USING *MYCOBACTERIUM TUBERCULOSIS* COMPLEX WHOLE GENOMES AND BLOOD BIOMARKERS TO STUDY TUBERCULOSIS TRANSMISSION

by

SAMUEL KIRIMUNDA

(Under the Direction of Christopher Whalen)

ABSTRACT

Background: There were over 10.0 million cases of tuberculosis (TB) in the world in 2018, a number that has been relatively stable in recent years. The burden of the disease varies enormously among countries per year with most cases occurring in the WHO regions of South-East Asia (44%), Africa (24%), and the Western Pacific (18%). Based on the epidemic theory, epidemics continue to occur when one index case is replaced by one or more cases. This study addresses the problem of ongoing transmission of Mycobacterium tuberculosis (Mtb) in Sub-Saharan Africa by exploiting Mtb whole genome sequences, drug resistance candidate gene mutations, and interferon-gamma (IFN-γ) cytokine as biomarkers.

Methods: A cross-sectional and longitudinal study conducted in Rubaga division in Kampala Uganda were the data sources for this study. The time cough symptoms started, social network distances, drug resistance candidate gene mutations, and genetic distances among TB cases were analyzed to determine the transmission tree, reproduction numbers, identifiability score, and factors associated with being a source of infections. Concentrations of Mtb-specific antigen-induced IFN-γ cytokine in blood were quantified in Mtb uninfected, recently infected and remotely infected individuals.

Results: A total of 15 transmission clusters with the largest cluster comprising of 12 members and the smallest cluster of 2 members were identified. There were 36 (58.1%) individuals who were identified as potential sources of infection. MTBC genomes were classified as 78 % (62/79) Lineage four (L4), 18 % (14/79) L2 and 4 % (3/79) L3. Based on a certainty score of 20%, 5 transmission clusters were identified. Ancestor genomes were inversely associated with mutations in the rrs and rrl genes and positively associated with mutations in gyrA, ribD and ethR genes. Lastly, mean Interferon-γ blood levels in recently infected individuals were intermediate between uninfected and those with established infection and accurately predicted 69% of recently infected individuals at an optimal cutoff value of 2.58 IU/ml.

Conclusion: An identifiability score, MTBC lineage four strain-specific drug resistance candidate gene mutations, and interferon-γ can sufficiently identify individuals who are sources of TB infection or persons recently infected with Mtb in the community.

INDEX WORDS: tuberculosis, transmission, whole genome sequencing, biomarkers, gene mutations, Uganda

USING *MYCOBACTERIUM TUBERCULOSIS* COMPLEX WHOLE GENOMES AND BLOOD BIOMARKERS TO STUDY TUBERCULOSIS TRANSMISSION

by

SAMUEL KIRIMUNDA

BSc., Makerere University, Uganda, 2008

MSc., Makerere University, Uganda, 2016

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2020

USING *MYCOBACTERIUM TUBERCULOSIS* COMPLEX WHOLE GENOMES AND BIOMARKERS TO STUDY TUBERCULOSIS TRANSMISSION

by

SAMUEL KIRIMUNDA

Major Professor:	Christopher Whalen
Committee:	Changwei Li
	Juliet Sekandi
	Fredrick Quinn

Electronic Version Approved:

Ron Walcott
Interim Dean of the Graduate School
The University of Georgia
May 2020

ACKNOWLEDGEMENTS

I wish to particularly acknowledge Professors Moses Joloba and Christopher Whalen for their mentorship and guidance on my academic journey. My Doctoral training was supported by the Fogarty International Center of the National Institutes of Health under Award Number D43TW010045.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF EQUATIONS

Chapter 1

INTRODUCTION

BACKGROUND

Tuberculosis (TB), caused by infection with *Mycobacterium tuberculosis (Mtb*), is the leading cause of death caused by a single pathogen globally. Inhalation of aerosols containing *Mtb* Complex (MTBC) bacilli by susceptible hosts is the main route of transmission of tuberculosis (Tom A Yates et al., 2016). There were over 10.0 million cases of tuberculosis (TB) in the world in 2018, a number that has been relatively stable in recent years. The burden of disease varies enormously among countries, from fewer than 5 to more than 500 new cases per 100,000 population per year (WHO, 2019). Approximately 2 billion people are latently infected with *M. tuberculosis* (Jajou et al., 2018) but only about 5 to 10% will develop TB disease in their life. Since 2015, the World health Organization created the 'End TB' global tuberculosis strategy framework setting ambitious milestones targeting a 50% reduction in tuberculosis incidence rate by 2025 and a 90% reduction by 2035 (World Health Organization, 2015b). Unfortunately, the annual reduction of TB incidence is 1.5%, lower than the 4-5% per year approximated as needed to achieve the 'End TB' targets (WHO 2016).

Epidemics of tuberculosis continue to occur when prevalent cases are replaced by incident cases (Whalen, 2016). Individuals who are latently infected with *Mtb* compromise a primary source of future TB cases, and their identification and treatment is important for intervention against the TB epidemic (Suliman et al., 2018; Whalen, 2016). In an attempt to curb the epidemic, various studies have attempted to characterize community transmission of tuberculosis using sequence data and have shown that some individuals infect a large number of

others and are thus ‘Super-spreaders’, while many infectious individuals do not transmit (Escombe et al., 2008; Kline, Hedemark, & Davies, 1995; RILEY et al., 1962; Snider, Kelly, Cauthen, Thompson, & Kilburn, 1985; van Geuns, Meijer, & Styblo, 1975; T. A. Yates, Tanser, & Abubakar, 2016; Ypma, Altes, Van Soolingen, Wallinga, & Van Ballegooijen, 2013).

Whole Genome Sequencing (WGS) of pathogens has become an essential tool for improving understanding of how tuberculosis is spread (Hatherell et al., 2016). Typically, samples are taken from patients in the field, the date of symptoms start and other epidemiological data are recorded, and the MTBC genome is sequenced. The phylogeny derived from the sequence data is used to infer likely transmission events (Stimson et al., 2019). Lately, these approaches mainly use MTBC phylogenetic clustering to signify on-going transmission (recent transmission) or the presence of unique strains as indicators of reactivation of latent TB infections to disease (Asiimwe et al., 2009; Gardy et al., 2011a; Weis et al., 2002).

Developing a new diagnostic assay or deploying the current tests to identify recent *Mtb* infection would allow for targeted treatment of those persons most likely to progress to active TB and is a priority among international TB agencies (Pai & Schito, 2015). To address this challenge, new techniques, including transcript microarrays, flow cytometry of intracellular cytokines, and multiplex micro bead-based immunoassay (Luminex assay) of cytokines, have recently been introduced and used to study TB diagnostics (Berry et al., 2010; Caccamo et al., 2010; Chegou, Black, Kidd, van Helden, & Walzl, 2009; Sutherland, de Jong, Jeffries, Adetifa, & Ota, 2010). Unfortunately, majority of these approaches were conducted in cohorts of households or close contacts of TB cases. Whereas the household is an environment of intense transmission of *Mtb*, it does not account for majority of new transmission of tuberculosis yet it

remains the focus of most TB transmission studies in the current literature (Crampin et al., 2006; Glynn et al., 2015; Martinez et al., 2017a; Verver et al., 2004; Whalen et al., 2011).

The main goal of this dissertation was to advance the current understanding of tuberculosis transmission by improving methods that are used to identify individuals newly infected or reactivated with *Mtb*. By using WGS and blood biomarkers, this work can inform policy measures to prevent tuberculosis transmission and identify recently infected individuals are most likely to develop clinical TB disease.

SPECIFIC AIMS

Aim 1: To adapt an outbreak investigation model to ascertain who infected whom in a TB endemic sub-Saharan African population.

Rationale: As a way to determine who infected whom in tuberculosis endemic settings, methodologies that use more than one study population characteristic have been proposed before (Auld et al., 2018; Guthrie et al., 2018; Martinez et al., 2017b). More advanced methods have been designed and implemented in tuberculosis outbreak investigation using WGS and epidemiologic data with varying emphasis on either data source (Gardy et al., 2011a; Walker et al., 2013b). However, in instances of both endemic and outbreak tuberculosis transmission, there is a shortage of methodologies able to concurrently utilize multiple sources of data to infer transmission (Teunis et al., 2013a). By assessing social distances among household and community contacts of index cases and controls, we investigated which case infected whom. To achieve this goal, we used the structure of the social network, WGS and serial interval estimates from the study population.

Aim 2: To study whether mutations in drug resistance candidate genes are associated with tuberculosis transmission in peri-urban Kampala, Uganda.

Rationale: In the past, traditional genotyping methods were not sufficient to distinguish MTBC strains in populations with outbreaks of tuberculosis (Gardy et al., 2011b; Zappala et al., 2018). However, the application of WGS technologies has advanced resistance prediction, outbreak detection, and genomic surveillance of MTBC (Merker et al., 2018). Drug resistance is currently considered a global emergence (World Health Organization, n.d.), and its currently a research field that has benefited from intensive studies aimed at learning important insights into drug resistant tuberculosis spread at a global (Mutreja et al., 2011), continental (Eldholm et al., 2015) and local scale (Bainomugisa et al., 2018).

In Aim 2, we used the TB profiler pipeline (Coll et al., 2015) to determine candidate gene mutations, genotypic drug resistance profiles and phylogenetic classification for all study participant whole genome sequences regardless of phenotypic drug resistance status. The TB profiler pipeline searches for small variants and big deletions associated with drug resistance in genes or variations on a gene that may relate to drug resistance as defined by a distinct biological pathway or findings from previous studies (candidate gene). We examined whether drug resistance candidate gene mutations were association with being an ancestor genome among transmission clusters in residents of peri-urban Kampala, Uganda. We hypothesized that there were drug resistance candidate gene mutations that can serve as potential markers of being an ancestor among transmission clusters within a pairwise SNP difference of less or equal to12.

Aim 3: To test whether interferon gamma cytokine concentration levels can identify recent infection with *Mycobacterium tuberculosis* in a community setting.

Rationale: Developing a new diagnostic assay or deploying the current tests so as to identify recent *Mtb* infection would allow for targeted treatment of those persons most likely to progress to active TB and is a priority among international TB agencies (Pai & Schito, 2015). To

address this challenge, new techniques, including transcript microarrays, flow cytometry of intracellular cytokines, and multiplex micro bead-based immunoassay (Luminex assay) of cytokines, have recently been introduced and used to study TB diagnostics (Berry et al., 2010; Caccamo et al., 2010; Chegou et al., 2009; Sutherland et al., 2010). Unfortunately, majority of these studies were conducted in cohorts of household or close contacts of TB cases. Whereas, the household is an environment of intense transmission of *Mtb*, it does not account for majority of new transmission of tuberculosis yet it remains the focus of most TB transmission studies in the current literature (Crampin et al., 2006; Glynn et al., 2015; Martinez et al., 2017a; Verver et al., 2004; Whalen et al., 2011).

In Aim 3, we assessed data from a community-based cohort of Ugandans who were *Mtb* uninfected at baseline and followed for 1-2 years until TST conversion (*Mtb* infection). We compared blood concentrations of *Mtb*-specific antigen induced interferon gamma (IFN-γ) to differentiate recently infected (TST converted) individuals from remotely infected and uninfected individuals in a community setting beyond the household of a TB case. We also assessed for the optimum cutoff of blood concentrations of *Mtb*-specific antigen induced interferon-γ for differentiating TST negatives from TST positive individuals.

Study conceptual framework: The three aims that formed the studies in this work were conceived in such a way as to impact TB clinical disease outcomes after exposure to *Mycobacterium tuberculosis* germs. Overall, without treatment, about 5 to 10% of infected persons will develop TB disease at some time in their lives (Figure 1). About half of those people who develop clinical TB will do so within the first two years of infection referred to as primary progressive disease and the other half arising after longer latency periods are classified as reactivation disease. Primary infection of TB refers to a state of immune sensitization of one's

body to ongoing or recent infection with *Mtb.* It is marked by presence of humoral immune response and cellular immune response markers. Individuals with an established primary infection usually have a positive tuberculin skin test and are Interferon Gamma Release Assay test positive.

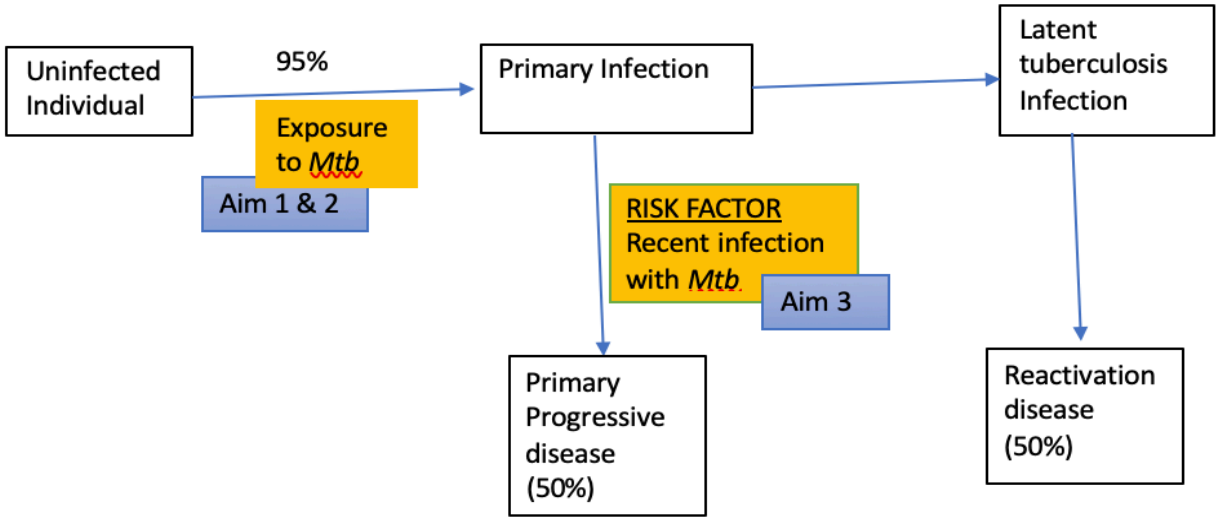


Figure 1: Study Conceptual Framework

Chapter 2

LITERATURE REVIEW

TUBERCULOSIS

Tuberculosis is caused by infection with *Mycobacterium tuberculosis* (Tom A Yates et al., 2016), and is the leading cause of death caused by a single pathogen globally ("WHO | Global tuberculosis report 2017," 2017). When a healthy individual inhales *Mtb* bacilli, the organism is engulfed by alveolar macrophages. Unlike other bacteria, *Mtb* is highly resistant to killing by non-activated macrophages and multiplies quite successfully inside these cells. It is has been shown that in persons with an uncompromised immune system, immunity to the *Mtb* bacilli does develop, and in most cases the invading bacilli are killed, and disease is averted (Smith, 2003). However, in a minority of cases, the host immune responses are not sufficient to eliminate the infection, which then proceeds to multiply in the manner of most pathogens, resulting in progressive primary TB disease or enter latent TB Infection (LTBI) state, in which bacilli can remain viable for decades, and may reactivate to disease in the future (Smith, 2003).

Latently infected individuals experience no adverse health effects and are not known to transmit *Mtb*, but face an ongoing risk of developing active tuberculosis through reactivation (Menzies et al., 2018). Because we cannot presently distinguish individuals who were briefly infected, but eliminated all viable bacilli, from those that continue to harbor LTBI, or accurately tell those who are LTBI infected and will progress to have LTBI reactivation, we rely on rough estimates. There is a high prevalence of latent tuberculosis infection in many settings (Houben & Dodd, 2016) and reactivation represents a substantial proportion of incident tuberculosis cases especially in settings in which transmission has been in sustained decline (Yuen, Kammerer,

Marks, Navin, & France, 2016). The risk of progressing to active disease also varies by individual characteristics, with infants (Marais et al., 2004), individuals with advanced HIV infection (Antonucci, Girardi, Raviglione, & Ippolito, 1995; Selwyn et al., 1989) and individuals with other conditions that heighten their risk for TB disease such as smoking, kidney disease, and diabetes (Bates et al., 2007; Chia, Karim, Elwood, & FitzGerald, 1998; Jeon & Murray, 2008; Lonnroth, Williams, Cegielski, & Dye, 2010) having elevated progression risks. This underscores the fact that LTBI infection is a defining feature of tuberculosis epidemiology and such factors as those affecting its dynamics should be considered in tuberculosis studies and modelling (Menzies et al., 2018).

Tuberculin skin test (TST): TST involves injecting a mixture of undefined *Mtb* antigens under the skin of the forearm. If the immune system of the individual has been sensitized to any of the antigens, a delayed type hypersensitivity response will be observed at the site. It is known that some of the antigens in the skin test mixture may cross-react with antigens of environmental mycobacteria which is thought to be a cause of false positive reactions (Wachtman, Miller, Xia, Curran, & Mansfield, 2011). Immunization with BCG vaccine also causes a reaction to the tuberculin skin test (Farhat, Greenaway, Pai, & Menzies, 2006). Furthermore, it is believed that individuals who have been briefly infected with *Mtb* bacilli but have successfully eliminated the infection through the development of an effective immune response (self-cured) continue to have a reaction to the skin test antigens.

Some people who develop a positive reaction at the site of the skin test are not infected with *Mtb*, and prophylactic treatment of such a group exposes them to the risk of medication side effects while offering no benefit. In patients with active disease, the presence of various symptoms, chest x-rays, and *Mtb* culture allows for an accurate diagnosis. In the case of LTBI,

there is virtually no way to confirm that a positive TST indicates a true infection. Hence, one of the major shortcoming of the present TST is its lack of specificity (Farhat et al., 2006). Lack of specificity implies that the rate of false positive results is high. As the incidence of active or latent TB lowers, an increasing proportion of the observed positive test results will be false positives, as the positive predictive value of the test decline. A poor positive predictive value is in fact a major problem with using the TST to guide prophylactic treatment recommendations in low incidence regions (Berkel, Cobelens, de Vries, Draayer-Jansen, & Borgdorff, 2005).

Cutoffs used for TST conversions are different from the cutoffs used for diagnosis of LTBI ("Targeted Tuberculin Testing and Treatment of Latent Tuberculosis Infection," 2000). Measurement of the long-term ability of a positive TST to predict development of active TB is difficult, requiring prolonged follow-up of unselected populations. Based on historical studies, there is a modest positive association between tuberculin reactivity and the risk of active TB (Watkins, Brennan, & Plant, 2000). However, a majority of individuals with positive TST results do not progress to active disease. As a result, many TST-positive individuals need to be treated in order to prevent one disease event (Landry & Menzies, 2008). Cutoff values represent a statistical attempt to minimize false-positive or false-negative readings and vary according to individual and epidemiologic factors, of which recent exposure to *M. tuberculosis* is the most heavily weighted (Dunn, Starke, & Revell, 2016). For instance, for children at highest risk of infection progressing to disease, an induration diameter of ≥5 mm is classified as a positive result. For other high-risk groups, an induration diameter of ≥10 mm is a positive result. For low-risk children, an induration diameter of ≥15 mm is a positive result (Committee on Infectious Diseases; American Academy of Pediatrics; David W. Kimberlin, MD, FAAP; Michael T. Brady, MD, FAAP; Mary Anne Jackson, n.d.). In work not published yet, our group has shown

that there are two TST optimal cutoffs for diagnosis of LTBI at 7.8 mm and 10 mm (Waldu et al, unpublished).

Interferon Gamma Release Assays (IGRAs): These tests detect cell mediated responses to *Mtb* infection by measuring IFN-γ released in response to antigens specific to *Mtb*. Two commercial IGRAs are available in many countries: the QuantiFERON-TB Gold In-Tube (QFT) assay (Cellestis/Qiagen, Carnegie, Australia) and the T-SPOT.TB assay (Oxford Immunotec, Abingdon, United Kingdom). Both tests are approved by the U.S. Food and Drug Administration (FDA) and Health Canada and are CE (Conformité Européenne) marked for use in Europe. The QFT assay is an enzyme-linked immunosorbent assay (ELISA)-based, whole blood test that uses peptides from the RD1 antigens ESAT-6 and CFP-10 as well as peptides from one additional antigen (TB7.7 [Rv2654c], which is not an RD1 antigen) in an in-tube format (Andersen, Munk, Pollock, & Doherty, 2000a, 2000b; Sorensen, Nagai, Houen, Andersen, & Andersen, 1995).

An individual is considered positive for M. tuberculosis infection if the IFN-γ response to TB antigens is above the test cutoff (after subtracting the background IFN-γ response of the negative control). The T-SPOT.TB assay is an enzyme-linked immunosorbent spot (ELISPOT) assay performed on separated and counted peripheral blood mononuclear cells (PBMCs) that are incubated with ESAT-6 and CFP-10 peptides. The result is reported as the number of IFN-γ producing T cells (spot-forming cells). An individual is considered positive for M. tuberculosis infection if the spot counts in the TB antigen wells exceed a specific threshold relative to the negative control wells. Indeterminate IGRA results can occur due to a low IFN-γ response to the mitogen (Pai et al., 2014).

EPIDEMIOLOGY OF TUBERCULOSIS

Modelling data suggest that a quarter of the world is infected with *Mtb* worldwide equating to approximately 2 billion people (R. M. Houben & Dodd, 2016). The World Health Organization estimated that 10.4 million (range 8.7-12.2 million) new active cases of tuberculosis occurred in 2015 in the world (World Health Organization., 2016). Roughly, 11% of these cases occurred in HIV-infected subjects, primarily impacting sub-Saharan Africa (World Health Organization., 2016). Tuberculosis accounted for an estimated 1.4 million deaths in HIV-uninfected people and 0.4 million deaths in HIV-infected individuals in 2015 (World Health Organization., 2016). Uganda is one of the 30 high tuberculosis/HIV burden countries as described by the World Health Organization (World Health Organization., 2016). The estimated incidence of Tuberculosis in Uganda in 2015 was 202 new cases (95% CI 120-304) per 100,000 population, totaling 79,000 new cases (95% CI 47,000-119,000 cases). More than a third of these cases occurred in HIV-infected subjects (26,000 cases).

Dynamic tuberculosis transmission modeling: There is a global effort to understand why certain individuals progress rapidly to disease and others don't (Houben & Dodd, 2016). Although the current realistic national tuberculosis control program estimations fail to capture key aspects about the disease, studies that forecast future disease trends using dynamic transmission models have come handy (Menzies et al., 2018). There are several classification model types and transition probabilities that have been suggested based on different theories and assumptions.

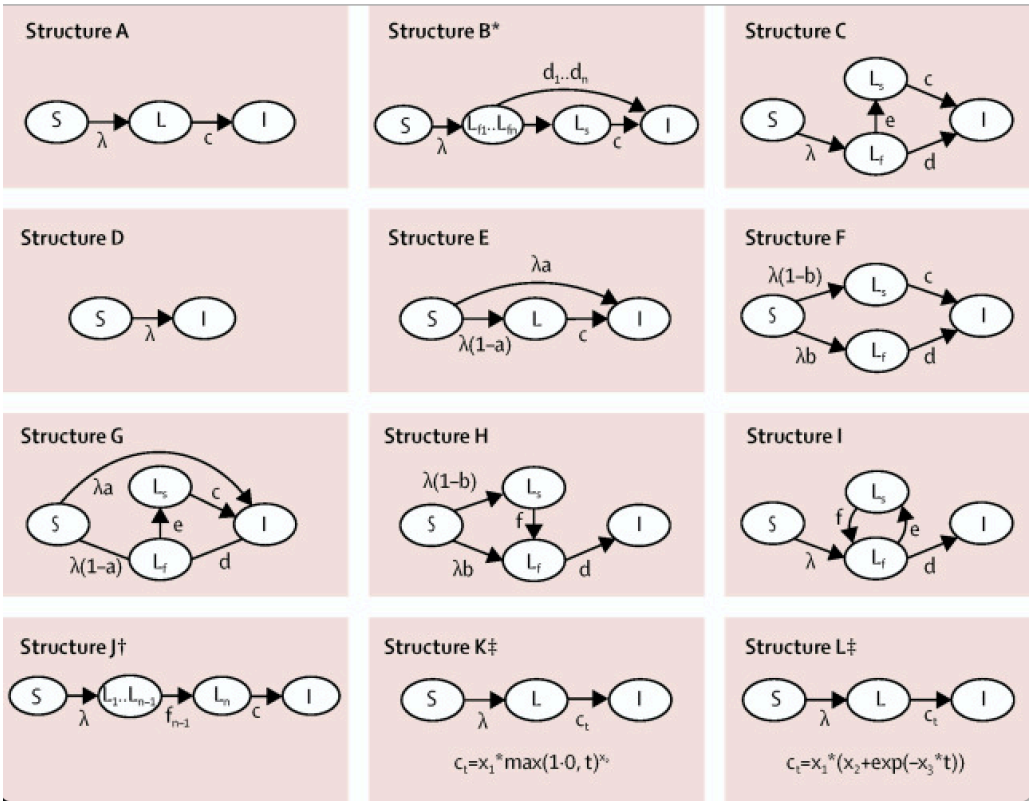


Figure 2 Dynamic model types and transition probabilities.

It has been shown that the model structures and parameter values as described by various studies scarcely reproduced the model predictions for tuberculosis incidence in the years following initial infection. There has also been a substantial disagreement between studies on the rate at which individuals progress to active disease after initial infection. A recent study (Ragonnet et al., 2017) that examined different model structures (Figure 2) found that structure E performed either worst or second worst among the six structures examined (depending on the data fitting method). Structure E performed better than structures A, D, and J, although the root mean squared error was still ten times worse than that of the other structures. The theory of our work was based on model structure E due to its simplicity and biological plausibility. This model allows for progression to primary disease following infection with no latency compartment.

Modeling of tuberculosis outbreak clusters and transmission: A transmission cluster is a group of closely related infections that is usually interpreted as resulting from recent transmission based on sequence relatedness (Poon, 2016). A basic approach to determining relatedness of sequences is by counting the number of single-nucleotide polymorphisms (SNPs) that differ between two sequences. If the SNP difference is fewer than a given threshold, such MTBC genomes are assumed to belong to the same cluster. Many existing methods to identify TB outbreak clusters rely on SNP thresholds, as surveyed recently (Hatherell et al., 2016). However, there is little agreement in the literature as to what a minimum threshold for transmission should be with recent studies adopting lower thresholds ranging from 2 to 50 SNP differences. Unfortunately the number of SNP differences between genomes does not directly imply a probability of recent transmission particularly when applied to different clinical settings (Walker et al., 2013b). Nevertheless, the use of a single SNP threshold is often employed in practice; for example the 12 SNP threshold, used for inferring likely transmission between a pair of TB cases by Public Health England (Walker et al., 2014) amongst others, is perhaps the most common probably because it has been demonstrated in a longitudinal study setting.

It is important to distinguish between the mutation rate, the rate at which spontaneous mutations occur, and the substitution rate, the rate of accumulation of changes in a lineage; this depends on both the mutation rate and the effects of selection and drift (Barrick & Lenski, 2013). In TB, the substitution rate is used for interpreting variants measured with sequencing technologies and it is important given the selection pressure due to antibiotics which can be substantial (Stimson et al., 2019). The background SNP accumulation rate for *Mtb* has been estimated at 0.5 SNPs/genome/year (Walker et al., 2013a). In order to identify transmission clusters of tuberculosis, Bayesian modeling approaches that use epidemiology data such as

sample collection dates in addition to the SNP threshold have been proposed. These include software packages currently implemented in R such as: Transcluster (Stimson et al., 2019), Transphylo, Outbreaker, Phybreak, Phyloscanner, bitrugs, TTsampler, BEAST among others (G. G. R. Murray et al., 2016). The key limitation of these methods is their reliance on bioinformatics expertise which is currently lacking in regions experiencing the highest burden of disease.

Tuberculosis social network analysis: Social Network Analysis (SNA) is an innovative approach to the collection and analysis of human behavior and *Mtb* transmission data (Cook et al., 2007). SNA has been used retrospectively to characterize *M. tuberculosis* outbreaks and highlights the importance of places of social aggregation in sustaining transmission (Barnes et al., 1997; Fitzpatrick et al., 2001; Klovdahl et al., 2001; McElroy et al., 2003; Sterling et al., 2000). The use of SNA methods to collect and interpret contact tracing data in order to determine whether important transmission patterns not otherwise detected by routine contact investigation would emerge was first demonstrated by Cook and colleagues in 2007. In their work, correlation between TST positive status and dense subgroup occurrence supported the value of collecting place data to help prioritize TB contact investigations (Cook et al., 2007). In addition, studies have shown that SNA alone cannot fully inform us of transmission events. Studies such as those done in Vancouver and Greenland incorporated SNA and microbial network data to confirm transmission (Gardy et al., 2011a). In these studies, microbial networks based on pre-WGS molecular genotyping methods alone were unable to pin point an index case in a network until WGS was used (Bjorn-Mortensen et al., 2016). It has been suggested that there is an appreciable increase in the ability to confirm transmission events when more data points are incorporated in

models (Teunis et al., 2013a). The methods of this work are based on WGS techniques, the most sensitive and specific platform currently used in the field of molecular epidemiology.

Whole Genome Sequencing and its role in epidemiology: WGS has become an essential tool for public health surveillance and molecular epidemiology of infectious diseases and antimicrobial resistance. At individual level, it has been used in relapses and reinfection differentiation (relapse means bad treatment and reinfection means bad disease control) (Revez et al., 2017). WGS has been used to investigate disease outbreak such as TB outbreak investigation in Vancouver and Greenland (Bjorn-Mortensen et al., 2016; Gardy et al., 2011a). An outbreak occurs when disease originating from a few index cases spreads fast in a very short time (Bennett, 2008).

