	30
59144 MPN_RV2	688	27	27
59144 MGIT_M	691	29	28
59144 MGIT_RPFB	696	27	30
59144 MGIT_RIPA/RPFB	690	23	25
59152 Sputum	372	19	11
59152 CFU	705	28	20
59152 MPN_M1	701	28	21
59152 MPN_M2	701	29	21
59152 MPN_RV1	705	26	22
59152 MPN_RV2	707	30	23
59152 MGIT_M	704	29	26
59152 MGIT_CF	704	32	28
59152 MGIT_RPFB	663	34	25
59154 Sputum	1360	0	0
59154 CFU	1177	55	48
59154 MPN_M1	1177	52	50
59154 MPN_M2	1180	57	53
59154 MPN_RV1	569	22	19
59154 MPN_RV2	1175	46	40
59154 MGIT_M	1179	55	51
59154 MGIT_CF	1177	55	53
59154 MGIT_RPFB	1179	56	51

59154 MGIT-RIPA/RPFB	1177	49	50
59161 Sputum	1255	2	0
59161 MPN-M1	1195	54	49
59161 MPN-RV1	1203	51	54
59161 MPN- RV2	1198	48	52
59161 MGIT-C	1195	45	50
59161 MGIT-B	499	20	24
59161 MGIT-AB	706	28	18
59162 Sputum	400	0	0
59162 MPN-M1	691	22	25
59162 MPN-M2	694	20	24
59162 MPN-RV2	693	23	27
59162 MGIT-M	692	23	25
59162 MGIT-C	1202	53	49
59162 MGIT-B	688	24	24
59164 Sputum	1205	47	54
59164 CFU	1220	55	58
59164 MPN_M1	1208	50	55
59164 MPN_M2	686	30	31
59164 MPN_RV1	696	26	28
59164 MPN_RV2	1200	45	54
59164 MGIT_M	1209	55	65
59164 MGIT-CF	1215	48	60
59164 MGIT-RPFB	663	48	55
59164 MGIT-RIPA/RPFB	1213	50	60
59165 Sputum	1220	3	0
59165 MPN-M2	1201	46	56
59165 MGIT-RV2	1208	52	52
59165 MGIT-M	1209	55	56
59165 MGIT-B	1214	52	52
59165 MGIT-AB	1207	51	51
59172 Sputum	121	0	0
59172 CFU	1185	49	58
59172 MPN_M1	675	25	24
59172 MPN-M2	679	30	21
59172 MPN_RV1	689	23	24
59172 MPN_RV2	1184	39	44
59172 MGIT-RPFB	779	39	35
59172 MGIT-RIPA/RPFB	1189	48	50

ORIGINALITY REPORT

12%	10%	7%	%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	wiredspace.wits.ac.za Internet Source	1%
2	hdl.handle.net Internet Source	1%
3	espace.library.uq.edu.au Internet Source	1%
4	etheses.bham.ac.uk Internet Source	<1%
5	bmcgenomics.biomedcentral.com Internet Source	<1%
6	open.uct.ac.za Internet Source	<1%
7	www.science.gov Internet Source	<1%
8	www.freepatentsonline.com Internet Source	<1%
9	core.ac.uk Internet Source	<1%