Phylogenetic trees: A phylogenetic tree is a graphical summary of the ancestral relationships between organisms. Phylogenetic trees are links from individuals, to populations, to species, to all biological diversity. Phylogenies have been used before in public health particularly in HIV investigations (Ou et al., 1992). A clade of samples on a phylogenetic tree are interpreted as monophyletic when they share a common ancestor. Interpretation of trees require training and are continuously miss interpreted by readers through mistakes such as looking along the tips to perceive relationships, counting nodes, and erroneously perceiving notions of relationships that do not exist. These can be avoided by relying on methods such as sign-post and grouping to correctly interpret relationships (Woese, 2000). Trees can also be rooted or unrooted based on a particular ancestor from which relationships to another member are built on. Based on the number of tree tips, there are multiple rooted trees which can emerge which creates a level of uncertainty. When a cladogram which is a display of the relationships

includes time, it is referred to as a phylogram. The time difference between two samples is the sum of the two-branch length from the tips to the shared node (Woese, 2000).

Branch length depend of the evolutionary model and gives the measure of the amount of evolution that has happened at a branch. In the case of maximum likelihood tree, they represent the expected or average number of substitutions along the branch. We sum branch length to measure the distance between pairs of nodes. Branch length are obtained by pairwise SNP counting of the difference between adjacent genomes. The number of substitution events are estimated through parsimony and Maximum likelihood/Bayesian model (Sober & Steel, 2004). This process occurs in sequences that are finite which leads to mutation saturation and it means that multiple hits occur at the same site. This implies that observed pairwise differences are likely an underestimate of the total divergence.

Most evolution models are based on substitution models and common parameters factored in these models include; the Base frequencies in a given gene, transition/transversion ratio, gamma distributed among site rate heterogeneity, and proportion of invariant sites.
The substitution models include the following (Sober & Steel, 2004):

a) Juke- Canto 1969 (JC69): This model assumes all bases have similar frequencies of 0.25 with a single substitution rate.
b) Kimura 1980 or Kimular-2-Parameter (K80): This model assumes all bases frequencies are the same at 0.25. However, it allows for unequal transition/transversion ratio.
c) Felsenstein 1981 (F81): This model assumes that bases have different frequencies.
d) Hesegawa, Kishino and Yano 1985 (HKY85): This model assumes that base frequencies are all different, allows for unequal transition/transversion ratio.

e) Generalized Time Reversible (GTR): This model assumes that bases have different frequencies and allows for individual substitution rates.

Molecular dating: This is the process of dating the emergence of clones before introduction to a novel environments. It is a process of placing at invent in history along a time stamped phylogenetic tree (Eppinger et al., 2014). Specifically, it describes the speed in units of substitution/site/time. This is conceptually the distance from the tip to root of a phylogenetic tree. This implies that temporal sampling is vital for studying evolution and when samples lack temporal sampling, it deters knowledge of molecular time and no clock signature is observed. A strict molecular clock assumes a constant speed with a constant mutation rate over time. In biological terms, that generally means neutral evolution (with no natural selection). These concepts have been incorporated in the BEAST and TransPhylo software that are based on Bayesian model (G. G. R. Murray et al., 2016).

Global genome-based phylogeny of the *Mycobacterium tuberculosis* complex (MTBC): Previously published data (Bos et al., 2014) shows that MTBC comprises of seven human-adapted lineages and several lineages are adapted to various wild and domestic animals. The human adapted lineage 2 (L2), lineage 3 (L3) and lineage 4 (L4) strains share a genomic deletion which is *M. tuberculosis (Mtb)* specific referred to as deletion 1 (TBD1) and these lineages are referred to as the modern lineages (Brosch et al., 2002). Specific lineages are geographically predominant in particular regions of the world which confirms adaptive evolution of particular MTBC lineages to be better transmitted in given geographical regions. *Mycobacterium tuberculosis* complex (MTBC) lineage 4 (L4) is geographically the most widespread cause of human tuberculosis. Genome based phylogeny further shows that L4 can be further subdivided into at least ten sub-lineages and that some of these sub-lineages are geographically restricted,

corresponding to ecological ‘specialists’, and others are globally distributed (‘generalists’). Among these sub-lineages, a specialists L4 Uganda genotype sub-lineage remains ecologically predominant in Uganda (Gagneux, 2018).

Antimicrobial resistance (Mycobacteria Antimicrobial Surveillance): Mycobacterium tuberculosis presents an opportunity to demonstrate the role of WGS in antimicrobial surveillance at an individual and population level. It takes up to seven weeks for culture results to be reported. Currently, WGS takes up to 12 days after culture for results of first line drugs to be reported. WGS is ideal in tuberculosis since many of the genomic targets associated with resistance are thought to be well characterized in Mycobacterium tuberculosis (Walker, Kohl, Omar, Hedge, Del Ojo Elias, Bradley, Iqbal, Feuerriegel, Niehaus, Wilson, Clifton, Kapatai, Ip, Bowden, Drobniewski, Allix-Béguec, Gaudin, Parkhill, Diel, Supply, Crook, Smith, Walker, Ismail, Niemann, Peto, & Modernizing Medical Microbiology (MMM) Informatics Group, 2015). WGS assays perform well for rifampicin and isoniazid resistance where the molecular markers of resistance are well defined. In the work flow shown in figure 3, the results are now reportable in days from weeks based on a method identifying resistance determinants from Mycobacterium whole genome sequences.

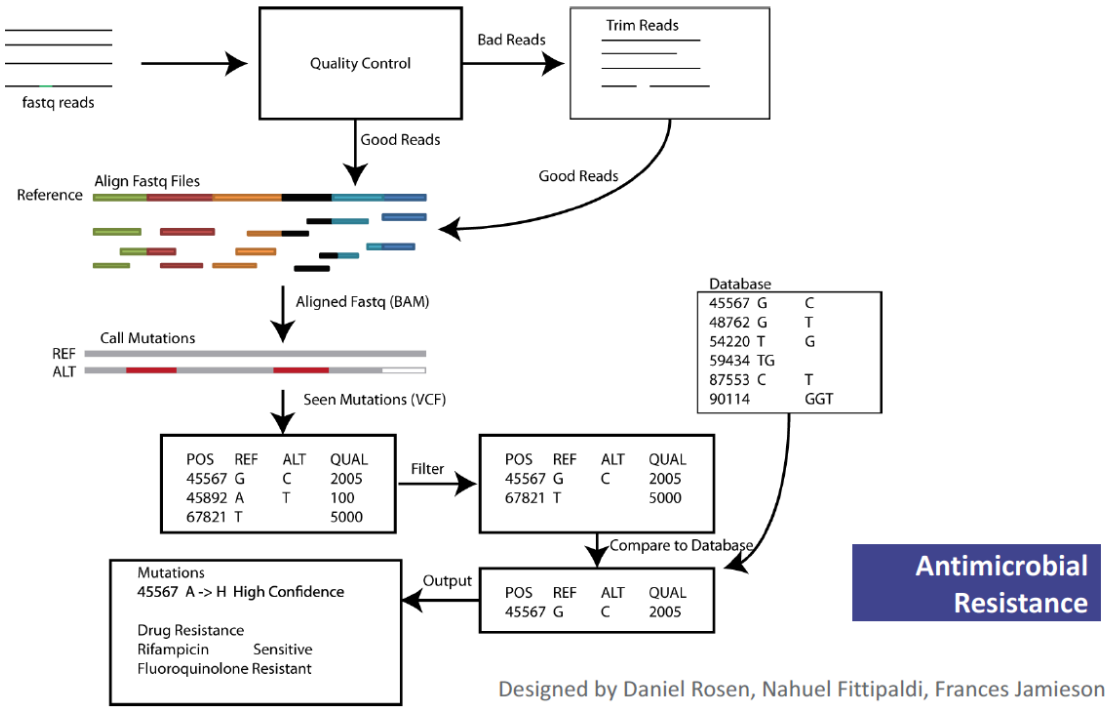


Figure 3 Procedure for predicting resistance of Tuberculosis from WGS

Biomarkers: The term "biomarker" refers broadly to any substance, structure, or process that can be measured in the body or its products and influence or predict the incidence of outcome or disease (Strimbu & Tavel, 2010). The measured response may be functional and physiological, biochemical at the cellular level, or a molecular interaction (Strimbu & Tavel, 2010). The use of biomarkers, and in particular laboratory measured biomarkers, in clinical research is somewhat newer, and the best approaches to this practice are still being developed and refined in tuberculosis (Suliman et al., 2018). For instance in work done by Suliman et al, they developed a simple 4- marker test that could be translated into a simple, rapid and affordable point-of-care test for field application in resource-limited settings, to identify individuals at high risk of developing TB (Suliman et al., 2018).

Mycobacterium tuberculosis Immune Biomarkers: Persons with active tuberculosis tend to have increased production of Interleukin-10 (IL-10) compared to those with latent infection

(Kim et al., 2014). *Mtb* uses the lipoarabinomannan (LAM) molecule in the cell wall to induce immune regulation pathways through transforming growth factor beta (TGFβ) and regulatory T cells that suppress protective immune responses (Dahl, Shiratsuchi, Hamilton, Ellner, & Toossi, 1996; Garg et al., 2008; Hirsch et al., 1996). This same molecule, when mannosylated, interacts with mannose receptors to drive dendritic cells to release the regulatory cytokine IL-10. The produced IL-10 inhibits antigen processing and presentation during tuberculosis infection (Chieppa et al., 2003) and suppresses Th1, Th2 and Th17 function (Almeida et al., 2009; Bobadilla et al., 2013; Hussain, Talat, Shahid, & Dawood, 2009). Thus, there is a delicate balance between pro-inflammatory immune responses that recognize *Mtb* infection, and conceal it by promoting granuloma formation, and regulatory immune responses that, while suppressing Th1, Th2 and Th17 function to prevent immunopathology, may support mycobacterial replication and disease progression (García Jacobo et al., 2014; Hussain et al., 2009; Lin & Flynn, 2010; Sutherland, Adetifa, Hill, Adegbola, & Ota, 2009). Based on this delicate physiological balance between the host physiology and microbial assault, there is an array of patient and pathogen proteins to assess for diagnostic potential.

In their work, Bark and colleagues assesses over 289 potential host biomarkers of recent infection and CLEC3B (Bark et al., 2017), ECM1(Bark et al., 2017), IGFALS (Bark et al., 2017), SELL (Bark et al., 2017) (LAM1), VWF (Bark et al., 2017) were the best in differentiating ‘progressors’ from ‘non-progressors’. In addition, interferon gamma inducible protein 10 (IP-10)(Biraro et al., 2016a), a chemokine (CXC motif) ligand 10 expressed by macrophages when stimulated by interferons and other pro-inflammatory cytokines (Farber, 1997; Luster, Unkeless, & Ravetch, n.d.; Ruhwald, Aabye, & Ravn, 2012) has been suggested as a key biomarker. IP-10 induces movement of monocytes and activated Th1 cells to sites of

inflammation through its interaction with the CXC chemokine receptor 3 (Dufour et al., 2002; Ferrero et al., 2003). IP-10 promotes Th1 immune responses by up regulating the expression of IFN-γ and is involved in the DTH immune responses (Kaplan, Luster, Hancock, & Cohn, 1987; Orme & Cooper, 1999). It has been suggested that production of this chemokine in response to MTBC antigens could be used as a marker of infection (Azzurri et al., 2005; Biraro et al., 2016b; Ruhwald et al., 2007).

Advent of biomarkers that are unique to *Mycobacterium tuberculosis* infection remain promising since they are not masked by host immune response. One such protein that is promising is the *Mycobacterium tuberculosis* Thymidine kinase whose production increases over 10 fold during replication of mycobacteria (Wayengera, Kateete, Asiimwe, & Joloba, 2018). Whereas growth and proliferation of *Mtb* presents a potential surrogate marker for TB disease progression. A noticeable gap, nonetheless, is the absence of easy to use and culture independent assays for detecting *Mtb* growth and proliferation. M. tuberculosis-thymidylate (a.k.a thymidine monophosphate, TMP) kinase or simply TMKmt, a phosphotransferase that catalyzes the phosphorylation of deoxythymidine monophosphate (dTMP) to the diphosphate precursor used to generate deoxythymidine-triphosphate (dTTP) that finally gets integrated into the growing M. tuberculosis DNA chain presents an opportunity (Wayengera et al., 2017). Secretory levels of TMKmt are characterized by a 10 to 20-fold increase after the G1/S transition, and remain high until about the time of cell division when they then decline rapidly in vitro.

Treatment of LTBI: One of the LTBI control strategy that public health policy makers have rolled out is the prescription of treatment for individuals that test positive for LTBI commonly referred to as Tuberculosis Preventive Therapy (TPT). There are indeed treatment regimens that can greatly reduce but not eliminate the risk of reactivation disease in latently

infected individuals (Lim et al., 2006; Wilkinson, 2000). These involve taking antibiotics for up to 24 months, and carries a potential risk of serious, or even fatal, side-effects. Extensive prophylactic treatment of latently infected people has been proposed and rolled out as the key to eliminating TB in the United States and elsewhere (Tang & Johnston, 2017). In populations with high prevalence of HIV, the risk of reactivation tuberculosis is significantly higher. In this most at risk population, a six-month course of isoniazid conferred short-term protection against tuberculosis among PPD-positive, HIV-infected adults (Whalen et al., 1997). Whereas IPT is effective, it's benefits are weakened by lack of valuable diagnostic tests that could predict which latently infected individuals are most at risk of developing active tuberculosis especially in endemic areas where infection is overwhelmingly high (Pai & Schito, 2015).

Chapter 3

MATERIAL AND METHODS

METHODOLOGY FOR AIM 1

Aim 1: To adapt an outbreak investigation model to ascertain who infected whom in a TB endemic sub-Saharan African population.

Study Design Overview: To address Aim 1, an outbreak investigation mathematical model was used to demonstrate who infected whom with a quantifiable degree/measure of certainty in a TB endemic setting of Kampala Uganda.

Study population and data source for Aim 1: From 2012 to 2019, a cross-sectional study (Community Health and Social Network of Tuberculosis - COHSONET study) of patients with tuberculosis was conducted among residents of Rubaga division of Kampala, Uganda. This Aim used data collected from the COHOSONET study for adapting the model. Data from the time cough symptoms started, social network (proxy for contact frequencies between subjects) and genetic distance (proxy for genetic similarity of pathogens isolated, cultured and whole genome sequenced from subjects sputum) was used to generate a matrix, V of transmission probabilities. In the transmission matrix V, element $v_{i,j}$ is the probability that subject i acquired his infection from subject j and v_i is the vector of transmission probabilities linking case i to any other case in the study (Figure 4).

$$\boldsymbol{V} = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix} = \begin{pmatrix} v_{1,1} & v_{1,2} & \cdots & v_{1,n} \\ v_{2,1} & v_{2,2} & \cdots & v_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ v_{n,1} & v_{n,2} & \cdots & v_{n,n} \end{pmatrix}$$

Figure 4 Transmission Matrix

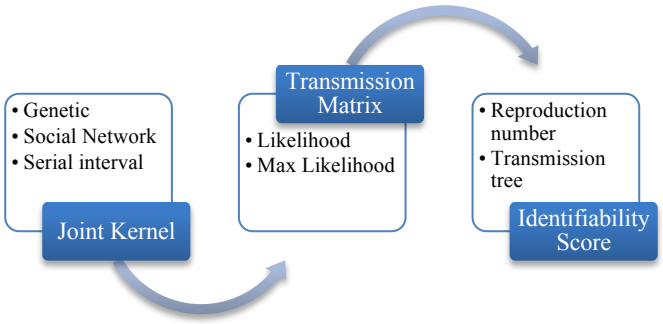


Figure 5: Aim 1 methodology conceptual framework

In the parent study, participants listed social contacts with whom they were close enough to carry out a normal conversation and the duration of such contacts. Demographic, social, and clinical characteristics were obtained through interviews performed by trained field workers using standardized questionnaires.

DNA sequencing and generating SNP table: Isolated genomic DNA of individual strains was sequenced on the Illumina platform (Illumina, San Diego, CA, USA) according to manufacturer instructions. Whole genome sequencing was performed at the CDC Atlanta microbiology laboratory using DNA extracted from sputum samples. Resulting FASTQ paired-end reads were processed using the CDC analysis pipeline for Mycobacterium tuberculosis complex genotyping from high throughput sequence to generate SNP tables and phylogenetic trees. The raw reads were trimmed and those that were <36bps were excluded. Next, Burrows-

Wheeler Aligner (Li & Durbin, 2009) was used to map the reads to the H37Rv Mycobacterium Tuberculosis reference genome (GenBank accession number NC_000962.3). Burrows-Wheeler Aligner (BWA) is a software package for mapping low divergent sequences against a reference genome. Samples with <75% coverage or <25x depth of coverage were discarded. Variants were called using SAMtools mpileup (http://www.htslib.org) and the Genome Analysis Toolkit (GATK; Depristo et al., 2011) (https://software.broadinstitute.org/gatk/). Only variants supported by at least five reads, including one in each direction were accepted. Recognizing that repetitive regions frequently contain erroneous variant calls, regions annotated as 'repetitive' elements (e.g. PPE and PE-PGRS gene families), insertions and deletions (InDels), and consecutive variants in a 12 bp window were excluded. Additionally, variants in drug resistance associated genes (i.e., mutations associated with drug resistance) were excluded. The resulting high-quality SNPs were concatenated to produce SNP alignments and SNP tables grouped according to the major MTBC lineages.

Serial time dependent arranged time kernel/matrix: We arranged the subjects according to the date they recalled to have started coughing as reported at enrolment thus the case with the earliest cough symptom appeared first in the matrix and the case with the latest cough symptom onset last (figure 6). Briefly, we subtracted the number of days the cases reported to have coughed at the time of enrollment from the date of enrolment to derive the original date of infectiousness. These new dates were arranged from the earliest to the last date and this arrangement of cases was adopted for the following steps of this analysis. To generate a symmetric matrix of difference of time in days, we subtracted the cough symptom dates of all cases from each subject. This resulted in a vector for each case. Binding these vectors as rows results in a 62 x 62 symmetric matrix of time in days. To transform this matrix to a kernel, we

divided each row with its own row sum so that the largest difference in days resulted in a higher probability respectively. This resultant matrix formed the time kernel, K^T where two cases (i,j) separated by a serial interval were linked (Glass, Mercer, Nishiura, McBryde, & Becker, 2011) by their respective probability.

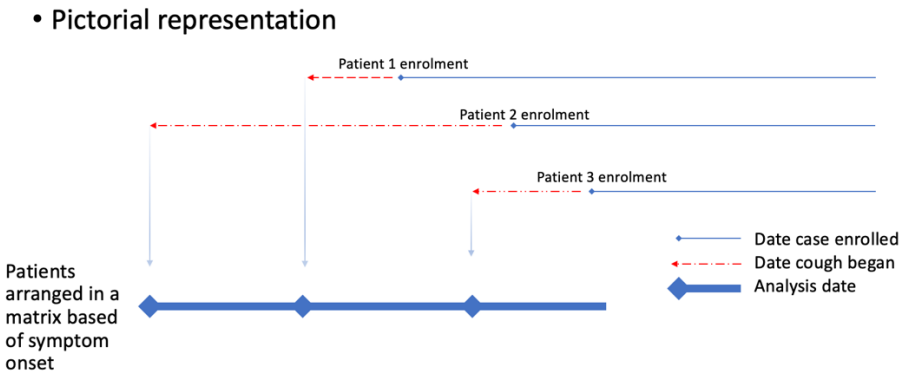


Figure 6: Generation of time kernel

Generating social kernel from social network distance: To generate probability of pairwise social links among cases in this study, social distances from the parent study network (networks of 123 cases and 124 controls and their first- and second-degree contacts) was analyzed. Briefly, trained social workers interviewed participants, beginning with the index case and their resident controls to generate a list of first- and second-degree contacts and meeting venues that allowed for the construction of ego-centric networks which were combined to form the network. The parent study network was constructed using machine learning (Dedupe software), Statnet software version 2018.10 and reliance on local content expertise.

To obtain social distances among this study subjects, a geodesic matrix (social distance matrix) for all members in the parent network was obtained in r-program. Lineage specific social distance matrices for the cases in the study were filtered and our analysis was focused on lineage four genotype, the most prevalent lineage in Uganda. The extracted matrix for cases was

arranged according to the dates of symptom onset as per the time interval kernel. By transforming this social distance matrix through its inverse, we obtained a social kernel, K^S. This work was done using r- program (version 3.6.0)

Generating genetic kernel from genetic distance matrix: To generate genetic distance based pairwise probabilities linking cases in the study, we used Whole Genome Sequence based SNP differences among lineage four specific MTBC pathogens. A SNP difference matrix for cases was obtained by importing aligned sequences into Geneious software (version 2019.2.1) as illustrated by Figure 7. Yielding a genetic kernel was done by transforming the genetic distance matrix through its inverse in r - program (version 3.6.0). The cases in the genetic kernel were arranged according to their respective dates of cough symptom onset to form a Genetic kernel, K^G.

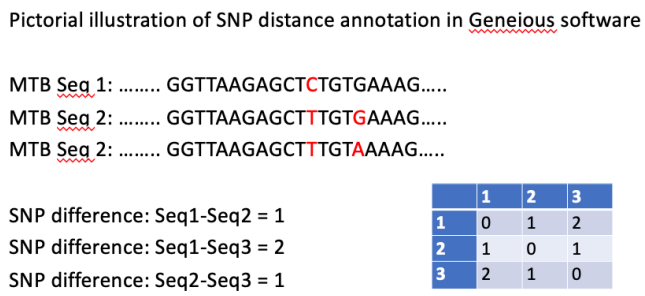


Figure 7. Generation of SNP difference

Element-wise multiplication of kernels to generate a transmission matrix: To generate a transmission matrix (pairwise likelihood of transmission) among cases in the study, we considered information from date of onset of cough (time kernel), Whole Genome Sequence SNP difference data (genetic distance kernel) and social network distance (social kernel) in a case of independent contribution to transmission. The elements of the likelihood matrix were found by element-wise multiplication of the three matrices assuming absence of correlation

between these quantified distances. Kernels were multiplied in r-program as per the formula below;

Transmission matrix/ Kernel, $V = K^T * K^S * K^G$

V= transmission matrix
K^T = Time matrix
K^S = Social distance matrix
K^G = Genetic distance matrix

Equation 1:Element-wise multiplication of kernels

The super-indexes denote kernels based on time (T), social (S) and genetic (G) data, respectively. By assuming that all cases have been found (the complete outbreak has been observed), rows were scaled to add up to 1 by dividing each row of the matrix with sums of row values in the transmission matrix.

Transmission tree: The transmission matrix was analyzed to generate a transmission tree. Briefly, the Maximum likelihood (ML) of the vector of transmission probabilities ($v_{i,j}$) for each individual i having acquired the infection from any of the members j in the study was considered. Since the ML of probabilities is 1, we identified the largest element in the vector for each individual and replaced it with 1. By replacing all the remaining values with 0, we generated a sparse matrix that was used as a directed adjacency matrix for the construction of a directed transmission tree to identify who infected whom among transmission clusters. Using this adjacency matrix, we plotted a directed tree using i-graph version 1.2.4.1 in r- program version 3.6.0.

Reproduction Number: The transmission matrix was analyzed to generate reproduction numbers for cases that infected others in the network. Briefly, the Reproductive number of a member (j) in this network refers to the number of cases (i) that probably got their infection

from j. Considering that the rows in the transmission matrix represent a vector of probabilities ($v_{i,j}$) for each individual i having acquired the infection from any of the members j in the study. Considering a ML of probabilities 1, we identified the largest element in the vector for each individual and replaced it with 1. By replacing all the remaining values with 0, we generated a sparse matrix with 1 and 0. The reproduction number was calculated by obtaining the column sum for each member j as per the formula below (Newman, 2010).

$$R_j = \sum_{i=1}^{n} v_{i,j}$$

Equation 2: Reproduction number formula

To calculate the overall network reproduction number, we added the column sums R_j and divided the total by the total number of column sum with $R_j > 0$.

Identifiability Score: was analyzed to generate an identification metric as a measure to evaluate the degree of certainty of identified links. Briefly, the identification metric was calculated based on the assumption that the infection of susceptible individual *i* by case *j* is a Bernoulli trial with variance;

Variance, $var_{i,j} = v_{i,j}(1 - v_{i,j})$

$q(i) = sum\ (var_{i,j})$

Equation 3: Identifiability Score formulae

Variance was used as the measure of how accurately subjects j in the observed population may be identified as the person who infected subject *i*. The sum of the variance for subject *i* then was considered as a measure of how accurately any of the subjects observed may be identified as having been the source of infection to subject *i*;

The quantity 1−q(*i*) was used as a score of the degree of identification of the infection parent of subject i. When there is only one non-zero element in row *i* of the transmission matrix, the parent node is perfectly known, and 1−q(*i*) = 1 indicating perfect identification. Using the formula below, we calculated the minimum possible identifiability score if the probability of infection from all 61 potential j sources of infection was equal. This would represent complete ignorance about who infected subject i, among those for whom a link should be possible (i.e. those with non-zero $v_{i,j}$)(Teunis et al., 2013b).

$$1 - q(i) = 1 - \sum_{i=1}^{n_i} \frac{1}{n_i}\left(1 - \frac{1}{n_i}\right) = \frac{1}{n_i}$$

Equation 4: Minimum identifiability Score formula

We used the minimum identifiability score to calculate the percentage identifiability metric. Briefly, we subtracted the minimum identifiability score from each score and divided the results by 1- the minimum identifiability score and multiplied the result by 100 to generate percentages scores.

Using identifiability score to determine strong links in the transmission network: We applied the Identifiability score on the overall network using i-graph version 1.2.4.1 in r-program version 3.6.0. We generated trees for individuals with a corresponding identifiability score of >5%, >10% and >20% by constructing subgraphs of the main network that included individuals with the corresponding identifiability scores respectively. This sub-graph was implemented first by dropping individuals having a score less than the identifiability score and re-construction of score specific networks.

METHODOLOGY FOR AIM 2

Aim 2: To study whether mutations in drug resistance candidate genes are associated with tuberculosis transmission in peri-urban Kampala, Uganda.

Study design overview: To address Aim 2, a case control study design (case and control MTBC genomes determined by the SeqTrack algorithms) was used to assess whether mutations in drug resistance candidate genes are associated with being an ancestor genome in transmission clusters. In the first step, the study was conducted among 62 participants lineages from the Parent study described in Aim 1 above and the results compared within a combined dataset including sequences collected at the same period in Uganda.

Whole genome database (primary study and NCBI deposited sequences): In addition to the 62 lineage four sequences from the parent study mentioned in Aim 1 above, we downloaded whole genome sequences using the Sequence Read Archive tool kit (Leinonen, Sugawara, & Shumway, 2011). To access the sequences, we searched the National Center for Biotechnology Information (NCBI) website with a search term "Mycobacteria Tuberculosis Uganda Whole Genome". We considered sequences of all the four bio-projects from studies that included patients from Kampala. We only included studies whose sequences were from sputum isolated pathogens and were sequenced by the Illumina sequencing platform, similar to the parent study. In summary, we downloaded 59 samples from a study that aimed at using WGS to characterize *Mtb* isolates from HIV-seropositive Ugandans with TB and CD4+ T cell counts of 0 – 1,150 cells/ul; work that was part of a grant to study community transmission of TB in urban Africa (Wampande et al., 2015). The second bio-project had 90 WGS from drug resistant *Mtb* isolates from Uganda. This study used WGS to complement the TB drug resistance national survey in Uganda (Ssengooba, Meehan, et al., 2016). The third bio-project aimed to understand the

heterogeneity of *Mtb* isolates obtained from sputum and blood compartments concurrently from 13 patients. This third study used WGS to reveal mycobacterial microevolution among concurrent isolates from sputum and blood in HIV infected TB patients (Ssengooba, de Jong, Joloba, Cobelens, & Meehan, 2016). The fourth study involved evolution of drug resistance by elucidating emergence and transmission of multidrug-resistant *Mtb* isolates using WGS; this study contributed 51 sequences (Clark et al., 2013). We excluded isolates with low phred quality scores. There were 168 isolates and of them, 129 isolates belonged to lineage 4 (Figure 28).

Study No.	**01**	**02**	**03**	**04**	**COHSONET**
Total Sequences	59	90	13	51	89
Good QC Seq	04	69	12	04	79
Lineage 4	04	52	07	03	62
Institution	Makerere University	Institute of tropical medicine, Belgium	Institute of tropical medicine, Belgium	LSHTM	University of Georgia, GHI/ CDC
Date of submission	18-July-2018	23-Feb-2016	6-Nov-2015	21-May-2013	Pending
Study design	UNK	Survey	Diagnostic	Longitudinal	Social Network
Accession No.	PRJNA481638	PRJEB10533	PRJEB10577	PRJEB224	N/A

Publication	Kateete, et al.	Ssengooba, et al. 2016	Ssengooba, et al. 2016	Clark TG, et al. 2013	Whalen, et al.

Table 9: Data sources for this study (4 Bio-projects/ sequences from NCBI)

Outcome variable (identification of ancestors genomes): The SeqTrack algorithm was used to create transmission networks among the sequences. Briefly, a SNP difference matrix for samples was generated using Geneious software (version 2019.2.1) from aligned sequences. Since the mutation rate of the MTBC in clinical settings has been estimated at 0.3-05 substitutions per genome per year (Eldholm & Balloux, 2016), we used a non-conservative mutation rate of 0.5 substitutions per genome per year. The study samples in this analysis were arranged according to their respective dates of cough symptom onset, a proxy for infectiousness. Briefly, we subtracted the number of days the cases reported to have coughed at the time of enrollment from the date of enrolment to derive the original date of infectiousness. Based on the cough onset dates, samples were arranged from the earliest to the last date and this arrangement of cases was adopted as the subject IDs in the analysis (similar to method in Aim 1, figure 6). The output of the SeqTrack analysis were a list of ancestor sequences, accompanying likelihood of links and the 'infected /descendant MTBC genome sequence '.

Exposure variable (mutations in candidate Genes): We used the TBprofiler pipeline (Coll et al., 2015) to identify candidate gene mutations, genotypic drug resistance profiles and phylogenetic classification of participants' MTBC Whole Genome Sequences. Briefly, the TBprofiler pipeline assessed study sequences for quality using FastQ (a step where poor quality sequences were eliminated) and aligned the sequences using Burrows-Wheeler Aligner (BWA). The aligned sequences variants were called using SAMtools, a suite of programs consisting of

BCFtools for writing Variant Call files (VCF), filtering and summarizing SNPs and short indel sequence variants and LoFreq a fast and sensitive variant-caller for inferring SNVs and indels from next-generation sequencing data. The variants, in form of Variant Call files were then compared to a drug-resistance database by Walker and colleagues (Walker, Kohl, Omar, Hedge, Del Ojo Elias, Bradley, Iqbal, Feuerriegel, Niehaus, Wilson, Clifton, Kapatai, Ip, Bowden, Drobniewski, Allix-Béguec, Gaudin, Parkhill, Diel, Supply, Crook, Smith, Walker, Ismail, Niemann, Peto, Davies, et al., 2015). The final output of this analysis was a report for each participant's MTBC genome.

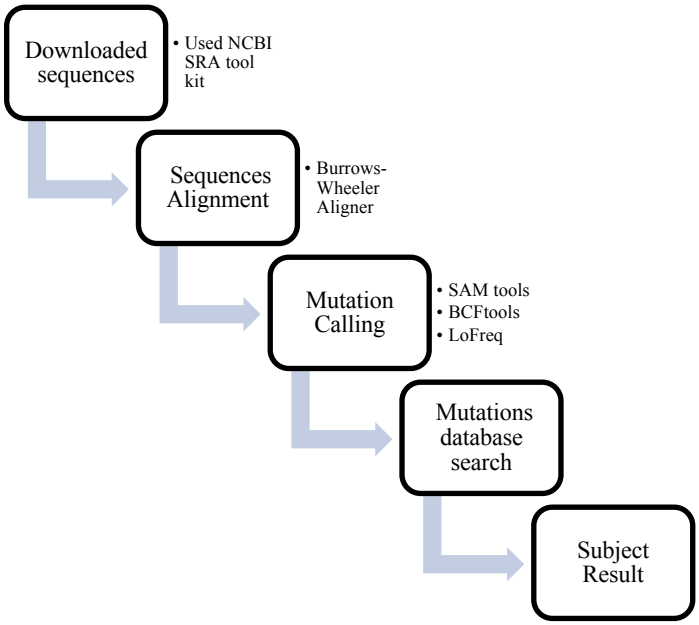


Figure 8: SeqTrack pipeline used to analyze participants WGS

Statistical analysis: Demographic characteristics of ancestor and non-ancestors sequences were compared using frequency tables, chi squared tests, Kruskal-Wallis chi-squared test and Fisher' s exact test in R-statistical package. Frequencies of mutations in candidate genes were tabulated for ancestors and non-ancestors. We calculated odds ratios and 95% confidence intervals (OR, 95% CI) for the association of ancestor status with candidate genes mutations

using logistic regression. We adjusted the associations for sex, smoking, duration of cough, and genotypic drug resistance (a proxy for phenotypic drug resistance).

METHODOLOGY FOR AIM 3

Aim 3: To test whether interferon gamma cytokine concentration levels can identify recent infection with *Mycobacterium tuberculosis* in a community setting

Study design overview: To address Aim 3, a case control study design was used to assess whether being LTBI positive or having recent infection with *Mtb* is associated with unit increase in blood concentration levels of IFN-γ cytokine after stimulation with *Mtb*-specific antigens.

Outcome variable (TST group classification): From a cohort of adults who were TST negative at baseline and followed up for 1-2 years, 55 subjects who tested TST negative and 64 who tested TST positive at the last visit of follow up were enrolled in this study. Among those who tested TST positive at the initial assessment into the cohort, 97 were assumed to be remotely infected and were enrolled as a third group in this study. All participants were residents of Rubaga division of Kampala, Uganda. Initial TST was performed by Mantoux method with 5 tuberculin units (TU) of purified protein derivative (PPD). After the initial evaluation, participants were evaluated after one year for active TB and LTBI. Whole blood collection for QuantiFERON-TB Gold In-Tube assay (QFT) was done at the last follow up visit. TST conversion was defined as contacts with an initial TST < 5 mm (diameter of induration) at baseline visit, who subsequently converted their skin test to positive (TST ≥ 10 mm and or an increment of 6 mm) after one year (Castellanos et al., 2018). If the above conditions were not met, subjects were classified as persistent TST negative.

Sample Processing: Whole blood samples for QFT testing were collected in three 1 mL tubes provided with the testing kits, transported to and received by the laboratory at room

temperature within approximately 2 hours of blood collection. The tubes were incubated at 37ºC for 16-24 hours when the plasma was separated and stored at -80° until Enzyme Linked Immuno-Sorbent Assays (ELISA) were performed.

Exposure variable (Measurement of IFN- γ by ELISA): ELISA assays to test Interferon gamma were carried out as per manufacturers instruction and optical densities (OD) values were used to compute IFN-γ concentrations in international units per milliliter (IU/mL) using QuantiFERON-TB Gold IT Analysis Software (version 2.17) (Qiagen, n.d.). All samples were evaluated by laboratory scientists who were blinded to the study participants and the clinical data. The background corrected IFN-γ concentrations were defined by subtracting the concentration in unstimulated supernatant from the corresponding concentration in *Mtb* antigen stimulated supernatant.

Statistical analysis: Demographic characteristics of recently infected uninfected and remotely infected participants were compared using frequency tables, chi squared tests, Kruskal-Wallis chi-squared test, Fisher' s exact test and Analysis of variance methods in R-statistical package. Quantitative concentrations of IFN-γ cytokine were tabulated for recently infected and remotely infected participants. We calculated odds ratios and 95% confidence intervals (OR, 95% CI) for the association of recent infection and unit change in concentration of IFN-γ using logistic regression. We adjusted the associations for age, HIV status, religion and education levels accordingly.

$$\log(\text{Odds}) = \Sigma_i \beta_0 + \Sigma_i \beta_i x_i + \Sigma_i \beta_i x_i$$

Equation 5: Logistic Regression model

Data was analyzed to generate test Receiver Operator Curves (ROC) to define the diagnostic performance of IFN-γ cytokine concentration to discriminate the three groups. The area under the ROC curve (AUC) was determined in r-program using the pROC package (Robin et al., 2011), together with the optimal cut-off value that maximized Youden's index (sum of % sensitivity and % specificity - 100). Youden's index, often used in conjunction with ROC analysis was used as a criterion for selecting the optimum cut-off point (Schisterman, Perkins, Liu, & Bondell, 2005). A cutoff values in IU/ml to discriminate between recently infected and uninfected or remotely infected groups was proposed.

Ethical Approval: Written informed consent was obtained from all participants prior to study inclusion for both the parent study and for studies downloaded from NCBI website (Bio-projects). Institutional review board clearance was obtained from Ethics Committees at Makerere University School of Public Health and the University of Georgia for the Parent study. Based on the NCBI Sequence Read Archive (SRA) study pages for downloaded bio-projects, there was no requirement for additional IRB approval and Institutional review board clearances were obtained by the original authors.

ORGANIZATION OF THE STUDY

Aim 1, 2 and 3 of the study are in chapters 4,5 and 6 respectively. Each of these chapters is in a manuscript-style format, with a stand-alone abstract, introduction, methodology, results, and discussion. Chapter 7 summarizes a synthesis of the study results, their implication and future direction. The references for the entire text come after the last chapter.

Chapter 4

EXTENDING A BASIC OUTBREAK INVESTIGATION MODEL TO EXAMINE TUBERCULOSIS TRANSMISSION LINKS USING SOCIAL NETWORK AND GENETIC EPIDEMIOLOGY METHODS IN KAMPALA, UGANDA

[1]Kirimunda S, Quinn F, Kakaire R, Chengwei L, Sekandi J, Whalen. To be submitted to American Journal of Epidemiology

ABSTRACT

Background: This study addresses the problem of ongoing transmission of *Mycobacterium tuberculosis (Mtb)* in Sub-Saharan Africa. There is a shortage of user-friendly methodologies able to concurrently utilize multiple sources of data to infer transmission in both endemic and outbreak tuberculosis settings. The current methods used to study transmission using WGS and social network data are sophisticated and require bioinformatics expertise to deploy (Alaridah et al., 2019; Roetzer et al., 2013). We hereby extend a basic outbreak investigation model to confirm who transmitted to whom in a TB endemic setting in Kampala, Uganda.

Methods: Using data from a community social network study, we studied who infected whom using social network, Whole-Genome Sequencing and serial interval characteristics of the study population. The duration of cough symptoms, social network distances and genetic distances were used to generate a matrix of transmission probabilities which was analyzed to get a transmission tree, reproduction numbers, and an identifiability score.

Results: The overall transmission network tree showed a total of 15 clusters with the largest cluster comprising of 12 members and the smallest cluster of 2 members. There were 36 (58.1%) individuals who were identified as potential sources of infection in the network and 26 individuals who did not infect anyone else. When a degree of certainty at cutoff of >5%, >10% and >20% was applied to the overall network, a total of 10 clusters, 5 clusters and 5 clusters remained respectively.

Conclusion: In this study, we adopted a basic outbreak investigation method capable of using multiple data sources to identify transmission links among TB cases. This approach can

easily be used by National TB control programs to identify and monitor outbreaks in endemic settings.

BACKGROUND

Tuberculosis (TB), caused by infection with *Mycobacterium tuberculosis* (*Mtb*) (Tom A Yates et al., 2016), is the leading cause of death caused by a single pathogen globally ("WHO | Global tuberculosis report 2017," 2017). Inhalation of aerosols containing *Mtb* Complex (MTBC) bacilli by susceptible hosts is the main route of transmission of TB ("WHO | Global tuberculosis report 2017," 2017). Based on epidemic theory and what is known about TB transmission, epidemics continue to occur when one index case is replaced by one or more cases (Whalen, 2016). Harnessing available resources to prevent new cases of disease remain crucial in elimination of TB.

Globally, an estimated 10.0 million people fell ill with TB in 2018, a number that has been relatively stable in recent years. The burden of disease varies enormously among countries, from fewer than five to more than 500 new cases per 100,000 population per year (WHO, 2019). In 2018, TB incidence in Uganda was 200 (range, 118-304) cases per 100,000 individuals per year compared to 202 per 100,000 in 2015 and remains among the 30 high HIV/TB burden countries (WHO, 2019).

Studies continue to show that transmission mostly occurs in the community and not in the household and that in fact susceptible individuals are exposed to infection by the time a new case is diagnosed (Auld et al., 2018; Martinez et al., 2017b; Sekandi et al., 2015; Whalen, 2016) (Kakaire et al, unpublished data). Various studies have attempted to characterize community transmission of TB using sequence data and have shown that some individuals infect a large number of others and are thus 'super-spreaders', while many infectious individuals do not transmit (Escombe et al., 2008; Kline et al., 1995; RILEY et al., 1962; Snider et al., 1985; van

Geuns et al., 1975; T. A. Yates et al., 2016; Ypma et al., 2013). Lately, these approaches mainly use *Mtb* Complex (MTBC) phylogenetic clustering to signify on-going transmission (recent transmission) or the presence of unique strains as indicators of reactivation of latent TB infections to disease (Asiimwe et al., 2009; Gardy et al., 2011a; Weis et al., 2002).

An important way to quantify the transmissibility of a pathogen is through the reproductive number, the average number of secondary infectious cases caused by one infectious individual (Galvani & May, 2005). However, for many infectious diseases, a heavily skewed distribution of new infections caused by an infected individual is often observed, with most diseased individuals infecting none or only a few others, and a few individuals infecting many others (Lloyd-Smith, Schreiber, Kopp, & Getz, 2005). In the transmission of TB, evidence shows that person-to-person transmission may be the primary driver of the epidemic in African settings (Auld et al., 2018).

Another approach to studying transmissibility of a pathogen is the generation interval, the time it takes to ‘generate' a secondary infectious person by an infected individual. This allows direct observation of the number of secondary infections caused by an infected host. This approach has not been widely used for TB, since the time of infectiousness of a primary case and disease in secondary cases can potentially be many years apart, or an infected person never develops disease (Vynnycky & Fine, 2000). However, there is documented evidence that this approach is relevant since an appreciable proportion of new cases are attributed to primary progressive disease especially among children in endemic areas (Auld et al., 2018; Martinez et al., 2017b).

There is now an increase in the quality of data available through sentinel and research programs. National tuberculosis control program data often consist of cases, their symptoms,

symptom onset dates, geo-spatial residence, genotype data of the pathogen and clinical data (World Health Organization, 2015b). Such data is sufficient to show when TB cases are clustered, in space and time (Shah, Auld, Brust, Mathema, Ismail, Moodley, Mlisana, Allana, Campbell, Mthiyane, Morris, Mpangase, van der Meulen, et al., 2017). This increase in volume of high quality data and notable advances in outbreak investigation tools avails an opportunity to harness our understanding of when and from whom new TB cases arise with greater certainty (Teunis et al., 2013b).

As a way to ascertain who infected whom in TB endemic settings, methodologies that use more than one study population characteristic have been proposed by previous researchers (Auld et al., 2018; Guthrie et al., 2018; Martinez et al., 2017b). More advanced methods have been designed and implemented in TB outbreak investigation using Whole-Genome Sequencing and epidemiologic data with varying emphasis on either data source (Gardy et al., 2011a; Walker et al., 2013b). However, in instances of both endemic and outbreak TB transmission, there is a shortage of methodologies able to concurrently utilize multiple sources of data to infer transmission (Teunis et al., 2013a). By analyzing data from a community social network study, we examined to know who infected whom using social network, Whole-Genome Sequencing and serial interval characteristics of the study population. We hereby extend a basic outbreak investigation model by Teunis and colleagues (Teunis et al., 2013b) to confirm who transmitted to whom in a TB endemic setting in Kampala, Uganda.

METHODS

Study Setting, population and data collection: The parent study was a cross-sectional study of 123 TB index cases residing in Rubaga Division of Kampala, Uganda and their contacts, called COHSONET (The Community Health and Social Network Study) conducted from 2012 to

2016 (Castellanos et al., 2018). Contact networks were an extended form of ego-centric sampling where 123 TB index cases and 124 randomly selected community controls that were matched on age group, sex and parish were asked to identify their respective social contacts. Cases with a positive culture isolate had DNA extracted and whole genome sequencing performed using the Illumina platform. Only those sequences which met a pre-set quality standard were included in this analysis (Figure 9).

In the parent study, participants listed social contacts with whom they were close enough to carry out a normal conversation and the duration of such contacts. The contacts (first level contacts), also listed their contacts (second level contacts i.e., the contacts of contacts). The respective second order/level egocentric social networks for each index (case or control) were merged to produce a merged social network using machine learning approach (Dedupe.io). Demographic, social, and clinical characteristics were obtained through interviews performed by trained field workers using standardized questionnaires. A clinical TB diagnosis questionnaire administered to cases at enrollment recorded if a case reported a cough symptom at the time of diagnosis and for how long the cough had lasted.

DNA sequencing and SNP tables: Isolated genomic DNA of individual strains from 102 index cases was sequenced on the Illumina platform (Illumina, San Diego, CA, USA) according to manufacturer instructions. Whole genome sequencing was performed at the CDC Atlanta microbiology laboratory using DNA extracted from sputum samples. Resulting FASTQ paired-end reads were processed using the CDC analysis pipeline for Mycobacterium tuberculosis complex genotyping from high throughput sequence to generate SNP tables and phylogenetic trees. The raw reads were trimmed and those that were <36bps were excluded. Next, Burrows-Wheeler Aligner (Li & Durbin, 2009) was used to map the reads to the H37Rv Mycobacterium

Tuberculosis reference genome (GenBank accession number NC_000962.3). Burrows-Wheeler Aligner (BWA) is a software package for mapping low divergent sequences against a reference genome. Samples with <75% coverage or <25x depth of coverage were discarded. Variants were called using SAMtools mpileup (http://www.htslib.org) and the Genome Analysis Toolkit (GATK; Depristo et al., 2011) (https://software.broadinstitute.org/gatk/). Only variants supported by at least five reads, including one in each direction were accepted. Recognizing that repetitive regions frequently contain erroneous variant calls, regions annotated as 'repetitive' elements (e.g. PPE and PE-PGRS gene families), insertions and deletions (InDels), and consecutive variants in a 12 base pair window were excluded. Additionally, variants in drug resistance associated genes (i.e., mutations associated with drug resistance) were excluded. The resulting high-quality SNPs were concatenated to produce SNP alignments and SNP tables grouped according to the major MTBC lineages.

Analytical model data input: Data from the time cough symptoms started, social distance/number of SNPs (proxy for contact frequencies between subjects) and genetic distance (proxy for genetic similarity of sputum collected, pathogen isolated, DNA extracted, and whole genome sequenced, and single nucleotide polymorphism (SNP) analyzed) was used to generate a matrix, V of transmission probabilities. Social network distance was defined as the geodesic distance (the length of the shortest path between pairs) in the social network. Genetic distance was defined by the number of SNPs separating *Mtb* genomes sequenced from DNA of pairs of individuals. In the transmission matrix V, element $v_{i,j}$ is the probability that subject i acquired his infection from subject j and v_i is the vector of transmission probabilities linking case i to any other case in the study (Figure 5).

Serial time dependent arranged time kernel/matrix: We arranged the subjects according to the self-reported date of when cough began; thus, the case with the earliest cough symptom appeared first in the matrix and the case with the latest cough symptom onset was last. Briefly, we subtracted the number of days the cases reported to have coughed at the time of enrollment from the date of laboratory TB diagnosis to derive the original date of infectiousness (Figure 6). These new dates were arranged from the earliest to the last date and this arrangement of cases was adopted for the following steps of this analysis. To generate a symmetric matrix of difference of time in days, we subtracted the cough symptom dates of all cases from each subject. This resulted in a vector for each case. Binding these vectors as rows results in a 62 x 62 symmetric matrix of time in days. To transform this matrix to a kernel, we divided each row with its own row sum so that the largest difference in days resulted in a higher probability, respectively. This resultant matrix formed is the time kernel, K^T where two cases (i,j) separated by a serial interval were linked (Glass et al., 2011) by their respective probability.

Generating social kernel from social network distance: To generate probability of pairwise social links among cases in this study, social distances from the parent study network (networks of 123 cases and 124 controls and their first- and second-degree contacts) was analyzed. Briefly, trained social workers interviewed participants, beginning with the index case and their resident controls to generate a list of first- and second-degree contacts and meeting venues that allowed for the construction of ego-centric networks which were combined to form the network. The parent study social network was constructed using machine learning (Dedupe software), Statnet software version 2018.10 and reliance on local content expertise.

To obtain social distances among study subjects, a geodesic matrix (social distance matrix) for all members in the parent network was obtained in r-program. Lineage specific social

distance matrices for the cases in the study were filtered and our analysis was focused on lineage four genotype, the most prevalent lineage in Uganda. The extracted matrix for cases was arranged according to the dates of symptom onset as per the time interval kernel. By transforming this social distance matrix through its inverse, we obtained a social kernel, K^S. This work was done using r- program (version 3.6.0)

Generating genetic kernel from genetic distance matrix: As previously mentioned, to generate genetic distance based pairwise probabilities linking cases in the study, we used WGS SNP analysis to identify differences among lineage four specific MTBC pathogens. A SNP difference matrix for cases was obtained by importing aligned sequences into Geneious software (version 2019.2.1). Yielding a genetic kernel was done by transforming the genetic distance matrix through its inverse in r - program (version 3.6.0). The cases in the genetic kernel were arranged according to their respective dates of cough symptom onset to form a Genetic kernel, K^G.

Element-wise multiplication of kernels to generate a transmission matrix: To generate a transmission matrix (pairwise likelihood of transmission) among cases in the study, we considered information from date of onset of cough (time kernel), WGS SNP difference data (genetic distance kernel) and social network distance (social kernel) in a case of independent contribution to transmission. The elements of the likelihood matrix were found by element-wise multiplication of the three matrices assuming absence of correlation between these quantified distances (Figure 4). Kernels were multiplied in r-program as per the formula below;

Transmission matrix/ Kernel, $V = K^T * K^S * K^G$

The super-indexes denote kernels based on time (T), social (S) and genetic (G) data, respectively. By assuming that all cases have been found (the complete outbreak has been

observed), rows were scaled to add up to 1 by dividing each row of the matrix with sums of row values in the transmission matrix.

Transmission tree/network: The transmission matrix was analyzed to generate a transmission tree. Briefly, the Maximum Likelihood (ML) of the vector of transmission probabilities ($v_{i,j}$) for each individual *i* having acquired the infection from any of the members *j* in the study was calculated. Since the ML of probabilities is 1, we identified the largest element in the vector for each individual and replaced it with 1. By replacing all the remaining values with 0, we generated a sparse matrix that was used as a directed adjacency matrix for the construction of a directed transmission tree to identify who infected whom among transmission clusters. Using this adjacency matrix, we plotted a directed tree using i-graph version 1.2.4.1 in r- program version 3.6.0.

Reproduction Number: The transmission matrix was analyzed to generate reproduction numbers for cases that infected others in the network. Briefly, the Reproductive number of a members (*j*) in this network refers to the number of cases (*i*) that probably got their infection from j. Considering that the rows in the transmission matrix represent a vector of probabilities ($v_{i,j}$) for each individual i having acquired the infection from any of the members j in the study. Considering a ML of probabilities as 1, we identified the largest element in the vector for each individual and replaced it with 1. By replacing all the remaining values with 0, we generated a sparse matrix with 1 and 0. The reproduction number was calculated by obtaining the column sum for each member j as per the formula below (Newman, 2010).

$$R_j = \sum_{i=1}^{n} v_{i,j}$$

Equation 6: Reproduction number formula

To calculate the overall network reproduction number, we added the column sums R_j and divided the total by the total number of column sum with $R_{j>0}$.

Identifiability Score: The transmission matrix was analyzed to generate an identification metric as a measure to evaluate the degree of certainty of identified links. The identification metric was calculated based on the assumption that the infection of susceptible individual *i* by case *j* is a Bernoulli trial with Variance; $var_{i,j} = v_{i,j}(1- v_{i,j})$

Variance was used as the measure of how accurately subject j in the observed population may be identified as the person who infected subject *i*. The sum of the variance for subject *i* then was considered as a measure of how accurately any of the subjects observed may be identified as having been the source of infection to subject *i*;

$q(i) = \text{sum}\ (var_{i,j})$

The quantity $1-q(i)$ was used as a score of the degree of identification of the infection parent of subject i. When there is only one non-zero element in row *i* of the transmission matrix, the parent node is perfectly known, and $1-q(i) = 1$ indicating perfect identification. Using the formula below, we calculated the minimum possible identifiability score if the probability of infection from all 61 potential j sources of infection was equal. This would represent complete ignorance about who infected subject i, among those for whom a link should be possible (i.e. those with non-zero $v_{i,j}$)(Teunis et al., 2013b).

$$1 - q(i) = 1 - \sum_{i=1}^{n_i} \frac{1}{n_i}\left(1 - \frac{1}{n_i}\right) = \frac{1}{n_i}$$

Equation 7: Minimum identifiability Score formula

We used the minimum identifiability score to calculate the percentage identifiability metric. We subtracted the minimum identifiability score from each score and divided the results

by 1- the minimum identifiability score and multiplied the result by 100 to generate the percentages shown by the percentage graph (Figure 18).

Applying the identifiability score on the transmission network: We applied the Identifiability score on the overall network using i-graph version 1.2.4.1 in r- program version 3.6.0. We generated trees for individuals with a corresponding identifiability score of >5%, >10% and >20% by constructing subgraphs of the main network that included individuals with the corresponding identifiability scores, respectively. This sub-graph was implemented first by dropping individuals having a score less than the set identifiability score and by re-construction of score based specific networks.

How we prevented cycles in the interpretation of the transmission network: In order for V to be translated into a proper transmission network, cycles were not admissible. Subjects could not mutually infect each other and larger cycles (i.e. loops involving more than two subjects) could not occur. By sorting all cases in order of their onsets of symptoms and preventing negative serial intervals, cycles were prevented from occurring. Based on serial interval, we resolved cases who were found to have infected each other based on serial interval timing. In this way, cycling in infection was eliminated and a directed transmission tree was generated successfully. In considering the transmission matrix, we assumed independence among sources of infection and that a source could infect many susceptible individuals as is the case with TB. A transmission link for subject i was therefore thought of as a random sample from a categorical distribution with probability vector v_i.

RESULTS

There were 123 index cases whose samples were collected and cultured on Lowenstein-Jensen media. DNA was extracted and WGS performed on 102 isolates of which 79 passed

quality controls and thus were included in this analysis. From the 79, there were 62 (78%) lineage four members, 14 (17%) were lineage two members, and 3 (4%) were lineage one members. This study analysis was MTBC lineage four specific, hence was limited to the 62 lineage four individuals.

Serial interval kernel: Analysis of the time kernel shows time difference to be relatively short with the longest time interval of 1,719 days (maximum probability of 0.08), mean time length of 362 days (mean probability 0.016) and a minimum time length recorded for 2 cases that started cough on the same day). The longest duration of time one subject coughed relative to other cases was significantly different resulting in a large probability value in the matrix (Figure 11).

Social distance kernel: The overall network had a total of 11,739 vertices (cases, control, and their contacts) from whom 62 MTBC lineage four diagnosed cases were filtered for this study. This study social network had a maximum social distance of 23 and a minimum social distance of 1 among cases. Notably, there were individuals who were not linked socially and upon inverse transformation of their social distances to obtain a kernel, we replaced their zero distances (Infinity values) with the minimum probability in the network. Since all individuals in the study were from the same peri-urban Division of Kampala, we assumed minimal social connections exist. Analysis of the social distance probability kernel shows a mean social linkage of 18 (probability of 0.0556) with the smallest social linkage occurring among first degree contact (maximum probability of 1.0) and a the longest social geodesic distance of 23 (minimum probability of 0.0435) among all possible pairwise links. There were 2 pairwise links with a maximum probability transmission of 1 arising from a social distance of 1 in the network. There

was an observable difference in the distribution of these probabilities with the second largest probability of 0.5 and majority having probabilities below 0.2 (Figure 12).

Genetic kernel: Analysis of the genetic kernel shows a mean SNP difference of 124 (probability of 0.008064739) with 7 pairs of the smallest SNP difference of 0-1 SNPs (maximum SNP difference of 1 for 5 pairs with a SNP difference of 0 and 2 pairs with a SNP difference of 1) and the largest SNP difference of 7,370 (minimum probability of 0.001356852) among all possible links (Figure 13). There were 7 pairwise links with a maximum probability of transmission of 1 arising from pairwise SNP difference of 0 or 1 among sequences (taking the inverse yielded a maximum probability of 1). There was an observable difference in the distribution of probabilities with7 pairs with a probability of 1 followed by 0.2 while majority had probabilities below >0.01 (>100 SNP difference).

Transmission matrix, V: In the transmission matrix, elements $v_{i,j}$ represent the probability that subject i acquired infection from subject j. Analysis of the matrix of transmission probabilities shows a mean transmission probability of 0.016 with a maximum transmission probability of 0.804 and a minimum transmission probability of 0.000008 among all possible 1,860 links. Compared to the time, social and genetic matrices (K^T, K^S ,K^G), the transmission matrix has a lower minimum probability of 0.000008 than 0.0014 for genetic kernel, 0.043 for the social distance kernels and 0.00006 for the time interval kernel (Figure 14).

Transmission tree/network: The overall transmission network tree shows a total of 15 clusters with the largest cluster comprising of 12 members followed by a 10-member cluster, and the smallest cluster comprising of 2 members. There were 26 network members who were identified as sources of infection in the network; the rest as either recipient (26 members) or both (10 members). The subject with the earliest and fifth ranked date of cough symptom onset appear

at the center of the two largest clusters with 11 and 9 members (Figure 15). Interestingly, the last individual to report symptoms in this network was probably infected by the fourth or sixth individual to report symptoms.

Reproduction Number: This analysis showed that there were 36 (58.1%) individuals who infected others (sources) in the network and 26 individuals who did not infect anyone else in this network. There are 3.2% of individuals with reproduction number 9 or 3 followed by 9.7% of individuals with a reproduction numbers of 2 with a majority (41.9%) having a reproduction number of 1.0 (Table 1). The members who did not infect anyone else in the population were 41.9% and the overall reproduction number of the network was 1.72 (Figure 16).

Identifiability Score: The identifiability score showed a population maximum identifiability score of 0.648, an average score of 0.0279, a minimum of 0.0221 and a minimum possible score of 0.0164. The majority of the scores were below 0.2 with only 11 links scored above 0.2 and only 2 scores above 0.5 (Figure 17 & 18).

Identifiability score adjusted networks/Trees: The transmission networks for individuals with a score >5%, >10% and >20%, shows a total of 10 clusters (12 links), 5 clusters (5 links) and 5 clusters (5 links), respectively (Figures 19, & 20 & 21). There were no links with graphical direction of infection inferred for scores >10% or >20. At score >5%, graphical direction of infection could only be inferred in 2 of the 12 possible links identified. Transmission networks for individuals when a score of >5%, >10% and >20% was applied showed an average SNP difference of 6.25 [range 0- 21], 5.2 [range 0- 14] and 5.2 [range 0- 14], respectively. Analysis of the social distance for scores >5%, >10% and >20% showed a social distance of 9.43 [range 6- No link], 8.5 [range 6- No link] and 8.5 [range 6- No link], respectively (Table 3 & 4). Even though the total number of network members decreased, there was no difference in the number

of clusters when a 10% score and 20% score was applied respectively. We therefore considered an identifiability score as ideal based on this data.

DISCUSSION

In this work, we deployed a basic outbreak investigation method to identify who infected whom in an endemic setting. To illustrate this methodology, we used data collected as part of a large cross-sectional study of TB transmission within social networks in urban Africa. The data we used was similar to that which is usually collected by TB national control programs, so our work may have pragmatic value for TB control. Among 62 MTBC lineage four isolated, we found a mean genetic distance of 5.0 SNPs, social geodesic distance of 9 steps between cases, and an average time difference of one year since start of cough among 10 individuals who were identified in transmission clusters. Considering a variance based degree of certainty cutoff of 20%, the mean genetic distance of identified links of 5.0 SNP difference was similar to that identified by Walker and colleagues in 2013 as stringent and confirmatory of TB transmission events (Walker, Kohl, Omar, Hedge, Del Ojo Elias, Bradley, Iqbal, Feuerriegel, Niehaus, Wilson, Clifton, Kapatai, Ip, Bowden, Drobniewski, Allix-Béguec, Gaudin, Parkhill, Diel, Supply, Crook, Smith, Walker, Ismail, Niemann, Peto, & Modernizing Medical Microbiology (MMM) Informatics Group, 2015).

The probability values of the time kernel were dependent on the time the case reported cough relative to other members and they were low relative to social network and genetic network probabilities. We did not expect time of cough to be a strong marker of infection given that date cough symptoms started was self-reported by patients thereby prone to recall bias. This low contribution by the serial time kernel to the overall transmission probability in our study was a new finding and needs to be investigated in context of other study populations.

Analysis of social distances among TB positive cases from this peri-urban Uganda community suggested that there were 1,006 (52.1%) socially unlinked cases probably suggestive of a smaller role social networks play in the transmission of TB. This finding is consistent with work done by others who have not found many linkages in TB cases based on their social networks alone (Walker et al., 2014). There were however few first-degree contacts; a proxy for social closeness (0.2% (4/1860), probability = 1). The observed large difference in the probability of being socially linked from the maximum to the second largest (probability = 0.5) confirms that whereas there were first degree contact closeness among these cases, this was rare which underscores the need to study the role of distant relationship such as community wide relationships in TB transmission.

Analysis of the genetic distances in this TB endemic community setting in peri-urban Uganda showed 0.8% (probability = 1) pairwise links with an extremely high probability of being linked genetically which is confirmative of genetic links. We observed an 80% drop in the probability of genetic links between the maximum to the second highest ranked link which probably confirms MTBC intra-genetic variability attributable to transmission accruing from members of the same sub-lineage, genetically distant from the main transmission network. This probably underscores the need to use this method to study members within a limited set number of SNP differences.

There were observable transmission links when serial time, genetic distance and social distance were combined and analyzed simultaneously. The distribution of the synthesized transmission probabilities shows variably relative to the component matrices. This is representative of the expected homogeneity effect of this method to incorporate and account for different sources of data. By this, we accounted for scenarios where individuals identified to

have strong genetic links but without documented social links. This observed homogeneity of the network is plausible given the variability of factors determining TB transmission. This model can equally be extended by giving weights to particular kernels in a quasi-Bayesian frame work or by introducing more kernels where pertinent data such as geo-spatial characteristics are available (Teunis et al., 2013b).

The different reproduction numbers identified among MTBC lineage four diagnosed cases suggests that this method could detect individuals and the secondary numbers of cases they infected .We found a difference in degree of certainty placed on links among cases based on the identifiability score metric. Whereas this score reflected a relatively low level of certainty placed on the assumed links, it captures our current uncertainties in the biology of infection and the assumptions required to infer a transmission links by models of transmission (Table 4). We propose the use of the 20% identifiability score based on our methodology as a stringent cut-off for detection of TB transmission events in an endemic setting. The use and interpretation of links among cases above a 20% certainty score cutoff in our study population highlights clusters most likely to belong to the same epidemic with a shared index case. We suggested that this score can be used to pin point clusters on ongoing transmission to public health for effected localized control measures based on cluster members.

The main limitation of this method is that it relies on an outbreak investigation method whose main assumption is that all (or most) cases are identified during a given period of time. By adapting it to an endemic setting where there are likely a mixture of outbreaks going on at any one time, cases may have arisen from different outbreaks, a challenge not accounted for in the analysis. As a way to limit the effect of this drawback, our study was conducted among residents of the same division of Kampala district, a confined geographical area. The fact that

our social network data was characterized by large geodesic distances and unlinked cases is probably proof that our cases arose from different epidemics with different index cases. To tackle this challenge, we assumed a minimal level of social connection existed where no links were identified among the study patients that resided in the same geographical location and we imputed the largest geodesic distances where links were absent. This therefore limited our method to studying individuals of a defined geographical area. However, the application of the certainty score was key since it most likely highlights clusters belonging to the same epidemic thereby removing links which have low transmission probabilities from the resultant network.

Another limitation of our study is that it did not include children or hospitalized individuals who were probability committed to health facilities. By not including children who are known proxies for recent infection, we limited the generalizability of this approach. Another limitation is that whereas we assumed transmission among linked cases in our study, there is no laboratory evidence for primary progressive disease and no prior status of infection was known for our study subjects. Lastly, just like we did not whole genome sequence all the isolates in the study, not all countries currently sequence their MTB samples which may limit the global application of this approach.

When considered individually, WGS or social network/epidemiological approaches alone cannot account for transmission events in most studies. In this study, we have advanced and optimized a basic outbreak investigation method to identify transmission links among TB cases in a Sub-Saharan Africa setting. This study uses an approach that combines epidemiological and WGS data with potential for extension to routinely collected data by national control programs. With evidence showing that person-to-person transmission may be the primary driver of the TB epidemic in an African endemic setting, we have used multiple data sources to identify

transmission links among TB cases. This method can be used by National TB control program to identify and monitor outbreaks in an endemic setting.

TABLES AND FIGURES

TABLES

Table 1: Frequency of Reproduction numbers

Reproduction Number	Number	Percentage
0.0 < R	26	41.9
1.0 = R	26	41.9
2 = R	6	9.7
3 = R	2	3.2
9 = R	2	3.2
TOTAL	62	100

Table 2: Distances of identified links for Identifiability score >20%

**For scenarios where direction of infection could not be specified among links, we used time since cough symptom onset to infer who infected whom. *In the social network, the least probability (0.043) for individuals with the largest social distance was used in cases where no connection (Inf) was recorded.

Paired Link (forward in time)*	Social distance (probability)	Genetic distance (probability)	Time difference (probability)
Case 6 →62	6 (0.167)	14 (0.071)	1092 (0.030)
Case 10→12	8 (0.125)	1 (1.000)	66 (0.003)
Case 11→16	Inf (0.043)**	1 (1.000)	124 (0.006)
Case 41→48	9 (0.111)	0 (1.000)	100 (0.006)

Case 47→59	11 (0.091)	0 (1.000)	289 (0.016)

Table 3: Distances of identified links for Identifiability score >5%

**For scenarios where direction of infection could not be specified among links, we used time since symptom on set to presume who infected whom. *In the social network, the least probability (0.043) for individuals with the largest social distance was used in cases where no connection (Inf) was recorded.

Paired Link (forward in time)*	Social distance (kernel probability)	Genetic distance (kernel probability)	Time difference in days (kernel probability)
Case 6 →62	6 (0.167)	14 (0.071)	1092 (0.031)
Case 10→12	8 (0.125)	1 (1.000)	66 (0.003)
Case 11→16	Inf (0.043)**	1 (1.000)	124 (0.006)
Case 41→48	9 (0.111)	0 (1.000)	100 (0.006)
Case 47→59	11 (0.091)	0 (1.000)	289 (0.016)
Case 28 →35	Inf (0.043)**	0 (1.000)	44 (0.003)
Case 5→7	Inf (0.043)**	0 (1.000)	118 (0.003)
Case 19→54	Inf (0.043)**	6 (0.167)	514 (0.031)
Case 23→36	21 (0.047)	0 (1.000)	97 (0.006)
Case 21→8*	Inf (0.043)**	21 (0.047)	297 (0.019)
Case 18→30	3 (0.333)	12 (0.083)	105 (0.006)
Case 8 →30*	8 (0.125)	20 (0.05)	354 (0.013)

Table 4:Average distances for score adjusted networks

*In calculating the average social distance, individuals who were not connected were excluded.

Identifiability Score	TB Cases (links)	Mean Genetic distance [range]	Mean Social distance [range]*	Mean Time(days) [range]
>5%	22 (12)	6.25 [0- 21]	9.43 [6 - unlinked]	267 [44-1092]
>10%	15 (5)	5.2 [0- 14]	8.5 [6 - unlinked]	334 [66-1092]
>20%	10 (5)	5.2 [0- 14]	8.5 [6 - unlinked]	334 [66-1092]

FIGURES

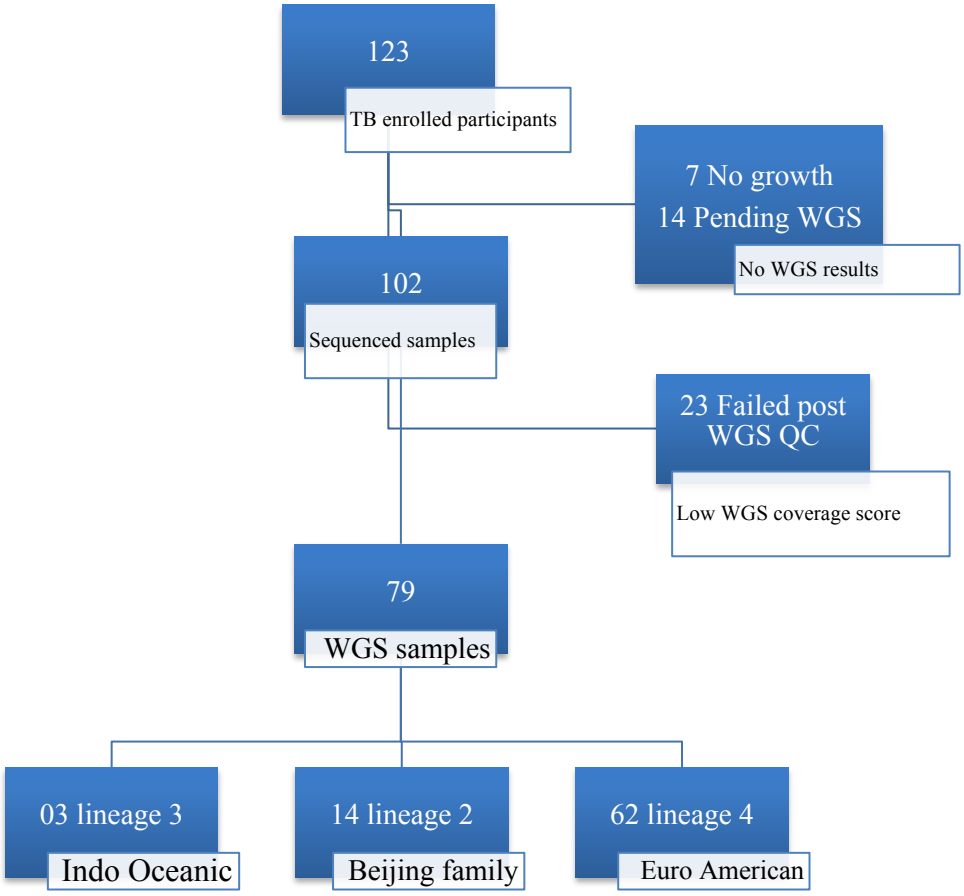


Figure 9: TB positive cases and lineage four samples

$$V = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix} = \begin{pmatrix} v_{1,1} & v_{1,2} & \cdots & v_{1,n} \\ v_{2,1} & v_{2,2} & \cdots & v_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ v_{n,1} & v_{n,2} & \cdots & v_{n,n} \end{pmatrix}$$

Transmission Matrix, *V*

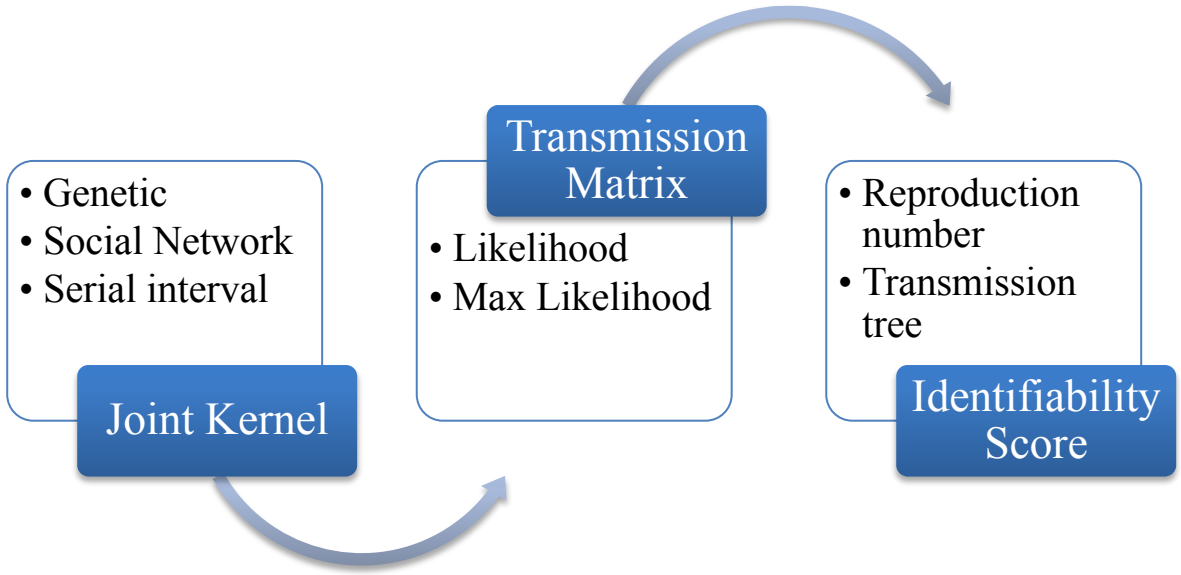


Figure 4: Summary of the methodology for this study analysis.

• Pictorial representation

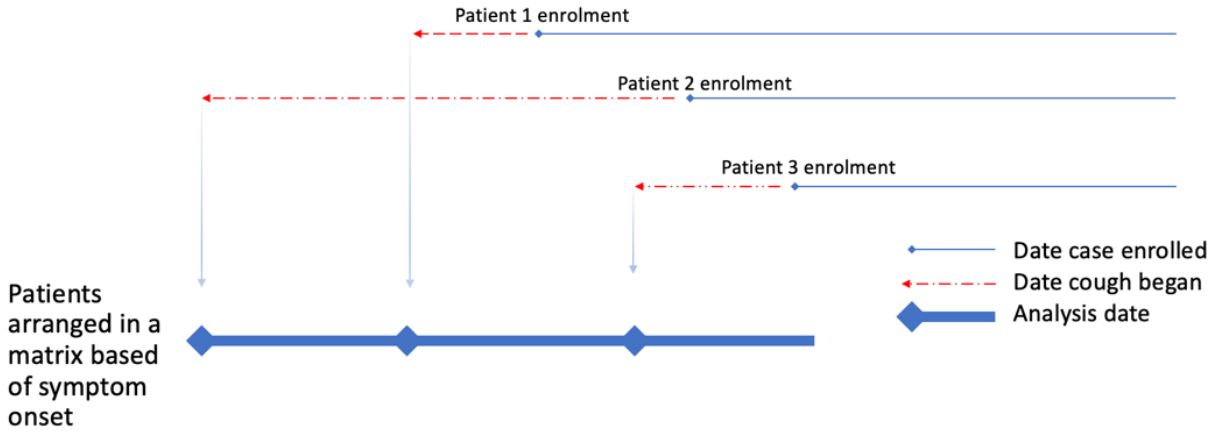


Figure 5: The method used to arrange the time kernel based on date of enrollment, reported days of cough and the date of symptom onset.

Pictorial illustration of SNP distance annotation in Geneious software

MTB Seq 1: …….. GGTTAAGAGCTCTGTGAAAG…..
MTB Seq 2: …….. GGTTAAGAGCTTTGTGAAAG…..
MTB Seq 2: …….. GGTTAAGAGCTTTGTAAAAG…..

SNP difference: Seq1-Seq2 = 1
SNP difference: Seq1-Seq3 = 2
SNP difference: Seq2-Seq3 = 1

	1	2	3
1	0	1	2
2	1	0	1
3	2	1	0

The standard procedure for generation of SNP difference in Geneious software version 2019.2.1.

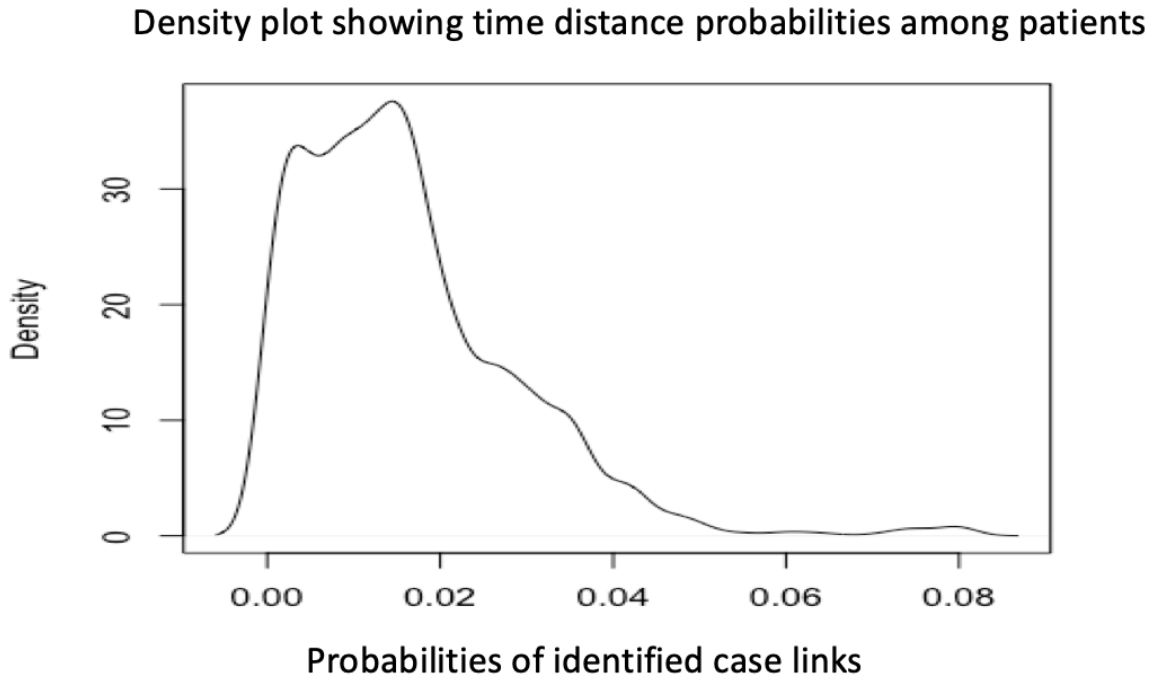


Figure 10: Probabilities based on difference in days since start of cough symptom

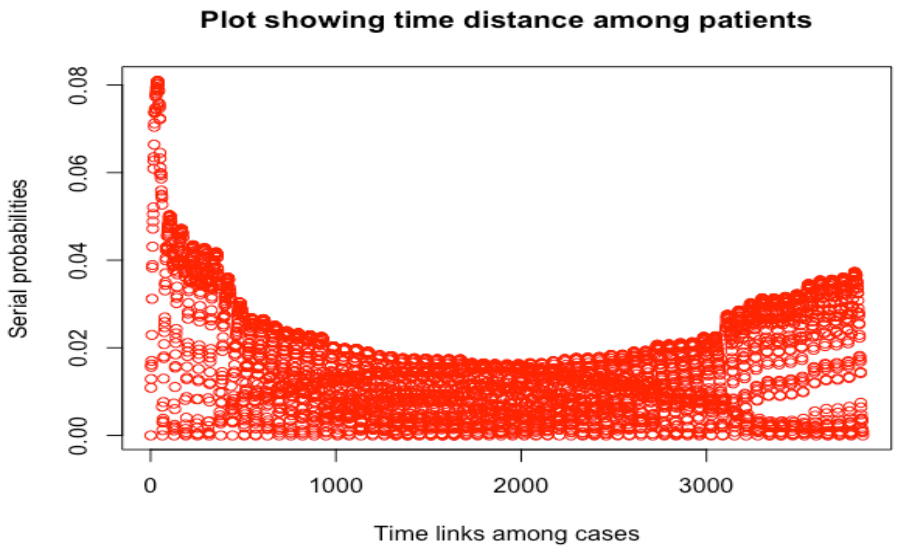


Figure 11: Time kernel probabilities on the y-axis and the pairwise link ID on the x- axis

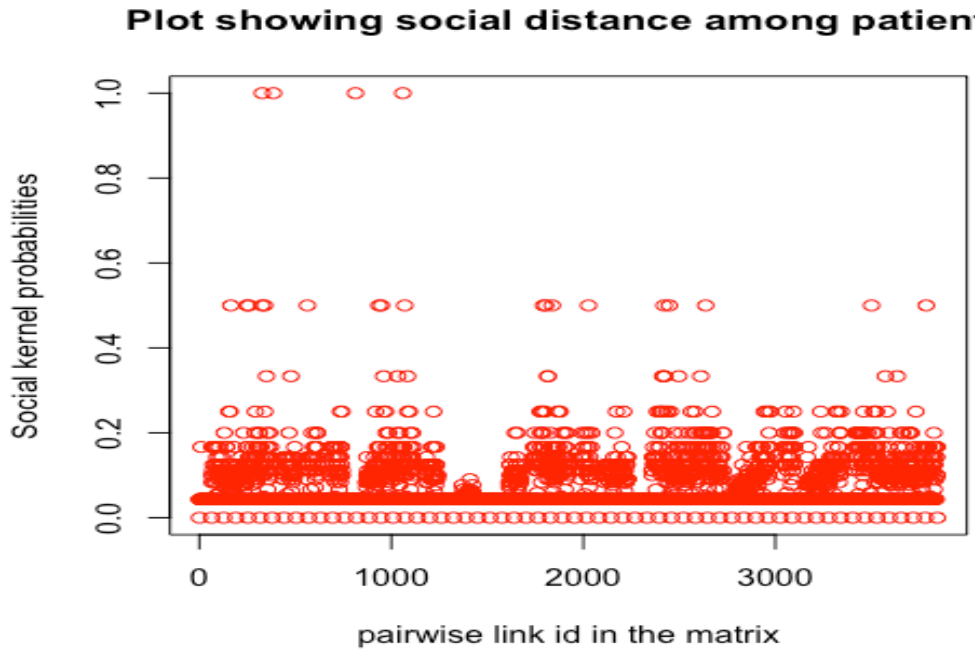


Figure 12: Social kernel (probability of links in a social network)

**the x-axis shows the pairwise link ID.

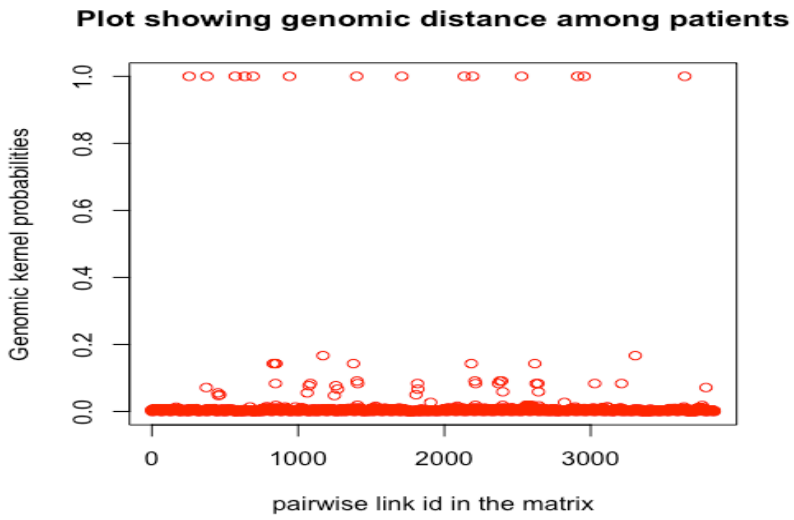


Figure 13: Genomic kernel (probability)

* pairwise link ID along the X-axis.

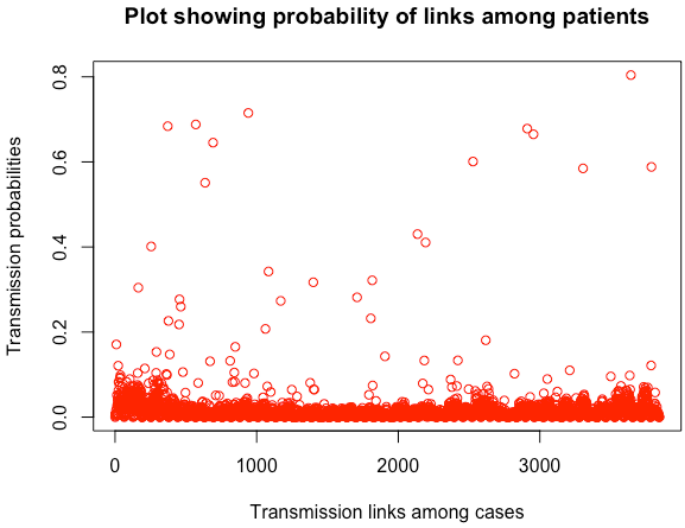


Figure 14: Transmission probability among cases

* pairwise links along the X-axis

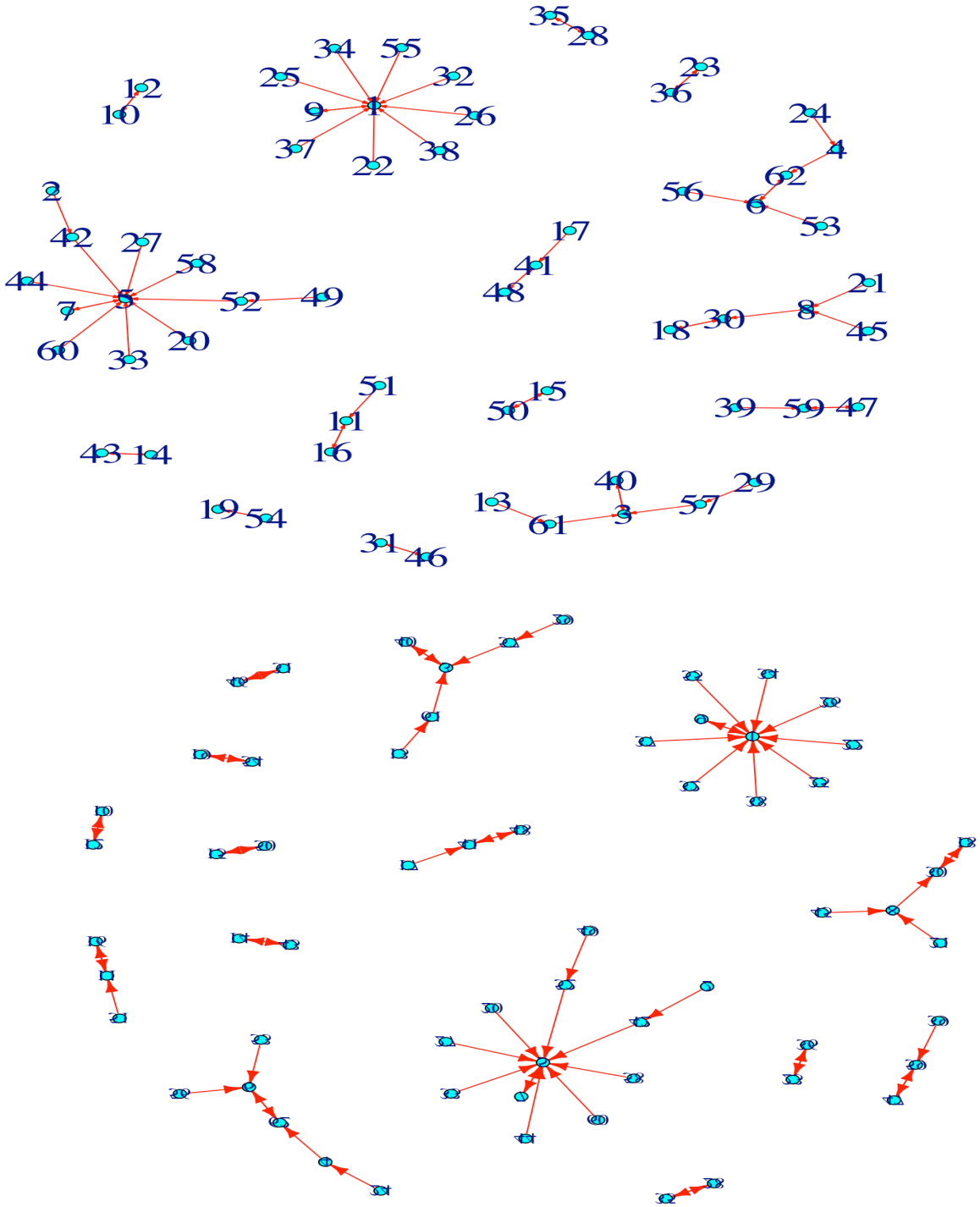


Figure 15: Transmission tree showing clusters of transmission among cases based on the Maximum Likelihood adjacency transmission matrix.

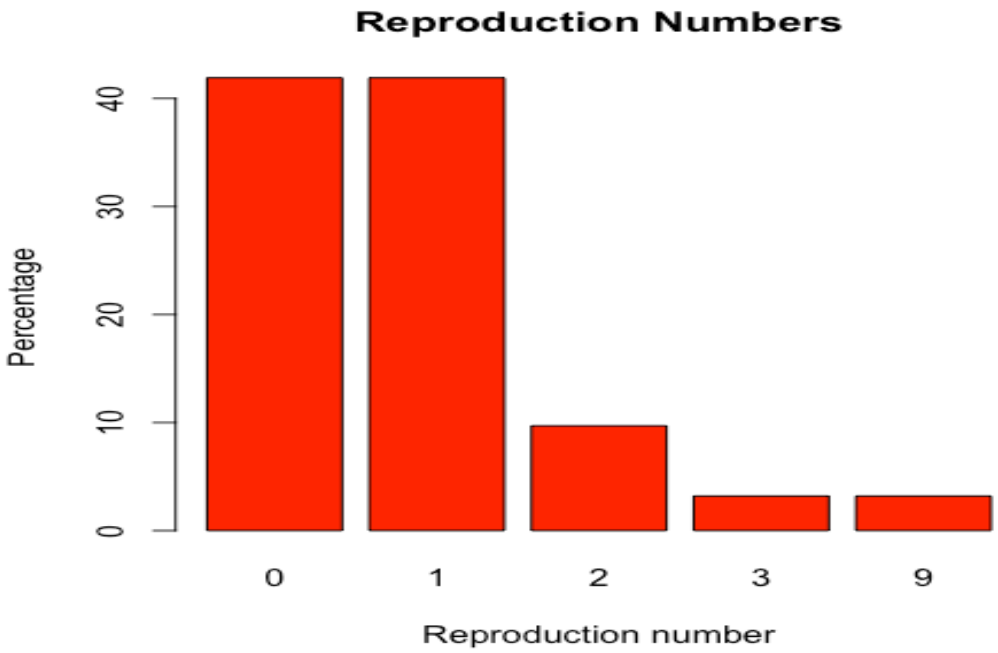


Figure 16: Reproduction numbers observed

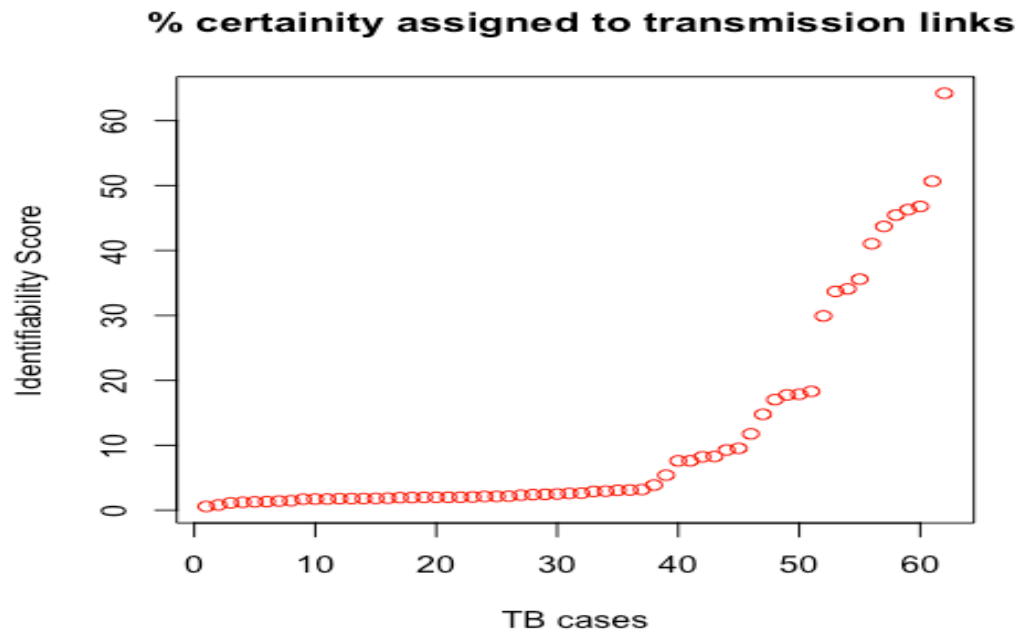


Figure 17: Identifiability score for each link inferred per case on the y-axis and the number of cases on the x-axis.

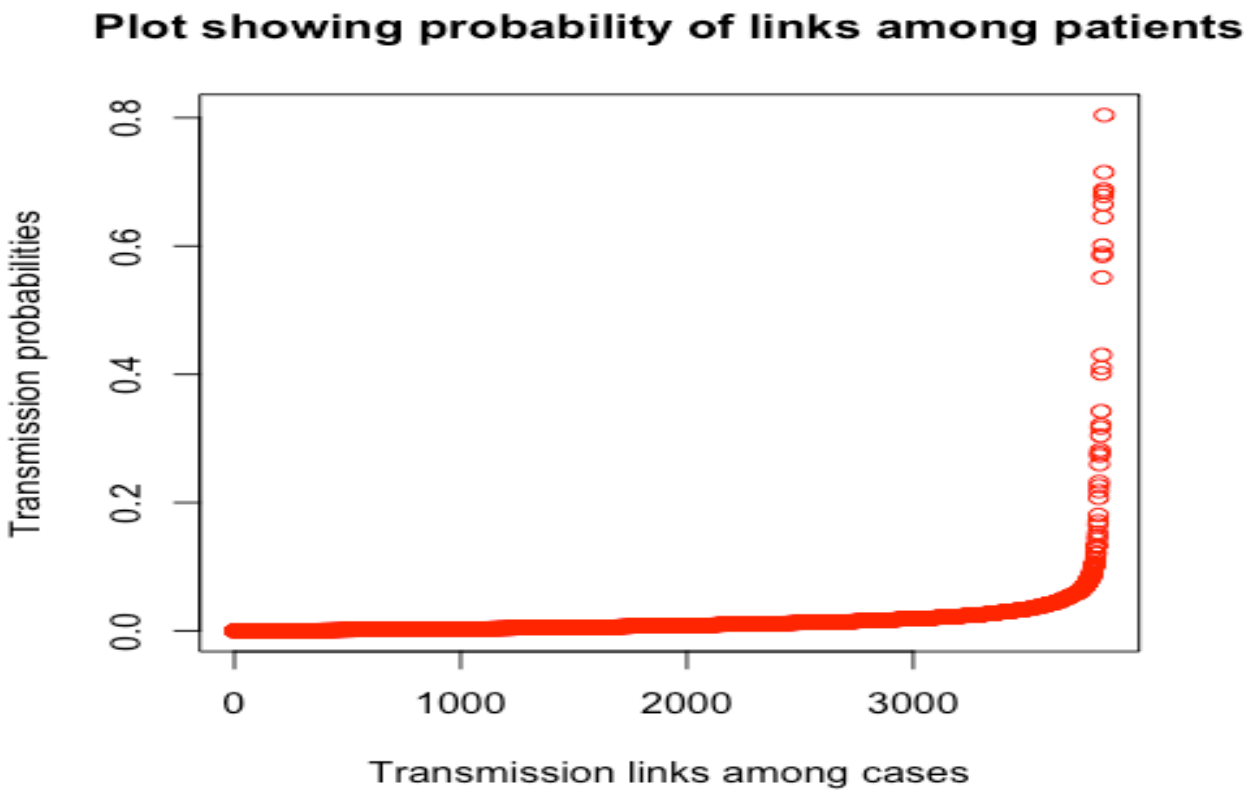


Figure 18: Percentage identifiability score

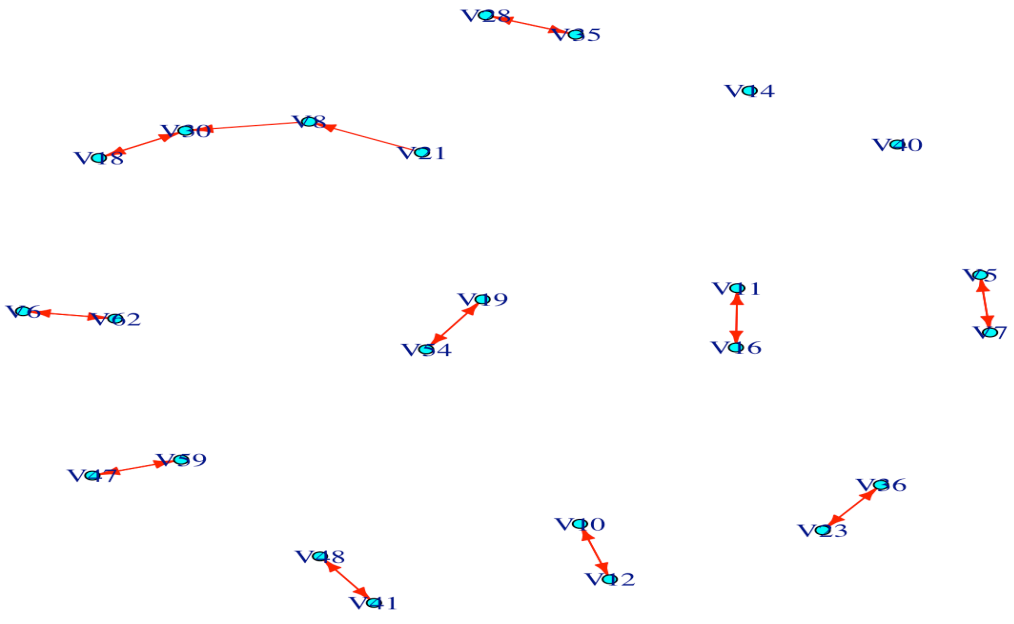


Figure 19: Links among cases with a score greater than 5%

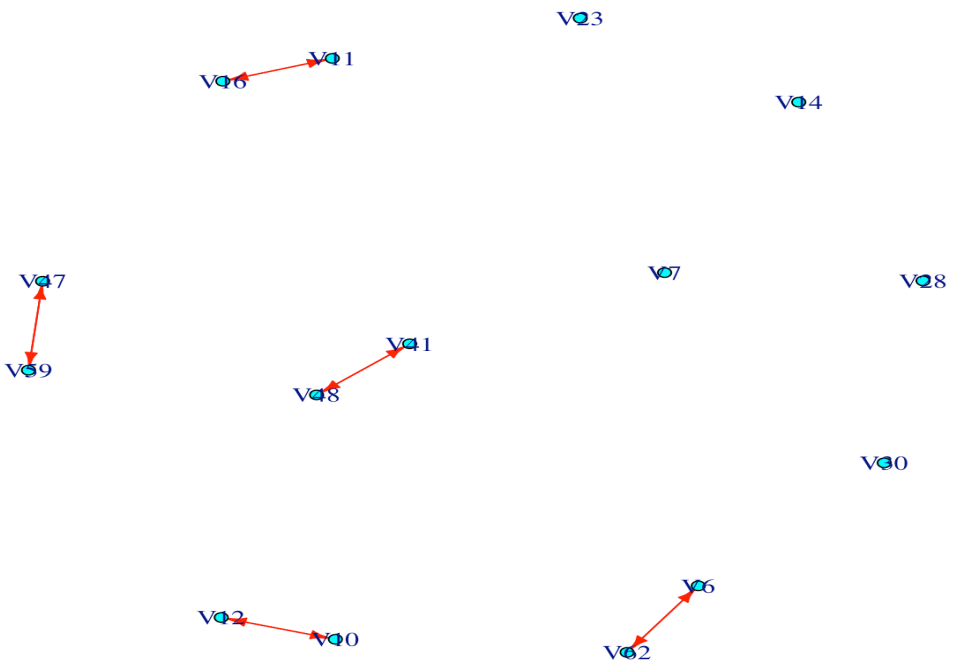


Figure 20: Links among cases with a score greater than 10%

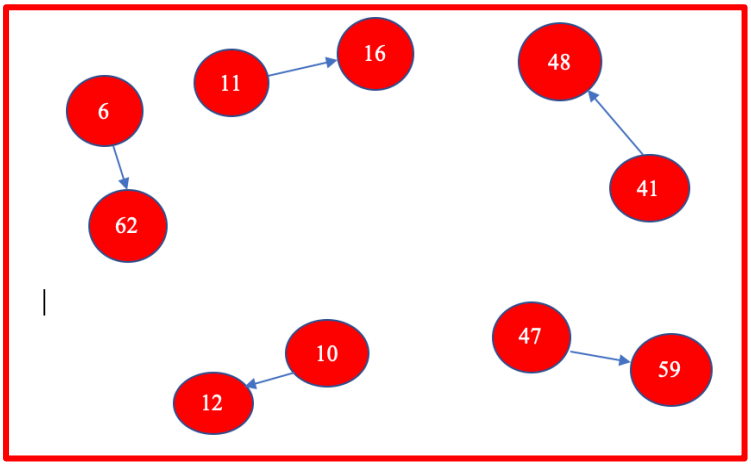


Figure 21: Clusters of transmission among cases based on a score greater than 20%.

Chapter 5

DRUG RESISTANCE CANDIDATE GENE MUTATIONS AND TRANSMISSION CLUSTERS IN PERI-URBAN KAMPALA, UGANDA

[1]Kirimunda S, Quinn F, Kakaire R, Chengwei L, Sekandi J, Whalen. To be submitted to American Journal of clinical microbiology

ABSTRACT

Background: This study addresses the problem of ongoing transmission of *Mycobacterium tuberculosis (Mtb)* in Sub-Saharan Africa. It is hypothesized that specific lineages that are geographically predominant in particular regions of the world have a higher chance of being transmitted in that region. We hypothesized that there were drug resistance candidate gene mutations that can serve as potential markers of being an ancestor among transmission clusters identified in the predominant MTBC lineage four genomes of tuberculosis cases from Kampala, Uganda.

Methods: This work studied MTBC Whole genome sequences defined transmission clusters from a cross-sectional study of 123 TB index cases residing in Rubaga division of Kampala, Uganda. An ancestor genome was defined as one in a pairwise SNP difference of less or equal to 12 by the SeqTrack algorithms. The TBprofiler pipeline was used to determine candidate gene mutations, genotypic drug resistance profiles, and phylogenetic classification for all study participant genomes regardless of phenotypic drug resistance status. The primary exposure was the number of mutations in each candidate gene, the outcome was being an ancestor (Yes = 1, No = 0). We performed a logistic regression analysis where the dependent variable was ancestry status and mutations in drug resistance genes the exposure.

Results: The distribution of lineages in the study population was 78 % (62/79) lineage 4, 18 % (14/79) lineage 2, and 4 % (3/79) lineage 3. Among the 62 participants with lineage four, 13 genomic clusters by 11 ancestors genomes within a pairwise SNP difference of less or equal to 12 of another isolate were identified. Ancestor status was inversely associated with carrying a

mutation in the *rrs* and *rrl* genes and positively associated with carrying mutations in *gyrA, ribD* and *ethR* genes.

Conclusion: We identified MTBC lineage four strain-specific drug resistance candidate gene mutations which could be used to develop SNP-typing assays to rapidly and inexpensively identify participant genomes that are most likely to be transmitted in an endemic Ugandan population, a potential marker for tracking *Mtb* transmission.

BACKGROUND

There were over 10.0 million cases of tuberculosis (TB) in the world in 2018, a number that has been relatively stable in recent years. The burden of disease varies enormously among countries, from fewer than 5 to more than 500 new cases per 100,000 population per year with most cases occurring in the WHO regions of South-East Asia (44%), Africa (24%) and the Western Pacific (18%) (WHO, 2019). In 2018, TB incidence in Uganda was 200 cases per 100,000 individuals compared to 202 per 100,000 in 2015 (WHO, 2019). The Global Health End TB strategy targeted to reduce TB incidence by 20% by 2020, a target that will not be achieved.

Based on epidemic theory and what is known about TB transmission, epidemics continue to occur when one index case is replaced by one or more cases (Whalen, 2016). Therefore, harnessing available resources to prevent new cases of disease remains crucial in elimination of TB and remains a key area of research. One approach is the use of current state-of-the-art phylogenetic methods for the reconstruction of pathogen genealogies (Cottam et al., 2008; Drummond & Rambaut, 2007; Grenfell et al., 2004; Templeton, 1998) based on the reconstruction of ancestries between hypothetical common ancestors (most recent common ancestor) and sampled isolates. Unfortunately, such approaches may be inappropriate when ancestors and their descendants are both present in the sample analyzed, as is likely in a sample of isolates drawn from typical public health TB program. Phylogenetic methods consider that sampled strains are all tips of an unknown genealogy, making it impossible for a sampled strain to be (directly or indirectly) ancestral to another sampled strain. By definition, these methods are thus unable to uncover ancestries between sampled isolates, and can fail to reconstruct the transmission tree of a pathogen.

Global genome-based phylogeny of the *Mycobacterium tuberculosis* Complex (MTBC) based on previously published data (Bos et al., 2014) shows that MTBC comprises of seven human-adapted lineages and several lineages are adapted to various wild and domestic animals. The human adapted lineages (L2, 3 and 4) share a genomic deletion which is *M. tuberculosis (Mtb)* specific referred to as deletion 1 (TBD1), referred to as modern lineages (Brosch et al., 2002). It is hypothesized that specific lineages are geographically predominant in particular regions of the world which may confirm adaptive evolution of particular MTBC lineages to be better transmitted in given geographical regions. MTBC L4 is geographically the most widespread cause of human TB. Genome based phylogeny further shows that L4 can be further subdivided into at least ten sub-lineages and that some of these sub-lineages are geographically restricted, corresponding to ecological 'specialists', and others are globally distributed 'generalists' (Gagneux, 2018). Among these sub-lineages, a specialists L4 Uganda genotype sub-lineage remains ecologically predominant in Uganda (Gagneux, 2018).

In this study, we adopted a method suggested by Jombart and colleagues to infer ancestry relationships among sample isolates. This approach uses the SeqTrack algorithm to reconstruct genealogies of sampled haplotypes or genotypes for which a collection date is available. By basing our analysis on genealogies, and not phylogenies, we directly address the concerns of phylogenetic analysis in inferring transmission (Figure 22). This approach proved more efficient than phylogenetic approaches for reconstructing transmission trees in densely sampled disease outbreaks (Jombart, Eggo, Dodd, & Balloux, 2011). This algorithm requires Whole Genome Sequences (WGS), Single Nucleotide Polymorphism (SNP) difference table, a genome mutation rate, date of infectiousness and names/identity of samples under analysis.

In the past, the problem of drug-resistant *Mtb* was primarily attributed to the *de novo* development of resistance linked to patient non-adherence and to the selection of drug resistant mutants during inadequate therapy (Dye, Williams, Espinal, & Raviglione, 2002). However, it has now become clear that the global epidemics of MDR-TB and XDR-TB are driven by a combination of *de novo* development and direct transmission of drug-resistant strains (Manson et al., 2017). In fact, it has been documented that person-to-person transmission is the primary driver of epidemics of extensively drug-resistant *Mtb* in Africa (Shah, Auld, Brust, Mathema, Ismail, Moodley, Mlisana, Allana, Campbell, Mthiyane, Morris, Mpangase, Van Der Meulen, et al., 2017). Understanding of drug resistance markers therefore offers an opportunity to study transmission since a bulk of the new cases of drug resistance in endemic settings is transmitted and not acquired. In addition, there is a notable emphasis on the problem of drug resistance evidenced by the fact that a majority of recent MTBC genomic sequences deposited in public databases are from MDR TB and national drug resistance surveys. Harnessing the increased volume of MTBC genomic sequences provides a novel opportunity to explore the question of transmission in context of drug resistance (Lukoye et al., 2013). Drug resistance is governed by multiple selective and neutral evolutionary forces acting concomitantly on the bacteria both within and between patients (Gagneux, 2018), yet little is known on the relative importance of these forces in different epidemiological settings.

In the past, traditional genotyping methods characterized by their low resolution nature were not sufficient to distinguish MTBC strains in populations with a high burden of TB which made it difficult to assess the degree of transmission let alone drug resistance profiling when required (Gardy et al., 2011b; Zappala et al., 2018). However, the application of WGS technologies has advanced resistance prediction, outbreak detection, and genomic surveillance of

MTBC (Merker et al., 2018). Several automated pipelines for resistance determination in MTBC genomes have been developed. These include; PhyResSE (Feuerriegel et al., 2015), CASTB (Iwai, Kato-Miyazawa, Kirikae, & Miyoshi-Akiyama, 2015), and TBProfiler (Coll et al., 2015) that can be run on the web, as well as two local software based, Mykrobe Predictor TB (Bradley et al., 2015) and KvarQ (Steiner, Stucki, Coscolla, Borrell, & Gagneux, 2014) all of which possess varying strengths and weaknesses (Schleusener, Köser, Beckert, Niemann, & Feuerriegel, 2017).

In this study, we used the TBprofiler pipeline to determine candidate gene mutations, genotypic drug resistance profiles and phylogenetic classification for all study participant genomes regardless of phenotypic drug resistance status. The TBprofiler pipeline searches for small variants and big deletions associated with drug resistance in genes or variations on a gene that may relate to drug resistance as defined by a distinct biological pathway or findings from previous studies (candidate gene). We examined whether TBprofiler defined drug resistance candidate gene mutations were associated with being an ancestor as defined by the SeqTrack algorithms among residents of peri-urban Kampala, Uganda. We hypothesized that there were drug resistance candidate gene mutations that can serve as potential markers of being an ancestor among transmission clusters within a pairwise SNP difference of less or equal to 12.

METHODS

Study setting, population, data collection and Whole Genome Sequence database:

The parent study was a cross-sectional study of 123 TB index cases residing in Rubaga division of Kampala, Uganda and their contacts, called COHSONET (The Community Health and Social Network Study) conducted from 2012 to 2016 (Kakaire et al., 2018, Castellanos et al., 2018). Briefly, second level egocentric contact networks were designed around 123 TB index cases and

124 randomly selected community controls that were frequency matched by age group, sex and parish. In this parent study, participants demographic, social, and clinical characteristics were obtained through interviews performed by trained study staff using standardized questionnaires (Figure 23). The date of diagnosis and the duration of cough were obtained during the initial interview of cases. Sputum samples were collected and grown in pure culture by a Mycobacteria culture laboratory to obtain one isolate per case from which genomic mycobacterial DNA was extracted.

DNA sequencing and generating SNP tables: Isolated genomic DNA of individual strains was sequenced on the Illumina platform (Illumina, San Diego, CA, USA) according to manufacturer instructions. Whole genome sequencing was performed at the CDC Atlanta microbiology laboratory using DNA extracted from sputum samples. Resulting FASTQ paired-end reads were processed using the CDC analysis pipeline for Mycobacterium tuberculosis complex genotyping from high throughput sequence to generate SNP tables and phylogenetic trees. The raw reads were trimmed and those that were <36bps were excluded. Next, Burrows-Wheeler Aligner (Li & Durbin, 2009) was used to map the reads to the H37Rv Mycobacterium Tuberculosis reference genome (GenBank accession number NC_000962.3). Burrows-Wheeler Aligner (BWA) is a software package for mapping low divergent sequences against a reference genome. Samples with <75% coverage or <25x depth of coverage were discarded. Variants were called using SAMtools mpileup (http://www.htslib.org) and the Genome Analysis Toolkit (GATK; Depristo et al., 2011) (https://software.broadinstitute.org/gatk/). Only variants supported by at least five reads, including one in each direction were accepted. Recognizing that repetitive regions frequently contain erroneous variant calls, regions annotated as 'repetitive' elements (e.g. PPE and PE-PGRS gene families), insertions and deletions (InDels), and consecutive variants in

a 12 bp window were excluded. Additionally, variants in drug resistance associated genes (i.e., mutations associated with drug resistance) were excluded. The resulting high-quality SNPs were concatenated to produce SNP alignments and SNP tables grouped according to the major MTBC lineages.

Whole Genome Sequence database of primary study and NCBI deposited sequences: The primary study was complete in terms of patient characteristics and variables and was the basis of this study. The results observed in the parent study lineage four analysis were replicated among lineage four NCBI downloaded sequences as part of sensitivity analysis. We downloaded genomic sequences of *M. tuberculosis* from the NCBI archive using the Sequence Read Archive tool kit (Leinonen et al., 2011). To access the sequences, we searched the National Center for Biotechnology Information (NCBI) website with a search term "Mycobacteria Tuberculosis Uganda Whole Genome". We considered sequences of all the four bio-projects from studies that included patients from Kampala (Table 9). We only included studies whose mycobacterial sequences were from sputum and were sequenced by the Illumina platform, similar to the parent study. In summary, we downloaded 59 samples from a bio-project that aimed at using WGS to characterize *Mtb* isolates from HIV-seropositive Ugandans with TB and CD4+ T cell counts of 0 – 1,150 cells/ul – research that was part of a grant to study community transmission of TB in urban Africa (Wampande et al., 2015). The second bio-project had 90 WGS from drug resistant *Mtb* isolates from Uganda. This study used WGS to complement the TB drug resistance national survey in Uganda (Ssengooba, Meehan, et al., 2016). The third bio-project aimed to understand the heterogeneity of *Mtb* isolates obtained from sputum and blood compartments concurrently from 13 patients. This third study used WGS to reveal mycobacterial microevolution among concurrent isolates from sputum and blood in HIV infected TB patients (Ssengooba, de Jong, et

al., 2016). The fourth study involved evolution of drug resistance by elucidating emergence and transmission of multidrug-resistant *Mtb* isolates using WGS; this study contributed 51 sequences (Clark et al., 2013). We excluded isolates with low phred quality scores. There were 168 isolates and of them, 129 isolates belonged to lineage 4 (Figure 24).

Transmission networks and identification of ancestors: To classify strains as ancestors, or not, we used SeqTrack algorithm to determine the ancestor and descendent sequences. Briefly, a SNP difference matrix was generated using Geneious software (version 2019.2.1) by importing aligned sequences as illustrated by (Figure 35: phylogenetic tree for 62 parent study samples). Since the mutation rate of the MTBC in clinical settings has been estimated at 0.3-0.5 substitutions per genome per year (Eldholm & Balloux, 2016), we used a non-conservative mutation rate of 0.5 substitutions per genome per year. The SeqTrack algorithm maximized the SNP difference to identify relations among cases. The study samples in this analysis were arranged according to their respective dates of cough symptom onset, a proxy for the onset of infectiousness (Figure 25 & 26). Briefly, we subtracted the number of days the cases reported to have coughed at the time of enrollment from the date of laboratory TB diagnosis to derive the original date of infectiousness. Based on the cough onset dates, samples were arranged from the earliest to the last date and this arrangement of cases was adopted as the subject IDs in the analysis. For cases in clusters where SNP differences were maximized as similar and ancestor genomes needed identification, the date of infectiousness was used by the program to identify true ancestor genomes. The output of the SeqTrack analysis were a classification of ancestor sequences, accompanying probability of links and a classification of the descendant genome sequence. Knowing that the number of SNP differences between genomes does not directly imply a probability of recent transmission (Walker et al., 2013b), the use of the 12 SNP threshold

for inferring likely transmission between a pair of TB cases as demonstrated under a longitudinal study setting by Public Health England (Walker et al., 2014) determined who was a true Ancestor among the ancestor- Non ancestor pairs identified by SeqTrack algorithms. In this analysis, an ancestor strain was defined as MTBC ancestor genome within a pairwise SNP difference that was less or equal to 12 with its corresponding Non-ancestor; all strains that did not meet this definition were considered non-ancestors.

Identification of mutations in candidate Genes: We used the TBprofiler pipeline (Coll et al., 2015) to identify candidate gene mutations, genotypic drug resistance profiles and phylogenetic classification of participants' MTBC WGS. Briefly, the TBprofiler pipeline checked study sequences for quality using FastQ (a step where poor quality sequences were eliminated) and aligned the sequences using Burrows-Wheeler Aligner (BWA). The aligned sequences variants were called using SAMtools, a suite of programs consisting of BCFtools for writing Variant Call files (VCF), filtering and summarizing SNPs and short indel sequence variants and LoFreq a fast and sensitive variant-caller for inferring SNVs and indels from next-generation sequencing data. The variants, in form of Variant Call files were then compared to a drug-resistance database by Walker and colleagues (Figure 31) (Walker, Kohl, Omar, Hedge, Del Ojo Elias, Bradley, Iqbal, Feuerriegel, Niehaus, Wilson, Clifton, Kapatai, Ip, Bowden, Drobniewski, Allix-Béguec, Gaudin, Parkhill, Diel, Supply, Crook, Smith, Walker, Ismail, Niemann, Peto, Davies, et al., 2015). The final output of this analysis was a report corresponding to each participant's MTBC genome.

Statistical analysis: Demographic characteristics of ancestor and non-ancestor sequences were compared using frequency tables, chi squared tests, Kruskal-Wallis chi-squared test and Fisher's exact test in R-statistical software. Frequencies of mutations in candidate genes were

tabulated for ancestors and non-ancestors. We calculated odds ratios and 95% confidence intervals (OR, 95% CI) for the association of ancestor status and candidate genes mutations using logistic regression ("Applied Logistic Regression - David W. Hosmer, Jr., Stanley Lemeshow, Rodney X. Sturdivant - Google Books," n.d.). We assessed for confounding, interaction and adjusted the associations for patient characteristics such as sex, smoking, duration of cough, HIV and genotypic drug resistance (a proxy for phenotypic drug resistance).

RESULTS

Study database: There were 319 TB cases whose genomic sequences were collected/downloaded and available for this study. There were 168 (52.7%) strains from 168 cases that were of good quality for analysis by the TBprofiler pipeline and were included in this study analysis. There were 129 (76.8%) of L4, 30 (17.9%) L3, 4 (2.4%) L2 and 3 (1.7%) L1 sequences among the 168 that were of good quality. Analysis was performed on 129 MTBC L4 isolates. The primary study L4 had 62 MTBC genome from 62 patients whose demographics and clinical characteristics (Table 6).

Characteristics of ancestor and Non-ancestor subjects of the study: In analyzing the 62 strains from the parent study with SeqTrack, we identified 11 ancestor strains and 51 non-ancestor genomes. The ancestors were slightly younger than non-ancestors (27.6 years versus 31.0 years). As observed in other studies, there were more males than females in both ancestors and non-ancestors (54.5% versus 45.5%, and 75.0% versus 25.0%, respectively). Both ancestors and non-ancestors reported cough that lasted longer than 3 weeks before enrollment into the study (90.9% versus 96.1%, P=0.45, respectively). Ancestors and non-ancestors had similar proportions of being a TB retreatment cases at enrollment (9.1% versus 9.0%, P= 1.0). Ancestors were less likely to report previous smoking than non-ancestors (9.1% versus 21.2%,

however ancestors did not report smoking at the time of enrollment compared non-ancestors (0% versus 11.5%, P= 0·35).

SeqTrack identified ancestors (Outcome variable): In analyzing the 62 strains from the parent study with SeqTrack, we identified 11 ancestor strains and 51 non-ancestor genomes. The identified 11 ancestors were from 13 clusters that were within a pairwise SNP distance equal to or less than 12 among L4 (18% vs 82%) (Figures 28 and 29, Table 6). Genotypically confirmed drug resistance was not different among ancestors and non-ancestors (9.1% versus 13.7%, P= 1.00). We observed a three group pairwise SNP arrangement with ancestor defined SNP difference between 0-12, a second increase from 18-37, and a third from 198-265 pairwise SNP differences (Figure 30).

Candidate gene mutations (Independent variable): There were a total of 32 candidate genes with varying gene lengths (Range: 561 – 3951 base pairs) whose mutations were identified in the study (Figure 33). The distribution of candidate gene mutations in the ancestors is broadly comparable to the distribution observed in non-ancestors (Table 7 and Figures 31 & 32). There were 157 gene mutations in ancestors compared to 760 mutations in non-ancestors. An average of 15.7 mutations per genome in ancestors compared to 14.6 in non-ancestors. The average mutations per genome did not differ among ancestor strains and non-ancestors (p = 0.86)

Four genes, *gyrA, rpsL, tlyA* and *embC* had mutations distributed in all ancestors but only *rpsL* and *tlyA* had 100% gene distribution in both ancestors and non-ancestors (Table 7, columns 3 and 5). The two most frequent candidate genes with mutations in the parent study ancestors and non-ancestors were *gyrA* (21% versus 18%), and *alr* genes (12.3% versus 13.8%).

Logistic regression results: At univariate logistic regression, ancestor status was positively associated with mutations in *gyr A* gene (OR=1.46, 95% CI 1.0-2.1), *ribD* gene

(OR=11.15, 95% CI 1.06-240.57), and *ethA* gene (OR=7.06, 95% CI 1.85-28.76) and inversely associated with mutations in the *rrl* gene (OR= 0.20, 95% CI 0.07-0.44) and the *rrs* gene (OR= 0.16, 95% CI 0.04-0.38). At multivariate logistic regression where we included all the genes that were statistically significant at univariate analysis, only mutations in the *rrl* gene (aOR= 0.23, 95% CI 0.08-0.54) and the *rrs* gene (aOR= 0.23, 95% CI 0.07-0.55) remained statistically significant (Table 8).

DISCUSSION

We analyzed MTBC Lineage four specific genomes from patients residing in a TB endemic division of Kampala, Uganda and found that 18% of the pairwise links among patients' genomes were within a SNP difference of 12. This is conventionally assumed to be real transmissions as previously demonstrated in a longitudinal study setting (Walker et al., 2013a). We studied isolates of 62 participants, observed 13 genomic clusters and 11 ancestors within a 12 SNP distance of another isolate. This was comparable to a study in the United Kingdom which assessed pairwise nucleotide differences in 247 patients with 84% of these patients that could not be genomically linked to another within a period of 6-year. This unlinked group was comparable to 82% of our study genomes that were considered unlinked by SeqTrack algorithm (Walker et al., 2013a).

Rather than relying on phylogenetic methods alone which consider that sampled strains are all tips of an unknown genealogy, making it impossible for a sampled strain to be ancestral to another sampled strain, by using the SeqTrack algorithm, we reconstructed transmission trees identifying ancestors based on time of infectiousness and SNP difference. We found a three group pairwise SNP arrangement in this population which may imply the role of MTBC genetic variability in transmission dynamics even within members of the same sub-lineage. This finding

remains to be studied with more participant genomes and among different sub-lineages. The distribution of lineages in our population was comparable to another genome study in the same study population that found 88 % (61/69) were infected with an MTBC strain of L4 strains confirming a high prevalence of L4 in the same study area (Wampande et al., 2015).

We observed a total of 13 ancestor-descendant clusters where two had 2 descendant genomes within a 12 SNP distance of another isolate, a representative proportion of 15 % (2/13) for infecting more than one individual in the study population. This proxy of superspreading is consistent with data showing that a few individuals infect more susceptible individuals with *Mtb* than others. In data from Victoria, Australia there were 9.9% super-spreading events similar to the 15% found in this work. This reaffirms that super-spreading events are responsible for a substantial majority of secondary infections and that heterogeneity of transmission and super-spreading are critical issues to consider in the design of interventions and models of TB transmission dynamics. (Melsew et al., 2019; Ypma et al., 2013).

Genotypic drug resistance in ancestors and non-ancestors (9.1% versus 13.7%) along with being a retreatment case, a proxy for phenotypic drug resistance (9.1% versus 14%) were not different. This is consistent with known facts that even though drug-resistance-conferring mutations in the MTBC are often associated with a fitness cost in the absence of the drug, some mutations show little or no cost, and these tend to be preferentially selected for in clinical settings (Gagneux et al., 2006; Sander et al., 2002). Moreover, compensatory evolution can overcome initial fitness deficits that are linked to particular resistance mutations (Comas et al., 2012). Also, epistatic interactions between mutations causing resistance to different drugs can lead to a reduction in the fitness cost that is associated with each individual mutation (Borrell et al., 2013). In summary, epistatic interactions between drug-resistance-conferring mutations,

compensatory mutations and the genetic background of the strain affect the biology and epidemiology of drug-resistant TB and remains a key area of research (Gagneux, 2018).

The two most frequent candidate gene with mutations in our study ancestors and non-ancestors were *gyrA* (21% versus 18%), and *alr* (12.3% versus 13.8%). The *gyrA* genes encodes for DNA gyrase which negatively supercoils closed circular double-stranded DNA in an ATP-dependent manner and also catalyzes the interconversion of other topological isomers of double-stranded DNA rings, including catenanes and knotted rings which catalyses ATP-dependent breakage, passage and rejoining of double-stranded DNA (Dejesus et al., 2017). Not much has been documented on the role of the (*alr*) gene in *Mtb* pathogenesis, however, it encodes for alanine racemase which provides the D-alanine required for cell wall biosynthesis by catalyzing the transformation of L-alanine to D-alanine (Strych et al., 2007). Four genes, *gyrA, rpsL, tlyA* and *embC* had mutations distributed in all ancestors but only *rpsL* and *tlyA* had 100% distribution in both ancestors and non-ancestors. This information in new and positively identifies housekeeping mutations in genomes found in our study population. The *rpsL* gene encodes for 30S ribosomal protein S12 which is involved in the translation initiation step. The *tlyA* gene encoded for 2-O-methyltranserase which methylates 16S and 23S rRNA and has contact-dependent hemolytic activity and is possibly involved in virulence (pore formation) (Johansen, Maus, Plikaytis, & Douthwaite, 2006). The *embC* gene encodes for integral membrane indolylacetylinositol arabinosyltransferase synthase which is involved in the biosynthesis of the mycobacterial cell wall arabinan and resistance to ethambutol (Belanger et al., 1996). Further study of these genes presents an opportunity for understanding the house keeping importance of their mutations especially the *rspL* gene which had one mutation in each of the 62 genomes studied.

In this study, being an ancestor was positively associated with carrying mutations in three genes; *gyrA*, *ribD* and *ethA*. The *ribD* gene encodes for a possible bifunctional enzyme riboflavin which is vital in the biosynthesis of diaminohydroxyphosphoribosylaminopyrimidine deaminase (a riboflavin-specific deaminase) which is critical in riboflavin biosynthesis (Sassetti, Boyd, & Rubin, 2003). Gene *ethA* encodes for monooxygenase which activates the pro-drug ethionamide to induced ethionamide sensitivity when overexpressed in *Mtb* (DeBarber, Mdluli, Bosman, Bekker, & Barry, 2000). The *gyrA* gene encodes for DNA gyrase which negatively supercoils closed circular double-stranded DNA in an ATP-dependent manner and also catalyzes the interconversion of other topological isomers of double-stranded DNA rings, including catenanes and knotted rings which catalyses ATP-dependent breakage, passage and rejoining of double-stranded DNA (Dejesus et al., 2017).

Whereas there are currently no direct biological links in the literature between the genes that were positively associated with being an ancestor and transmission in our study, all the drug resistance candidate genes identified are vital for the survival of *Mycobacterium* species. We know that although DR TB arises as secondary (acquired) due to poor treatment and management of cases by hospital systems in sun-Saharan Africa, it is now transmitted as primary drug resistance due to poor TB control measures (Shah, Auld, Brust, Mathema, Ismail, Moodley, Mlisana, Allana, Campbell, Mthiyane, Morris, Mpangase, van der Meulen, et al., 2017). These mutations is drug resistance candidate genes which are also relevant for the survival of the bacteria probably point to evolutionary fitness and hence plausibly foster the transmissibility of *Mtb*. Our finding needs further research since mutations in MTBC drug resistance candidate genes have recently been demonstrated to display epistatic relations with other genes when

subject to environmental and host factors (Coll et al., 2018). These finding call for further research into the genetic interactions of these mutations with environment and host factors.

Ancestor status was inversely associated with carrying a mutation in the *rrs* and *rrl* genes. The *rrs* gene encodes for ribosomal RNA 16S subunits and *rrl* with ribosomal RNA 23S subunits. These gene mutations are generally associated with different patterns of resistance or susceptibility to capreomycin, kanamycin, amikacin, and viomycin. For instance, a A1400G mutation of the *rrs* gene was identified in *Mtb* strains and high level cross resistance has been reported to rise from such a mutation since *Mtb* has one single copy of the rrs gene (Alangaden et al., 1998). This may imply that mutations in this gene make the organisms less fit for transmission probably explaining why it was negatively associated with ancestors. This is consistent with well stablished hypothesis that drug resistance strains of *Mtb* are less fit and unable to survive in a gene pool (Dye et al., 2002).

This study had several important limitations. The study population was restricted to outpatient health centers and did not include hospitalized individuals, so ascertainment bias could have occurred (Delgado-Rodríguez & Llorca, 2004), limiting the generalizability of our results. Second, whereas we assumed transmission among cases in our study, there is no laboratory evidence for primary progressive disease, besides no prior status of infection was known for our study subjects. Given that this was a cross-sectional study, we cannot prove the directionality of the infection, even if the method of choice used in this analysis relies on the time of cough symptom to infer directionality. Third, although we studied mutation in drug resistance conferring candidate genes in all study participant genomes, phenotypic drug resistance was not known. However self-reported retreatment status served as a proxy for phenotypic drug resistance. This self-reporting, along with other self-reported data bore a risk for

recall, and response bias as a potential for information bias (Delgado-Rodríguez & Llorca, 2004). Fourth, in identifying ancestors, we utilized a convenience cut-off of equal of less than 12 SNP which may understate the level of transmission (M. Murray & Alland, 2002; Van Soolingen, 2001).

A Fifth drawback is that we used H37Rv as the reference genome in the analysis of this work. This reference genome does not represent the entire genomic repertoire of MTBC, including "core genes" shared by all strains and "accessory genes" that are not present in all strains (Vernikos, Medini, Riley, & Tettelin, 2015). The use of pan-genomes has previously been proposed as a better reference in the study of mycobacterial transmission (Yang et al., 2018). Lastly, we observed a limited number of outcomes (11 ancestors) that limited the robustness of potential statistical models used. Because of this, our work was limited to univariate models and correction for multiple testing could not successfully be employed. This explains the wide confidence intervals around the measures of effect from the logistic models.

As part of sensitivity analysis, univariate logistic regression analysis of the study sample genomes along with NCBI downloaded sequences collected in the same period of the study was performed. The genes identified in this study in context of transmission clusters and drug resistant mutations remained significant in the larger sample analysis which confirm generalizability of such a marker for transmission.

In this exploratory proof of concept study, we studied drug resistance conferring candidate gene mutations as a reason for the successful transmission among L4 isolates. We show that genomes that were readily transmitted are positively associated with mutations in *gyrA, ribD* and *ethR* genes that are vital for the survival and virulence of the bacteria. These data suggest that strain specific virulence factor variations could be important for the successful

transmission of L4 ancestor genomes in Uganda (Folkvardsen et al., 2018). If combined with strain-specific SNP typing and targeted WGS, the mutations identified in this study could be deployed to investigate *Mtb* clusters in an outbreak or in retrospective settings. In conclusion, we have identified L4 strain-specific drug resistance candidate gene mutations which could be used to develop SNP-typing assays to rapidly and inexpensively identify participant genomes that are most likely to be transmitted in an endemic Ugandan population, a potential marker for tracking *Mtb* transmission (Stucki et al., 2012).

TABLES AND FIGURES

TABLES

Table 5 Demographic and clinical characteristics of the study population

*P-values are by the fisher exact method.

Variables	Ancestors, n (%)	Non-ancestors, n (%)	P value*
Characteristics			
BMI (standard deviation)	20.9 (3.7)	19.6 (3.0)	0.28
Mean age, (standard deviation)	27.6 (7.5)	31.0 (9.8)	0.22
Sex			0.27
Male	6 (54.5)	38 (75.0)	
Female	5 (45.5)	13 (25.0)	
Mtb Sub-lineage*			0.90
L4.1	1 (9.1)	2 (3.9)	
L4.2	0 (0.0)	1 (2.0)	
L4.3 (LAM)	2(18.2)	9 (17.6)	
L4.4	0(0.0)	2(3.9)	
L4.6 (Uganda)	8(72.7)	29 (56.9)	
L4.7	0(0.0)	1 (2.0)	
L4.8	0(0.0)	4 (7.8)	
Genotypic Drug resistance*			1.00
Yes	1 (9.1)	7 (13.7)	
No	10 (90.9)	44 (86.3)	
TB Retreatment status			1.00
Yes	1 (9.1)	9 (14.0)	
No	10 (90.9)	45 (86.0)	
Smoking			0.35
Current smoker	0 (0.0)	6 (11.5)	
Former smoker	1 (9.1)	11 (21.2)	
Never smoked	10 (90.9)	35 (67.3)	
Other medical conditions			0.25
Asthma	1 (10.0)	0 (0.0)	

HIV	0 (0.0)	6 (85.7)	
Diabetes	0 (0.0)	1 (14.3)	
Alcohol			0.74
Yes	4 (36.4)	24 (47.1)	
No	7 (63.6)	27 (52.9)	
Duration of cough			
2-3 week	1 (9.1)	2 (3.9)	0.45
>3 week	10 (90.9)	49 (96.1)	

Table 6 Thirteen pairwise links of genomes classified as ancestors by the SeqTrack algorithm above the red line and their corresponding descendant genomes (table is truncated at a SNP difference of 95).

Ancestor Genome	**Descendant Genome**	**SNP difference**
20242	20269	0
21083	21104	0
20951	21128	0
21560	21609	0
21619	22193	0
20751	20814	1
20436	20495	1
20851	22038	6
20547	20951	7
20547	20440	7
20951	21179	11
20818	21074	12
21618	21792	12
20818	21535	13
20191	22283	14
20599	20818	18
21172	21617	37
20951	21192	54
21192	22264	56
21192	21466	64
20814	21847	74
21192	21619	75
21192	21126	88
21192	21618	93
21381	21083	95

Table 7 Mutations in drug resistance candidate genes in ancestors vs non-ancestors and their distribution among members of each group

Gene n = 32	Mutations in Ancestors, n = 157 (%)¶	Ancestors with mutation, n (%)	Mutations in Non-Ancestors, n = 760 (%)¶	Non-ancestors with mutation, n (%)	**χ^2, df, P-value**
rrl	5(2.7)	3(27)	126 (12.2)	11 (22)	68.619, 31, 0.0001156*
rrs	4(2.1)	3 (27)	126(12.2)	10 (20)	
gyrB	5(2.7)	3 (27)	20(1.9)	14 (27)	
gyrA	45(24.1)	11(100)	184(17.8)	49 (96)	
rpoB	6(3.2)	2 (18)	23(2.2)	14 (27)	
rpoC	5(2.7)	4 (36)	25(2.4)	19 (37)	
Rv0678	0(0.0)	0 (0)	6(0.6)	6 (12)	
rpsL	12(6.4)	11 (100)	52(5.0)	51 (100)	
rplC	0(0.0)	0 (0)	4(0.4)	3 (6)	
embR	10(5.3)	5 (45)	53(5.1)	24 (47)	
fabG1	1(0.5)	1 (9)	2(0.2)	2 (4)	
inhA	2(1.1)	2 (18)	9(0.9)	8 (16)	
rpsA	1(0.5)	1 (9)	8(0.8)	7 (14)	
tlyA	12(6.4)	11 (100)	54(5.2)	51 (100)	
katG	3(1.6)	2 (18)	10(1.0)	8 (16)	
pncA	0(0.0)	0 (0)	3(0.3)	3 (6)	
kasA	2(1.1)	2 (18)	9(0.9)	8 (16)	
eis	0(0.0)	0 (0)	2(0.02)	2 (4)	
ahpC	0(0.0)	0 (0)	1(0.01)	1 (2)	
folC	1(0.5)	1 (9)	15(1.5)	12 (24)	
ribD	2(1.1)	2 (18)	1 (0.1)	1 (2)	
thyX	0(0.0)	0 (0)	2(0.2)	2 (4)	
thyA	3(1.6)	2 (18)	12(1.2)	11 (23)	
ald	3(1.6)	3 (27)	19(1.8)	18 (35)	
fbiA	1(0.5)	1 (9)	8(0.8)	7 (14)	
alr	33(17.6)	9 (81)	134(13.0)	39 (76)	
embC	14(7.5)	11 (100)	59(5.7)	50 (98)	
embA	7(3.7)	5 (45)	26(2.5)	20 (39)	
embB	0(0.0)	0 (0)	15(1.5)	11 (23)	

ethA	5(2.7)	2 (18)	4(0.4)	4 (8)
ethR	2(1.1)	2 (18)	6(0.6)	6 (12)
gid	3(1.6)	3 (27)	14(1.4)	13 (25)

Table 8 Logistic regression model results for ancestor status given a mutation in a candidate gene. Abbreviations: OR (95% CI): Odds ratio (95% Confidence Interval); LR: Log-likelihood ratio test. LR X2 value (P-value) is the difference with the null model

**Adjustment included controlling for sex, smoking, duration of cough, and genotypic drug resistance

Genes	Univariate model (Crude OR (95% CI))	**Multivariate model (aOR (95% CI))
rrl	**0.20 (0.07-0.44)**	**0.23 (0.08-0.54)**
rrs	**0.16 (0.04-0.38)**	**0.23 (0.07-0.55)**
gyrB	1.4 (0.46-3.48)	1.7 (0.50-4.90)
gyrA	**1.46 (1.0-2.1)**	1.48 (0.97-2.21)
rpoB	1.45 (0.53-3.4)	0.78 (0.24-2.18)
rpoC	1.11 (0.37-2.70)	1.27 (0.38-3.50)
Rv0678	2.5e-06	1.8e-06
rpsL	1.29 (0.65-2.39)	1.31 (0.62-2.60)
rplC	2.6e-6	7.3e-6
embR	1.04(0.49-2.00)	0.89 (0.37-1.93)
fabG1	2.77(0.13-29.05)	1.99 (0.09-22.65)
inhA	1.23 (0.14-4.82)	0.41 (0.05-2.29)
rpsA	0.69 (0.03-3.78)	0.71 (0.04-4.56)
tlyA	1.24 (0.62-2.29)	1.27 (0.60-2.50)
katG	1.67 (0.37-5.51)	2.48 (0.50-9.57)
pncA	2.60e-06	5.28e-06
kasA	1.23 (0.19-4.82)	1.61 (0.22-7.43)
eis	7.06e-06	1.58e-06
ahpC	7.08e-06	6.55e-06
folC	0.36 (0.02-1.81)	0.69 (0.04-3.87)
ribD	**11.15 (1.06-240.57)**	4.013 (0.29-99)
thyX	7.07e-06	2.40e-06

thyA	1.39 (0.31-4.41)	1.48 (0.32-5.14)
ald	0.87 (0.21-2.58)	0.96 (0.21-0.05)
fbiA	0.69 (0.04-3.78)	0.69 (0.04-3.93)
alr	1.44 (0.93-2.16)	1.25 (0.78-1.98)
embC	1.33 (0.70-2.38)	1.44 (0.72-2.72)
embA	1.50 (0.59-3.34)	1.02 (0.36-2.55)
embB	9.45e-07	3.33e-07
ethA	**7.06 (1.85-28.76)**	6.18 (0.90-43.21)
ethR	1.85 (0.27-8.09)	2.02 (0.27-10.30)
gid	1.19 (0.27-3.68)	1.46 (0.31-5.10)
LR X^2 value (P-value)		45.49(3.144e-09)

FIGURES

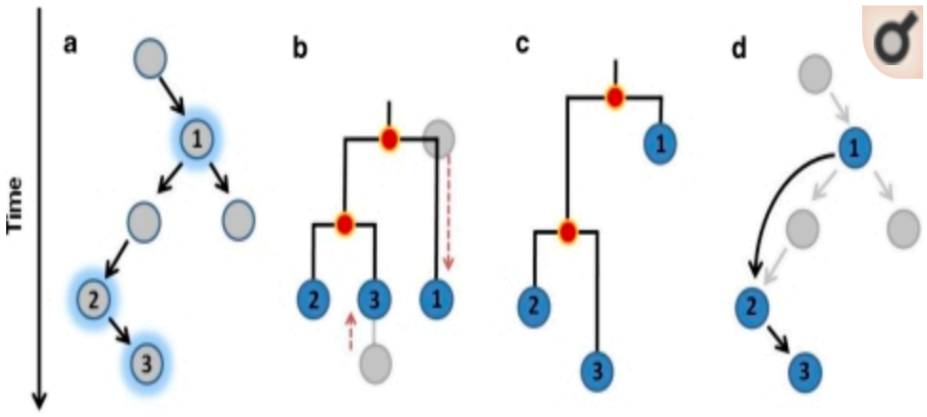


Figure 22: direct ancestry reconstruction method implemented in SeqTrack.

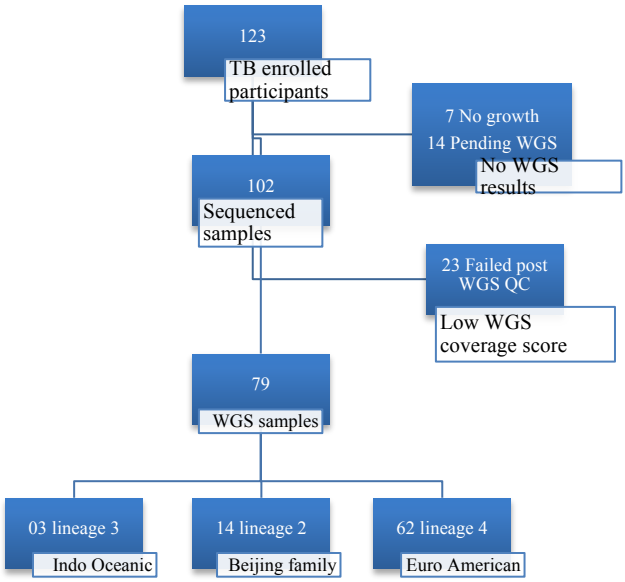


Figure 23: TB positive cases and lineage four samples included in this analysis

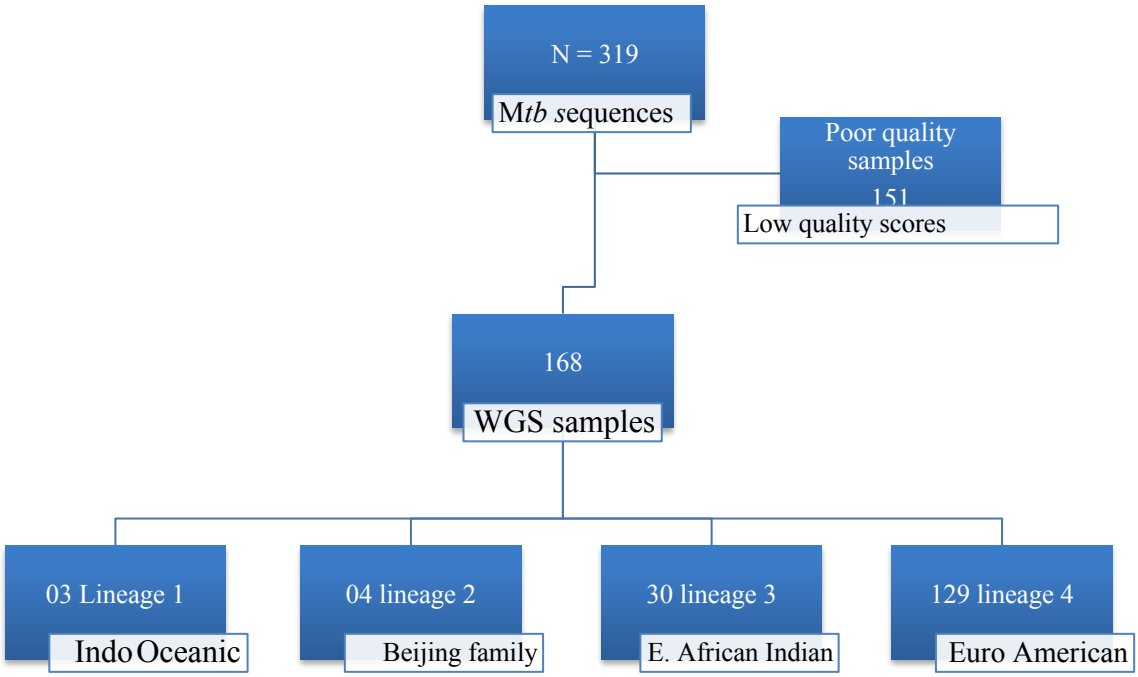


Figure 24: MTBC sequenced genomes included in Aim 2 study

Pictorial illustration of SNP distance annotation in Geneious software

MTB Seq 1: …….. GGTTAAGAGCTCTGTGAAAG…..
MTB Seq 2: …….. GGTTAAGAGCTTTGTGAAAG…..
MTB Seq 2: …….. GGTTAAGAGCTTTGTAAAAG…..

SNP difference: Seq1-Seq2 = 1
SNP difference: Seq1-Seq3 = 2
SNP difference: Seq2-Seq3 = 1

	1	2	3
1	0	1	2
2	1	0	1
3	2	1	0

Figure 25: Diagram illustrates the procedure for generation of SNP difference tables in Geneious software

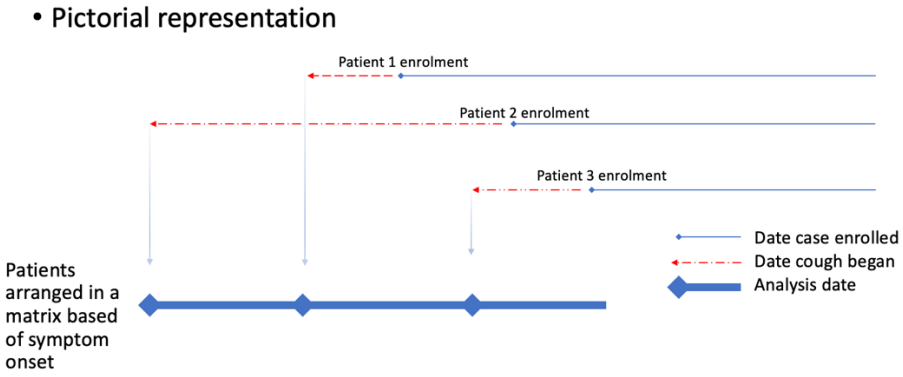


Figure 26: Diagram illustrating the method used to derive the start time for infectiousness

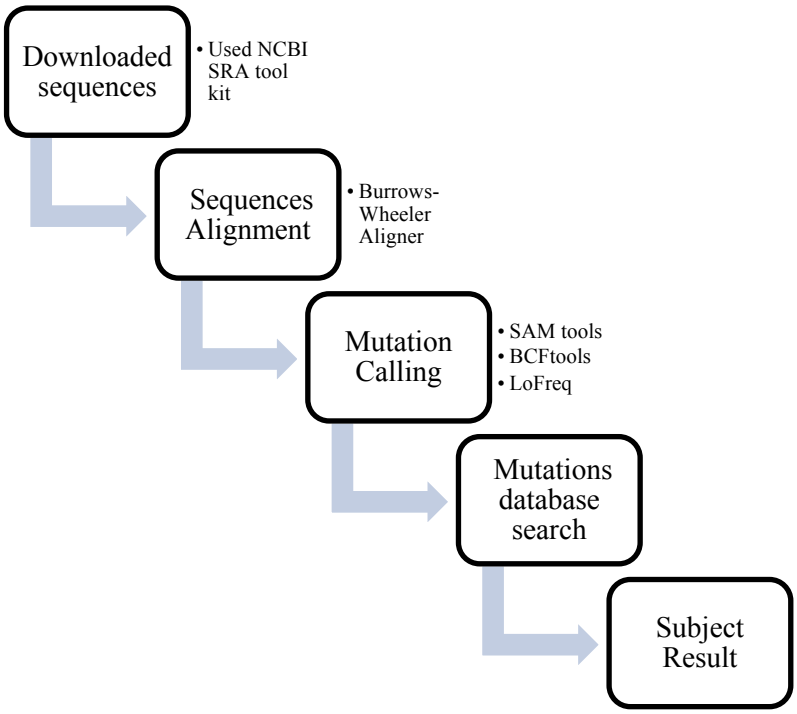


Figure 27: Summary of the TBprofiler pipeline steps

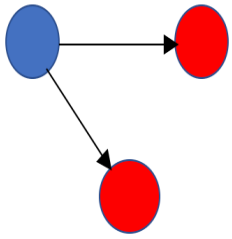
Cluster Type	Number of clusters (%)
	2 (18.2)
	9 (81.8)

Figure 28: Network with ancestors illustrated as blue and non-ancestors as red

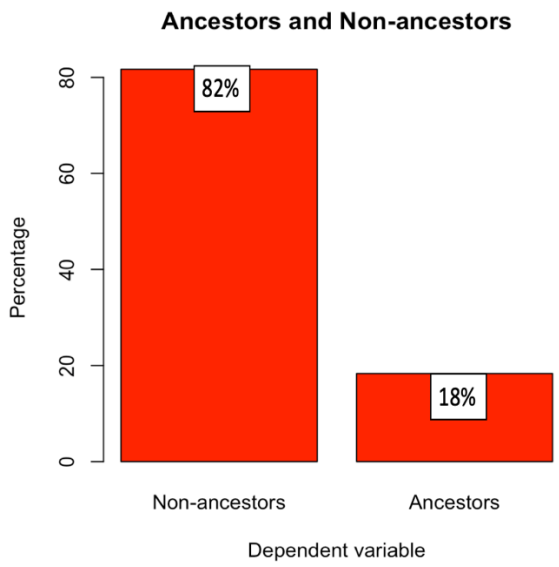


Figure 29: Distribution of identified ancestors in the population

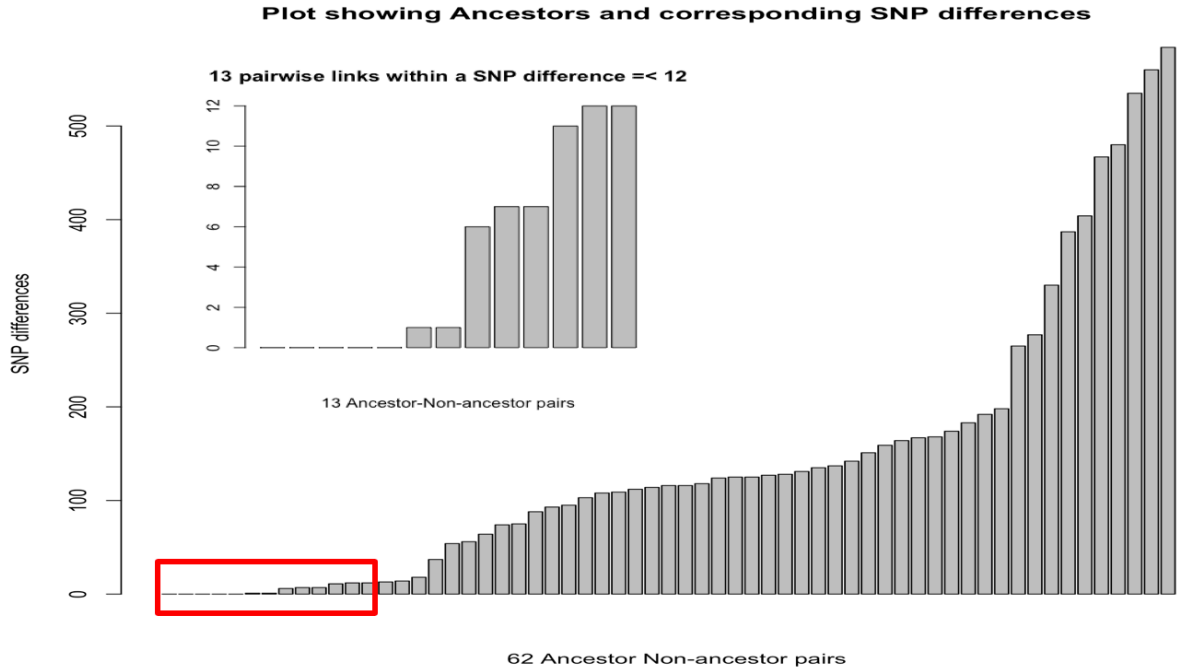


Figure 30: Distribution of SNP differences among 62 pairs of identified ancestors and non-ancestors

*The red box shows the pairwise SNP difference clusters whose ancestors were classified as true ancestors in this study.

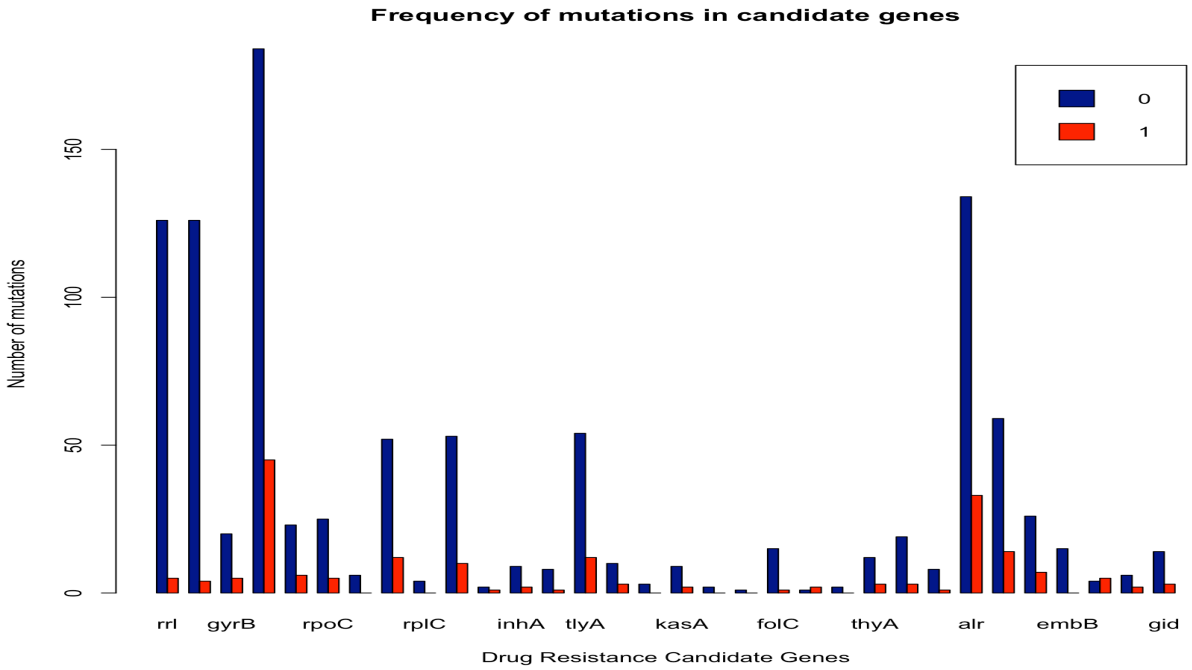


Figure 31: Candidate genes and the count of mutations in each gene among ancestors and non-ancestors

* Ancestors are represented by red bars (1), and non-ancestors are represented by blue bars (0).

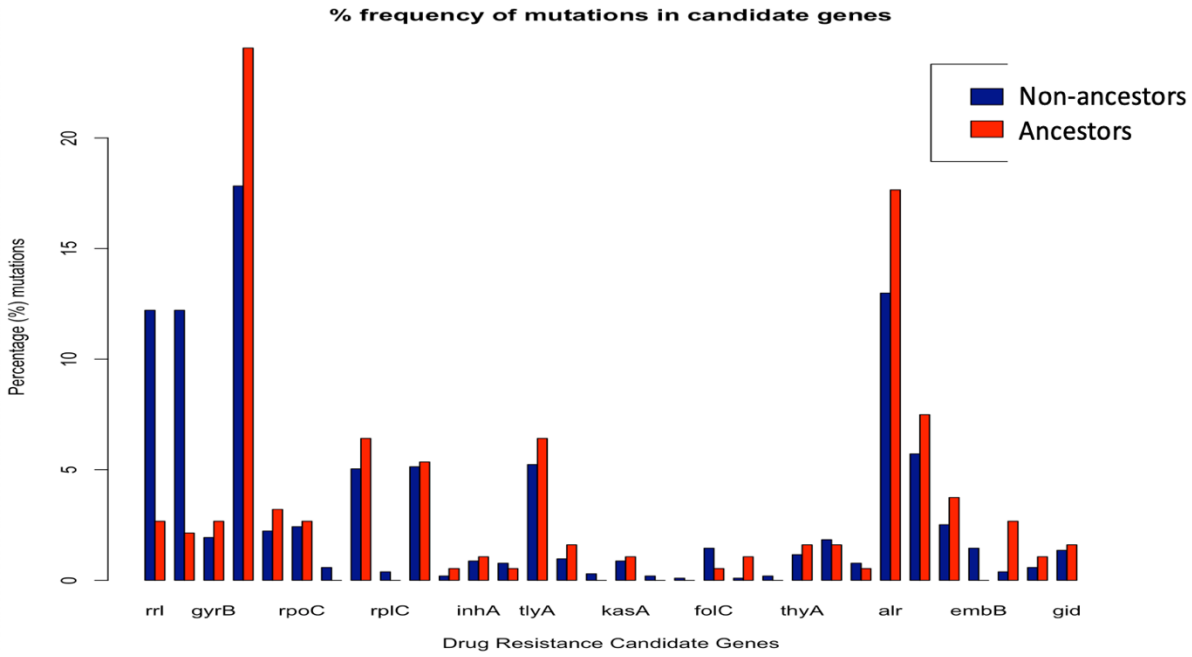


Figure 32: Candidate genes and mutations in each gene among ancestors and non-ancestors expressed as percentages

* Ancestors are represented by red bars and non-ancestors are represented by blue bars.

SUPPLIMENTALY TABLES AND FIGURES

Figure 33: Gene names, gene products and length of genes identified among study participant genomes

Supplementary Table S1 (Source: https://mycobrowser.epfl.ch/genes)

Gene	**Gene Product**	**Gene length**
rrl	Ribosomal RNA 23S	3138
rrs	Ribosomal RNA 16S	1537
gyrB	DNA gyrase (subunit B) GyrB (DNA topoisomerase (ATP-hydrolysing)) (DNA topoisomerase II) (type II DNA topoisomerase)	2028
gyrA	DNA gyrase (subunit A) GyrA (DNA topoisomerase (ATP-hydrolysing))	2517

	(DNA topoisomerase II) (type II DNA topoisomerase)	
rpoB	DNA-directed RNA polymerase (beta chain) RpoB (transcriptase beta chain) (RNA polymerase beta subunit)	3519
rpoC	DNA-directed RNA polymerase (beta' chain) RpoC (transcriptase beta' chain) (RNA polymerase beta' subunit).	3951
Rv0678	Conserved protein	498
rpsL	30 ribosomal protein S12	375
rplC	50 ribosomal protein L3	654
EmbR	Transcriptional regulatory protein EmbR	1167
fabG1	3-oxoacyl-[acyl-carrier protein] reductase FabG1 (3-ketoacyl-acyl carrier protein reductase) (mycolic acid biosynthesis a protein)	744
inhA	NADH-dependent enoyl-[acyl-carrier-protein] reductase InhA (NADH-dependent enoyl-ACP reductase)	810
rpsA	30S ribosomal protein S1 RpsA	1446
tlyA	2'-O-methyltransferase TlyA	807
katG	Catalase-peroxidase-peroxynitritase T KatG	2223
pncA	Pyrazinamidase/nicotinamidase PncA (PZase)	561
kasA	3-oxoacyl-[acyl-carrier protein] synthase 1 KasA (beta-ketoacyl-ACP synthase) (KAS I)	1251
eis	Enhanced intracellular survival protein Eis, GCN5-related N-acetyltransferase	1209
ahpC	Alkyl hydroperoxide reductase C protein AhpC (alkyl hydroperoxidase C)	588

folC	Probable folylpolyglutamate synthase protein FolC (folylpoly-gamma-glutamate synthetase) (FPGS)	1464
ribD	Possible bifunctional enzyme riboflavin biosynthesis protein RibD: diaminohydroxyphosphoribosylaminopyrimidine deaminase (riboflavin-specific deaminase) + 5-amino-6-(5-phosphoribosylamino)uracil reductase (HTP reductase)	777
thyX	Probable thymidylate synthase ThyX (ts) (TSase	756
thyA	Probable thymidylate synthase ThyA (ts) (TSASE	792
ald	Secreted L-alanine dehydrogenase Ald (40 kDa antigen) (TB43)	1116
fbiA	Probable F420 biosynthesis protein FbiA	996
alr	Alanine racemase Alr	1227
embC	Integral membrane indolylacetylinositol arabinosyltransferase EmbC (arabinosylindolylacetylinositol synthase)	3285
embA	Integral membrane indolylacetylinositol arabinosyltransferase EmbA (arabinosylindolylacetylinositol synthase)	3285
embB	Integral membrane indolylacetylinositol arabinosyltransferase EmbC (arabinosylindolylacetylinositol synthase)	3285
ethA	Monooxygenase EthA	1470
ethR	Transcriptional regulatory repressor protein (TetR-family) EthR	651
gid	Probable glucose-inhibited division protein B Gid	675

Supplementary Table S2: Ancestors and their descendant genomes and IDs used to generate phylogeny based transmission tree .

Netid	Network id	Ancestors	weight	date	Ances.date
20217	1	NA	NA	1/27/12	NA
20080	2	1	174	8/26/13	1/27/12
20094	3	1	192	10/16/13	1/27/12
20026	4	1	560	8/19/13	1/27/12
20269	5	7	0	2/26/14	2/24/14
20191	6	3	480	1/13/14	10/16/13
20242	7	2	131	2/24/14	8/26/13
20599	8	4	404	3/9/14	8/19/13
20375	9	22	167	6/18/14	3/18/14
20436	10	22	584	7/23/14	3/18/14
20814	11	16	1	11/25/14	10/30/14
20495	12	10	1	7/29/14	7/23/14
20626	13	14	109	9/9/14	8/22/14
20547	14	3	118	8/22/14	10/16/13
20819	15	14	114	12/1/14	8/22/14
20751	16	4	183	10/30/14	8/19/13
21381	17	14	142	1/9/15	8/22/14
20818	18	8	18	12/15/14	3/9/14
20851	19	9	159	12/15/14	6/18/14
20908	20	7	128	1/27/15	2/24/14

21535	21	18	13	2/1/15	12/15/14
21010	22	1	127	3/18/14	1/27/12
20951	23	14	7	2/17/15	8/22/14
21191	24	4	151	6/19/15	8/19/13
20941	25	23	103	2/23/15	2/17/15
20969	26	23	137	2/23/15	2/17/15
21061	27	22	164	3/26/15	3/18/14
21104	28	35	0	4/21/15	4/17/15
21075	29	12	198	4/24/15	7/29/14
21074	30	18	12	5/29/15	12/15/14
21172	31	23	467	4/6/15	2/17/15
21095	32	6	277	4/16/15	1/13/14
21076	33	50	124	4/21/15	3/12/15
21466	34	42	64	9/30/15	6/26/15
21083	35	17	95	4/17/15	1/9/15
21128	36	23	0	5/25/15	2/17/15
21164	37	50	116	6/15/15	3/12/15
21230	38	23	168	6/26/15	2/17/15
21179	39	23	11	6/28/15	2/17/15
21126	40	42	88	7/5/15	6/26/15
21609	41	48	0	11/20/15	10/31/15

21192	42	23	54	6/26/15	2/17/15
20440	43	14	7	10/20/15	8/22/14
21379	44	27	125	8/28/15	3/26/15
21510	45	18	125	10/15/15	12/15/14
21617	46	31	37	11/17/15	4/6/15
21619	47	42	75	11/24/15	6/26/15
21560	48	42	112	10/31/15	6/26/15
21618	49	42	93	11/30/15	6/26/15
21676	50	26	116	3/12/15	2/23/15
21847	51	11	74	6/15/16	11/25/14
21792	52	49	12	3/7/16	11/30/15
21931	53	6	265	6/8/16	1/13/14
22038	54	19	6	6/29/16	12/15/14
22039	55	42	535	6/29/16	6/26/15
22054	56	53	330	8/4/16	6/8/16
21988	57	53	387	6/16/16	6/8/16
21996	58	42	108	6/24/16	6/26/15
22193	59	47	0	8/9/16	11/24/15
22108	60	55	135	8/12/16	6/29/16
22264	61	42	56	9/19/16	6/26/15
22283	62	6	14	10/11/16	1/13/14

Figure 34: Distribution of mutations within genomes of ancestors

Gene	20242	20436	20547	20751	20818	20851	20951	21083	21560	21618	21619
rrl	0	0	0	0	1	3	0	0	0	0	1
rrs	0	0	0	0	1	0	2	1	0	0	0
Rv0005	0	0	0	1	3	1	0	0	0	0	0
Rv0006	4	5	4	3	4	4	4	5	4	4	4
Rv0667	0	1	0	0	0	5	0	0	0	0	0
Rv0668	0	1	0	1	2	1	0	0	0	0	0
Rv0678	0	0	0	0	0	0	0	0	0	0	0
Rv0682	2	1	1	1	1	1	1	1	1	1	1
Rv0701	0	0	0	0	0	0	0	0	0	0	0
Rv1267c	1	0	2	0	2	3	2	0	0	0	0
Rv1483	0	1	0	0	0	0	0	0	0	0	0
Rv1484	1	0	0	0	0	1	0	0	0	0	0
Rv1630	0	0	0	0	1	0	0	0	0	0	0
Rv1694	2	1	1	1	1	1	1	1	1	1	1
Rv1908c	0	0	0	0	2	0	0	0	0	0	1
Rv2043c	0	0	0	0	0	0	0	0	0	0	0
Rv2245	1	0	0	0	1	0	0	0	0	0	0
Rv2416c	0	0	0	0	0	0	0	0	0	0	0
Rv2428	0	0	0	0	0	0	0	0	0	0	0

Rv2447c	0	0	0	0	1	0	0	0	0	0	0
Rv2671	0	0	0	0	1	1	0	0	0	0	0
Rv2754c	0	0	0	0	0	0	0	0	0	0	0
Rv2764c	0	0	0	1	2	0	0	0	0	0	0
Rv2780	0	1	0	1	1	0	0	0	0	0	0
Rv3261	0	0	0	1	0	0	0	0	0	0	0
Rv3423c	4	3	4	0	0	5	4	4	4	4	1
Rv3793	1	2	1	1	2	1	2	1	1	1	1
Rv3794	1	0	0	0	2	2	0	1	0	0	1
Rv3795	0	0	0	0	0	0	0	0	0	0	0
Rv3854c	0	0	1	0	0	4	0	0	0	0	0
Rv3855	0	0	0	1	1	0	0	0	0	0	0
Rv3919c	0	0	0	1	1	0	0	1	0	0	0

Table 9: Data sources for this study (4 Bio-projects/ sequences from NCBI)

Study No.	**01**	**02**	**03**	**04**	**COHSONET**
Total Sequences	59	90	13	51	89
Good QC Seq	04	69	12	04	79
Lineage 4	04	52	07	03	62
Institution	Makerere	Institute of	Institute of	LSHTM	University of

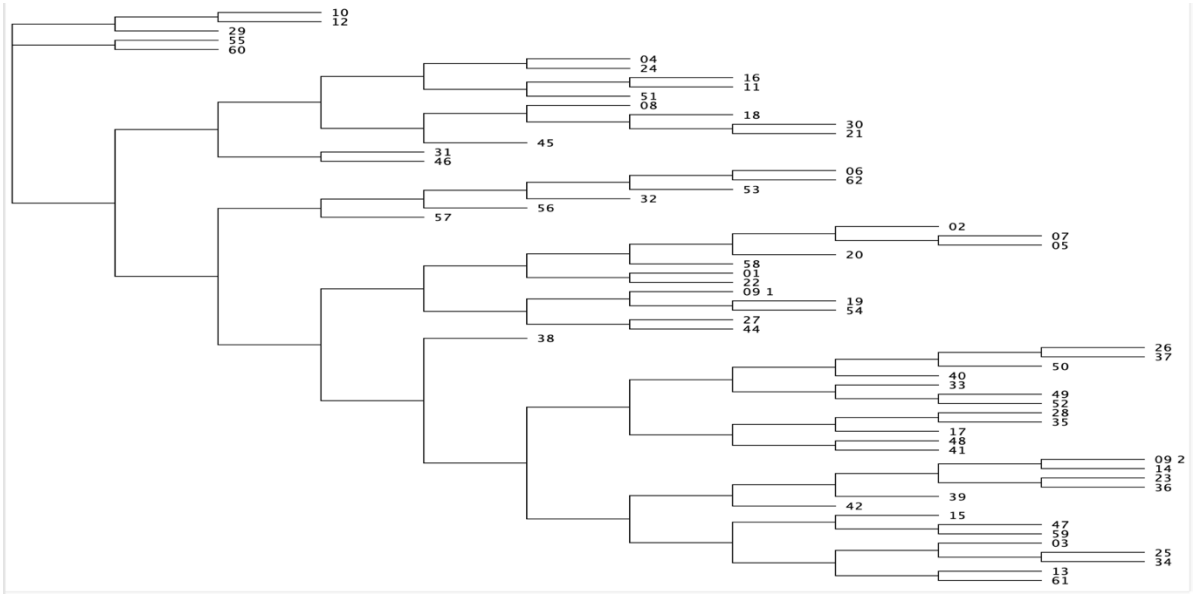
	University	tropical medicine, Belgium	tropical medicine, Belgium		Georgia, GHI/ CDC
Date of submission	18-July-2018	23-Feb-2016	6-Nov-2015	21-May-2013	Pending
Study design		Survey	Diagnostic	Longitudinal	Social Network
Accession No.	PRJNA481638	PRJEB10533	PRJEB10577	PRJEB224	N/A
Publication	Kateete, et al.	Ssengooba, et al. 2016	Ssengooba, et al. 2016	Clark TG, et al. 2013	Whalen, et al.

Table 10: Statistically significant candidate genes

Gene	Product	Association to ancestry
rrl	Ribosomal RNA 23S	Negative (OR= 0.20)
rrs	ribosomal RNA 16S	Negative (OR= 0.16)
gyrA	DNA gyrase (subunit B)	Positive (OR= 1.46)
ribD	Riboflavin biosynthesis protein RibD:	Positive (OR= 11.15)
ethR	Glucose-inhibited division protein B Gid	Positive (OR= 7.06)

Figure 35: Phylogenetic tree showing the relationships among genomes

*Inferred ancestors are generally closer in the phylogenetic tree; e.g ancestor for id 5 is 7 and they are closer on the phylogenetic tree).


10
12
29
55
60
04
24
16
11
51
08
18
30
21
45
31
46
06
62
53
32
56
57
02
07
05
20
58
01
22
09 1
19
54
27
44
38
26
37
50
40
33
49
52
28
35
17
48
41
09 2
14
23
36
39
42
15
47
59
03
25
34
13
61

Chapter 6

USING INTERFERON GAMMA CYTOKINE CONCENTRATION LEVELS TO IDENTIFY RECENT INFECTION WITH *MYCOBACTERIUM TUBERCULOSIS* IN A COMMUNITY SETTING

[1]Kirimunda S, Quinn F, Kakaire R, Chengwei L, Sekandi J, Whalen C. To be submitted to International Journal of Tuberculosis and Lung Disease

ABSTRACT

Background: This study addresses the problem of ongoing transmission of *Mycobacterium tuberculosis (Mtb)* in Sub-Saharan Africa. Inability to identify and treat *Mtb* infected individuals who are most likely to progress to disease constitutes an important impediment to TB control efforts. Tests for accurate diagnosis of recent infection, a marker of progression to disease are currently unavailable. We aimed to determine whether concentrations of interferon gamma (Interferon-γ) cytokines produced in response to *Mtb*-specific antigen stimulation could serve as a biomarker to distinguish individuals that are recently infected with *Mtb* from uninfected and remotely infected adults in the community.

Methods: From a longitudinal prospective cohort of adults who were TST negative at baseline and followed up for 1-2 years in Kampala Uganda, we randomly sampled 55 subjects who tested TST negative and 64 who tested TST positive at the last follow-up visit. From among those who tested TST positive at the initial assessment into the cohort, we randomly selected 97 remotely infected individuals. Measurement of *Mtb* specific antigen induced interferon-γ concentrations in blood was done using the QuantiFERON-TB Gold In-Tube assay (QFT) platform for all the three study groups.

Results: After adjusting for background interferon-γ levels, mean interferon-γ quantities were significantly higher in recently infected individuals compared to the uninfected and significantly lower than in remotely infected groups. Interferon-γ measurement resulted in the accurate prediction of 69% of recently infected individuals compared to remotely infected

individuals at an optimal cutoff value of 2.58 IU/ml with at a specificity of 73% and specificity of 67%.

Conclusion: The mean interferon-γ blood levels in converters were intermediate between uninfected and those with established infection. We demonstrated that the concentrations of interferon-γ after stimulation with *Mtb*-specific antigens could serve as a biomarker to discriminate recently infected from remotely infected groups despite having the same mean TST reading in the community.

BACKGROUND

There were over 10.0 million cases of tuberculosis (TB) in the world in 2018, a number that has been relatively stable in recent years. The burden of disease varies enormously among countries, from fewer than 5 to more than 500 new cases per 100,000 population per year (WHO, 2019). In 2018, tuberculosis incidence in Uganda was 200 cases per 100,000 individuals compared to 202 per 100,000 in 2015 (WHO, 2019). The Global health End TB strategy targeted to reduce TB incidence by 20% by 2020, a target that is bound not to be achieved (World Health Organization, 2015a)**.**

Approximately 2 billion people are latently infected with *M. tuberculosis* (Jajou et al., 2018), but only about 5 to 10% will develop TB disease in their life. Based on epidemic theory and what is known about tuberculosis transmission, TB epidemics continue to occur when one index case is replaced by one or more cases from a pool of latently infected individuals (Whalen, 2016). These latently infected persons are therefore the primary source of future TB cases, and their identification and treatment is important for intervention against the TB epidemic (Suliman et al., 2018; Whalen, 2016). However, it is not possible to diagnose and treat all 2 billion people who are infected, we need a way to identify who among those with LTBI are likely to progress to disease. If we had such are test, we would be able to limit the numbers who are given preventive therapy, thereby making it more feasible particularly in regions with high background LTBI prevalence like Sub-Saharan African where the majority of TB cases are likely due to recent infections from ongoing TB transmission (Chin et al., 1998; Kang et al., 2007; Verver et al., 2004).

Diagnosis of LTBI is primarily by two WHO-approved tests based on immunological activity suggestive of current or previous infection, commonly measured by either the tuberculin skin test (TST) or interferon gamma release assays (IGRAs) (Farhat et al., 2006). IGRAs measure IFN-γ secreted by the patient's T-lymphocytes or the number of IFN-γ-secreting lymphocytes upon *ex vivo* stimulation with 3 *Mtb* specific peptide antigens; early secreted antigenic target 6 (ESAT-6), culture filtrate protein 10 (CFP-10) and TB7.7 that are not found in BCG vaccine strains or most nontuberculous mycobacterium (NTM) species with exception of *M. kansasii, M. szulgai, and M. marinum* (Andersen et al., 2000a). Unlike TST, IGRA tests have a cutoff and an individual is considered positive for *M. tuberculosis* if the IFN-γ response to these TB antigens is above the test cutoff (after subtracting the background IFN-γ response of the negative control) regardless of the patient's exposure history. Due to substantial variability and poor reproducibility of IGRA results (van Zyl-Smit, Zwerling, Dheda, & Pai, 2009), WHO has in the past issued a recommendation against their use, advocating for research and development of newer serologic biomarkers for TB (Steingart et al., 2007; Whalen, 2005).

In addition to having a single cutoff value regardless of the patient's exposure history or immune status, the current approved LTBI tests are unable to differentiate between LTBI and active TB, nor distinguish recent from remote infection. The latter is an important distinction since recent infection is a strong risk factor for progression to active TB, and in some high incidence areas, the majority of TB cases are likely due to recent infections from ongoing TB transmission (Chin et al., 1998; Verver et al., 2004). In high-burden settings these assays show low sensitivity because of malnutrition, and immunosuppression as well as low specificity due to a high background prevalence of LTBI (Kang et al., 2007). This has been demonstrated by work previously done in the same study cohort as ours to test the performance of the QuantiFERON-

TB Gold In-Tube assay compared to TST. We found no remarked difference but rather a moderate performance similarity, underscoring the need to improve these tests for an endemic setting like Uganda (Castellanos et al., 2018). It has been proposed previously that to maximize the positive predictive value of these existing tests, LTBI screening should be reserved for those who are at sufficiently high risk of progressing to disease (Pai et al., 2014).

Developing a new diagnostic assay or deploying the current tests to identify recent *Mtb* infection would allow for targeted treatment of those persons most likely to progress to active TB and is a priority among international TB agencies (Pai & Schito, 2015). To address this challenge, new techniques, including transcript microarrays, flow cytometry of intracellular cytokines, and multiplex micro bead-based immunoassay (Luminex assay) of cytokines, have recently been introduced and used to study TB diagnostics (Berry et al., 2010; Caccamo et al., 2010; Chegou et al., 2009; Sutherland et al., 2010). Unfortunately, the majority of these studies were conducted in cohorts of household or close contacts of TB cases. Whereas the household is an environment of intense transmission of *Mtb*, it does not account for majority of new transmission of tuberculosis yet it remains the focus of most TB transmission studies in the current literature (Crampin et al., 2006; Glynn et al., 2015; Martinez et al., 2017a; Verver et al., 2004; Whalen et al., 2011).

Here, we report results from a non-household community-based cohort of Ugandans who were *Mtb* uninfected at baseline and followed for 1-2 years until conversion (*Mtb* infection). We compare concentrations of *Mtb* specific antigen induced Interferon gamma (IFN-γ) to differentiate recently infected (TST converted) individuals from remotely infected and uninfected individuals. We also assessed for the optimal QFT cutoff to differentiate TST

negative and TST positive individuals in a community setting beyond the household of a TB case.

METHODS

Study setting, study population, and data collection: From a longitudinal prospective cohort of adults who were TST negative at baseline and followed up for 1-2 years, we sampled 55 subjects who tested TST negative and 64 who tested TST positive at the last visit of follow up. Among those who tested TST positive at the initial assessment into the cohort, 97 were assumed to be remotely infected and were enrolled as a third group in this study. All participants were residents of Rubaga division of Kampala, Uganda.

Initial TST was performed by the Mantoux method with 5 tuberculin units (TU) of purified protein derivative (PPD). After the initial evaluation, participants were evaluated after one year for active TB and LTBI. Whole blood collection for QuantiFERON-TB Gold In-Tube assay (QFT) was done at the last follow up visit (Figure 35 & 36). TST conversion was defined as a participant with an initial TST < 5 mm who had a TST >= 10 mm or with an increment of 6 mm by the end of follow-up. (Castellanos et al., 2018). If the above conditions were not met, subjects were classified as persistent TST negative. All subjects with a positive TST at the end of follow up were considered positive and offered treatment with isoniazid preventive therapy (IPT) (10 - 20 mg/kg or a maximum dose of 300 mg/day) for nine months. However, no LTBI subjects were receiving LTBI treatment at the time of sampling. The study was approved by the Makerere University School of Public Health Institutional Review Boards in Uganda and all participants provided written informed consent.

Sample Processing: Whole blood samples for QFT testing were collected in 1 mL tubes provided with the testing kits, transported to and received by the laboratory at room temperature

within approximately 2 hours of blood collection. The tubes were incubated at 37ºC for 16-24 hours when the plasma was separated and stored at -80° until Enzyme Linked Immuno-Sorbent Assays (ELISA) were performed.

Measurement of IFN-g by ELISA: ELISA assays to test Interferon gamma were carried out as per manufacturers instruction and optical density (OD) values were used to compute IFN-γ concentrations in international units per milliliter (IU/mL) using QuantiFERON-TB Gold IT Analysis Software (version 2.17) (Qiagen, n.d.). All samples were evaluated by laboratory scientists who were blinded to the study participants and the clinical data. The background corrected IFN-γ concentrations were defined by subtracting the concentration in unstimulated supernatant from the corresponding concentration in *Mtb* antigen stimulated supernatant.

Statistical Analysis: The sample size for each group was determined using Open Epi online tool to sufficiently detect group differences at a statistical significant level of 0.05 with power of 80% (Sullivan, Dean, & Minn, 2009). Demographic characteristics of recently infected, uninfected and remotely infected participants were compared using frequency tables, chi-squared tests, Kruskal-Wallis chi-squared test, Fisher' s exact test and Analysis of variance (ANOVA) methods in R-statistical package. Quantitative concentrations of IFN-γ cytokine were tabulated for recently infected and remotely infected participants. We calculated odds ratios and 95% confidence intervals (OR, 95% CI) for the association of recent infection and unit change in concentration of IFN-γ using logistic regression ("Applied Logistic Regression - David W. Hosmer, Jr., Stanley Lemeshow, Rodney X. Sturdivant - Google Books," n.d.). We adjusted the associations for age, HIV status, religion and education levels accordingly.

Data were analyzed to generate test Receiver Operator Curves (ROC) to define the diagnostic performance of IFN-γ cytokine concentration to discriminate between the three

groups. The same methods were used to asses for the optimal cutoff to differentiate TST negative (TST reading < 10 mm) and TST positive (TST reading > 10mm). The area under the ROC curve (AUC) was determined in R using the pROC package (Robin et al., 2011), together with the optimal cutoff value that maximized Youden's index (sum of % sensitivity and % specificity - 100). Youden's index, often used in conjunction with ROC analysis was used as a criterion for selecting the optimum cutoff point (Schisterman et al., 2005). A cutoff values in IU/ml to discriminate between recently infected and uninfected or remotely infected groups were proposed.

RESULTS

Study population: A total of 216 participants (64 recently infected, 55 uninfected, and 97 remotely infected) were analyzed in this study. Recently infected subjects were slightly younger than the uninfected and remotely infected subjects with mean age of 26.8 years versus 27.9 and 30.6 years respectively (Table 11). There were more males than females among the recently infected, uninfected and remotely infected groups in this study (58.2% versus 57.8% and 56.7% males respectively). None of the remotely infected subjected were HIV positive, however 4.7% and 14.5 % of those recently infected and uninfected tested HIV positive respectively (p-value = <0.05). There was minimal difference in the final mean TST reading among recently infected subjects compared to those remotely infected (15.0 mm versus 16.2 mm, P= 0.019) relative to the uninfected group with a mean TST reading of 0.57 mm (Figure: 40 & 41). IGRA diagnosis (QFT) identified 56.2% of the recently infected as LTBI positive compared to 82.5% among those remotely infected and 14.5% in the uninfected group (46.2% versus 82.5% versus 14.5%, p-value = 2.2e-16) (Table 11).

Interferon-γ concentration in recently infected group compared to other groups: There was no statistically significant difference in the mean concentration of IFN-γ in the unstimulated supernatants among all three groups (0.42 IU/ml versus 0.47 IU/ml versus 0.45IU/ml, p-value= 0.26, Figure 43). The concentration of IFN-γ was high in unstimulated supernatant of 2 remotely infected individuals resulting in indeterminate results while other groups did not have any indeterminate results. The mean absolute concentration of IFN-γ was highest among those remotely infected, followed by recently infected and lowest among the uninfected group (5.12 IU/ml versus 2.74 IU/ml versus 0.47 IU/ml) respectively (Table 12 & Figure: 42, 43, 44, & 45). The odds ratio of being recently infected versus remotely infected (OR= 1.33, 95% CI: 0.14-1.64) or versus the uninfected group (OR = 0.87, 95% CI: 0.78-0.94) given a unit increase in IFN-γ concentration was consistent with our hypothesis.

Receiver Operator Curve/Area Under the Curve (ROC/AUC) analysis Results: An AUC of 0.69 (0.61-0.78) and an optimal concentration cutoff of 2.58 IU/ml was ideal for discrimination of recently infected participants and remotely infected ones (Figure: 47). This optimal cut off had a 67.0% sensitivity and 73.0% specificity and a Youden's index of 0.4. An AUC of 0.86 (0.8 - 0.92) and an optimal concentration cutoff of 0.84 IU/ml was ideal for discrimination of uninfected participants and remotely infected ones (Figure: 48) with a sensitivity of 78% and specificity of 89% with a Youden's index of 0.67 (Table 13).

DISCUSSION

In this study, the interferon-γ blood levels in converters were intermediate between uninfected and those with established infection. We found that interferon-γ concentration levels were different between persons with a recent TST conversion from those remotely infected despite having similar TST readings. In this work, we successfully use interferon-γ levels as a

marker for recent infection in a community setting. When compared to the currently recommended diagnostic cutoff values, the optimal cutoff value for our study population was lower but the performance characteristics did not differ. To the best of our knowledge, these are new findings and have not been demonstrated before in our study setting.

By conducting our study in a community cohort and not the household of TB cases or their contacts, this study is representative of operational research and edges the field of tuberculosis transmission control closer to a programmatic setting, informative of a novel approach to deploy interferon-γ testing in TB control (Cobelens, van Kampen, Ochodo, Atun, & Lienhardt, 2012). Whereas the household is an environment of intense transmission of *Mtb*, it does not account for majority of new transmission of tuberculosis yet it remains the focus of most TB transmission studies (Martinez et al., 2017a; Verver et al., 2004; Whalen et al., 2011). Our results demonstrate a potential approach for the investigation of new infections away from the household in a programmatic settings.

Although the participants in our study were randomly selected for blood collection and inclusion in this study, the mean TST readings of recently infected individuals and remotely infected ones was similar (TST = 15.0mm versus 16.2mm, standard deviation of 3.3 and 4.0 respectively). This was in support of our hypothesis that TST readings did not differ between recently infected and remotely infected groups. Nevertheless, misclassification cannot be ruled out due to already known limitations of the TST such as false-positives due to non-tuberculous mycobacterium (NTM) infection and prior BCG vaccination. Our study was limited to adults older than 15 years and the effect of BCG should be minimal since BCG is offered during infancy in Uganda. False-negatives in TST results may have affected the classification of patient groups in this study because of limited sensitivity in HIV patients (Farhat et al., 2006) who were

significantly more in the uninfected group than those remotely infected. Unlike the TST, IGRA tests are currently approved to have only one cutoff value regardless of the patient's exposure history or immune status (Steingart et al., 2007; Whalen, 2005). In this work, we have explored and demonstrated the use of an alternative cutoff value to delineate recent infection from remote infection in adults.

There was no difference in the background mean interferon-γ levels before stimulation with *Mtb* specific antigen among the three groups (0.42 IU/ml versus 0.47 IU/ml versus 0.45 IU/ml) however, 2 remotely infected subjects had a high concentration of interferon-γ resulting in indeterminate LTBI results. The effect of indeterminate results where the unstimulated interferon-γ level was characteristically high was probably due to ongoing infection at the time of sampling which requires further investigation in the context of recent infection. It is notable however that these high background concentrations were independent of HIV since they occurred in the remotely infected participants who were HIV negative.

We observed a unit increase in the concentration of interferon-γ from the uninfected to the recently and remotely infected groups respectively. This result supports the hypothesis that IGRA responses are related to the bacillary burden and antigenic load present in the body (Wallis et al., 2010). This result which suggests that concentration of interferon-γ, a marker of inflammation is associated with *Mtb* infection is not new and it is a finding comparable to results that showed that early infection was associated with specific host response processes related to inflammation and immune-response which were predictive of TST conversion (Bark et al., 2017). Diagnosis of recent infection has also been previously demonstrated by studying cytokines that are associated with interferon-γ such as interferon-γ induced protein 10 (IP-10) (Biraro et al., 2016a).

Our results showed that interferon-γ correctly classified 69% of recently infected participants from remotely infected individuals at an optimal cutoff of 2.58 IU/ml. This optimal cutoff is more than the current QFT test kit manufacturer's suggested 0.35 IU/ml cutoff for the diagnosis of LTBI (Qiagen, n.d.). When the manufacturer's cutoff was deployed to differentiate recently infected individuals from those remotely infected in the study population, it showed a better sensitivity of 85% (95% CI, 76-91) and a dismal specificity of 44% (31-56) with a low Youden's index of 0.27 compared to 0.4 by our proposed cutoff (Table 14). Results comparing remotely infected and uninfected showed an optimal cutoff of 0.84 IU/ml, the sensitivity of 78% and a specificity of 0.89% and an AUC of 0.86 (0.81-0.92). This was comparable to results by Won and colleagues who compared *Mtb* infected and uninfected controls of 0.5 IU/ml, sensitivity of 87.3% and a specificity of 100% and an AUC of 0.954 (0.91-0.98) although their work was from a relatively small cohort with a cross-sectional design (Won et al., 2017). Our work therefore partly validates similar earlier results but in a longitudinal setting (Chegou et al., 2009).

The optimal interferon-γ concentration cutoff for differentiating LTBI (defined as having a TST reading greater or equal to 10 mm) was 0.21 IU/ml compared to the current 0.35 IU/ml as defined by QFT manufacturers. There was no difference in performance characteristics when the proposed cutoff value was applied compared to when the current recommended cutoff value of 0.35 IU/ml was used (Youden's index of 0.57 versus 0.54 respectively). Since the performance characteristics were not different when either cutoff was used, there remains a need for more research to understand why the QFT test has poor performance characteristics in TB endemic populations of sub-Saharan Africa.

This study had several important limitations. The study population was restricted to adults and we did not include children so ascertainment bias could have occurred (Delgado-Rodríguez & Llorca, 2004) limiting the generalizability of our results. Second, whereas we assumed incident *Mtb* infection among study subjects in our study based on TST results, there is no laboratory gold standard to that effect. A positive TST response could be due to exposure to similar antigens from other non-tuberculous mycobacteria or BCG vaccination or a previous infection that has been cleared (Pai et al., 2014). Third, this study relied on a remotely infected group that is assumed to have been retrospectively infected for more than 1-2 years before blood collection. By combining subjects from different time periods, we probably introduced potential for confounding by temporal trends. Lastly, when assessing the results of AUC, it is recommended that all metrics of a model be measured on a dataset separate from the one used to build the model. Independence of test data is crucial because reusing the data on which a model was built (the "training data") to measure accuracy may overestimate the parameters of the model in future clinical applications (Efron, 1986). Our sample size, though larger than many previous studies, was not sufficient for 'training-test' data partition to be successfully employed. Additionally, using the AUC alone as a metric assumes that a false-positive result is just as bad as a false-negative result (Meurer & Tolles, 2017) and this is not intuitively correct for *Mtb* infection.

In conclusion, we show that mean blood level concentration of interferon-γ in converters were intermediate between uninfected and those with established infection. We demonstrated that the concentrations of interferon-γ after stimulation with *Mtb*-specific antigens could serve as a biomarker to discriminate recently infected from remotely infected individuals with the same TST readings in the community.

TABLES AND FIGURES

TABLES

Table 11: Demographic and clinical characteristics of the study population

P-values are by the fisher exact method, Chi-squared , t-test and ANOVA tests. *p-values are for comparison of remotely infected and recently infected groups. HIV, human immunodeficiency virus; TST, tuberculin skin test.

Variables	TST Converters, n = 64 (%)	TST negative, n = 55 (%)	TST positive, n = 97 (%)	P-values
Characteristics				
Mean age, (standard deviation)	26.8 (7.2) Range:18-47	27.9 (7.2) Range:18-49	30.6 (9.9) Range:15-61	0.0001067
TST reading, (standard deviation)	15.0 (3.3) Range:10-25	0.22 (1.16) Range:0-7.1	16.2 (4.0) Range:10.2-29.1	0.019*
Sex				0.982
Male	37 (57.8)	32 (58.2)	55 (56.7)	
Female	27 (42.2)	23 (41.8)	42 (43.3)	
HIV				2.591e-07
Positive	3 (4.7)	8 (14.5)	0 (0.0)	
Negative	54 (84.4)	47 (85.5)	97 (100)	
Unknown	7 (10.9)	0 (0.0)	0 (0.0)	
Marital Status				
Single/Never married	29 (45.3)	25 (45.5)	37 (38.1)	0.78
Married	26(40.6)	18 (32.7)	42 (43.3)	
Separated	1 (1.6)	4 (7.3)	3 (3.1)	
Divorced	6 (9.4)	5 (9.1)	12 (12.4)	
Widowed	2 (3.1)	3 (5.5)	3 (3.1)	
Religion				0.01
Roman Catholic	15 (23.4)	29 (52.6)	28 (28.9)	
Anglican	22 (34.4)	13 (24.6)	23 (23.7)	

Muslim	14 (21.9)	11 (19.3)	30 (30.9)	
Adventist	3 (4.7)	0 (0.0)	2 (2.1)	
Others (No response)	9 (14.1)	2(3.5)	14 (14.4)	
Occupation				
Agriculture	0 (0)	1 (1.8)	3 (3.1)	0.09
Bar worker	0 (0)	0(0)	1 (1.0)	
Boda-boda (Motor cyclists)	3 (4.8)	5 (8.8)	9 (9.3)	
Casual laborer	5 (8.1)	5 (8.8)	5 (5.2)	
Construction	1 (1.6)	2 (3.5)	3 (3.1)	
Fishing	0 (0)	0	0 (0)	
Government employed	1 (1.6)	0(0)	1 (1.0)	
Hair dresser	3 (4.8)	0 (0)	1 (1.0)	
Housework	3 (4.8)	2 (3.5)	13 (13.4)	
Housekeeper	0 (0)	1 (1.8)	2 (2.1)	
Home brewing	1 (1.6)	1 (1.8)	0 (0)	
Mechanic	11 (17.2)	9 (16.4)	4 (4.1)	
Medical worker	0 (0)	0 (0)	0 (0)	
Military/police	2 (3.2)	0 (0)	3 (3.1)	
Shopkeeper	2 (3.2)	2 (3.5)	1 (1.0)	
Student	8 (12.9)	5 (8.8)	9 (9.3)	
Trading/vending	11 (17.7)	14 (24.6)	32 (33.0)	
Truck driver	2 (3.2)	0 (0)	3 (3.1)	
Unemployed	1 (1.6)	2 (3.5)	3 (3.1)	
Waitress	5 (8.1)	2 (3.5)	2 (2.1)	
Others	5 (7.8)	4 (7.2)	2 (2.1)	
Highest Education				0.002
No formal Education	1 (1.6)	1 (3.6)	1 (1.0)	
Primary Education	15 (23.4)	18 (32.7)	50 (51.5)	
Post primary education	48 (75.0)	35 (63.6)	46 (47.4)	
Tribe				0.87
Ganda	45 (71.0)	39 (70.2)	64 (66.0)	
Ankole	5 (8.0)	5 (8.8)	8 (8.0)	
Others	14 (21.0)	11 (21.0)	27 (28.0)	
Income (UGX)				0.18
Less than 100,000	18 (28.1)	19 (34.5)	45 (46.0.)	
100,000 - 200,000	19 (39.7)	16 (29.1)	22 (23.0)	
More than 200,000	27 (42.2)	20 (36.4)	29 (30.0)	

Table 12: Showing the LTBI IGRA QFT results and interferon-γ concentration

P-values are by fisher exact and ANOVA methods. Abbreviations: LTBI, latent TB infection, QFT, QuantiFERON_-TB Gold In-Tube; IU/ml, international units per milliliter. *p-values are for comparison of remotely infected and recently infected groups

Variables	TST Converters, n = 64 (%)	TST negative, n = 55 (%)	TST positive, n = 97 (%)	P-values
QFT LTBI Diagnosis				2.2e-16
Negative	28 (43.8)	47 (85.5)	15 (15.5)	
Positive	36 (56.2)	8 (14.5)	80 (82.5)	
Indeterminate	0 (0.0)	0 (0.0)	2 (2.0)	
Interferon Gamma (IU/ml)				
Unstimulated QFT concentration, mean (SD)	0.42 (0.9) Range: 0.03-6.5 Median: 0.145	0.47 (1.0) Range: 0.03-6.0 Median: 0.16	0.45 (1.1) Range: 0.03-10 Median:0.14	0.26*
Quantitative interferon-γ (IFN-γ), mean (SD)	2.50 (3.6) Range:-0.71-10 Median: 0.74	0.46 (2.1) Range:-4.64-10 Median: 0.01	5.12 (4.0) Range:-9.8-10 Median: 5.52	2.94e-06 *

Table 13: Showing Logistic regression and Receiver Operator Curve analysis results

Abbreviations: OR (95% CI): Odds ratio (95% Confidence Interval); AUC, area under the curve; CI, confidence interval; FN, false negative; FP, false positive; TN, true negative; TP, true positive.

Characteristic	Positives vs Negatives	Converters vs Negatives	Converters vs Positives	TST(-) vs TST(+)
Odds Ratio (95% CI)	1.5 (1.31-1.80)	1.30 (0.16-1.50)	0.86 (0.8-0.94)	1.4 (1.2-1.7)
AUC (95% CI)	0.86 (0.80-0.92)	0.71 (0.61-0.80)	0.69 (0.61-0.78)	0.80 (0.72-0.85)
Optimal cutoff (IU/ml)	0.84	0.61	2.58	0.21
Youden's index	0.67	0.37	0.4	0.57
Sensitivity (95% CI)	0.78	0.56 (0.43-0.69)	0.67 (0.57-0.76)	0.75 (0.68-0.82)
Specificity (95% CI)	0.89	0.85 (0.73-0.93)	0.73 (0.61-0.84)	0.82 (0.69-0.91)
True Positives	76	36	65	121
False Negatives	21	28	32	40

False Positives	6	8	17	10
True Negatives	49	47	47	45

Table 14: Receiver Operator Curve analysis results comparing the proposed cutoff values for differentiating TST(-) and TST (+) and recent from remote infection.

Abbreviations: OR (95% CI): Odds ratio (95% Confidence Interval); AUC, area under the curve; CI, confidence interval; FN, false negative; FP, false positive; TN, true negative; TP, true positive, *Cutoff values of 0.32 and 0.33 IU/ml were closest to the currently recommended cutoff of 0.35 IU/ml.

Characteristic	Converters vs Positives*	Converters vs Positives	TST(-) vs TST(+)*	TST(-) vs TST(+)
ROC Area Under the curve (AUC)	N/A	0.69 (0.61-0.78)	N/A	0.80 (0.72-0.85)
cutoff (IU/ml,	0.32*	2.58	0.33*	0.21
Youden's index	0.27	0.4	0.54	0.57
Sensitivity (95% CI)	0.85 (0.76-0.91)	0.67 (0.57-0.76)	0.73 (0.65-0.79)	0.75 (0.68-0.82)
Specificity (95% CI)	0.44 (0.31-0.56)	0.73 (0.61-0.84)	0.82 (0.69-0.91)	0.82 (0.69-0.91)
True Positives	82	65	117	121
False Negatives	15	32	44	40
False Positives	36	17	10	10
True Negatives	28	47	45	45

FIGURES

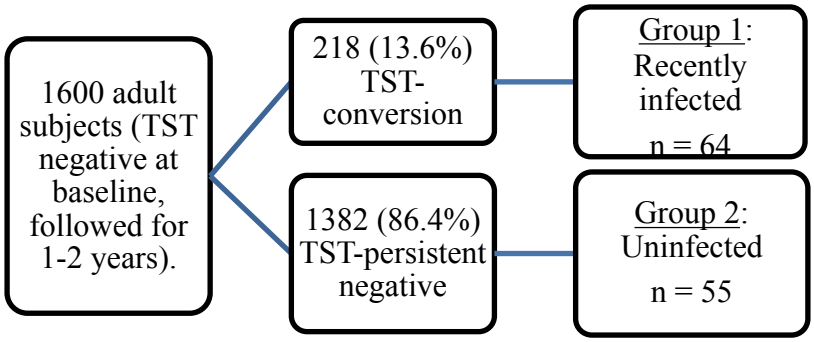


Figure 36: Longitudinal parent study flowchart showing uninfected and recently infected groups included in this study

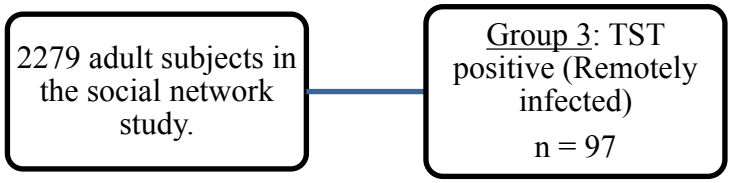


Figure 37: Cross-sectional social network parent study parent study flowchart showing remotely infected group included in this study

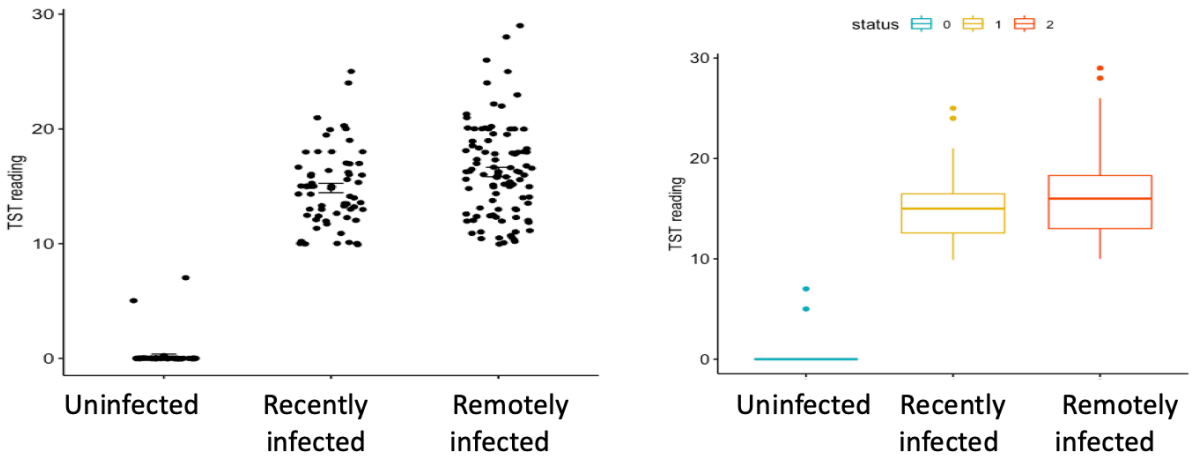


Figure 38: Plots showing TST reading in millimeters for uninfected participants, recently infected and remotely infected groups.

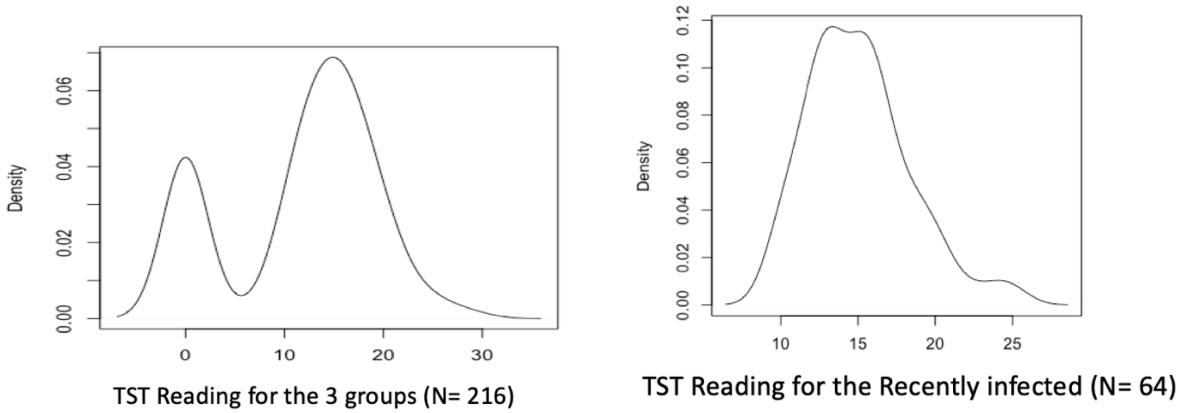


Figure 39: Graph showing final TST reading density plots for the all population (left) and recently infected group (right).

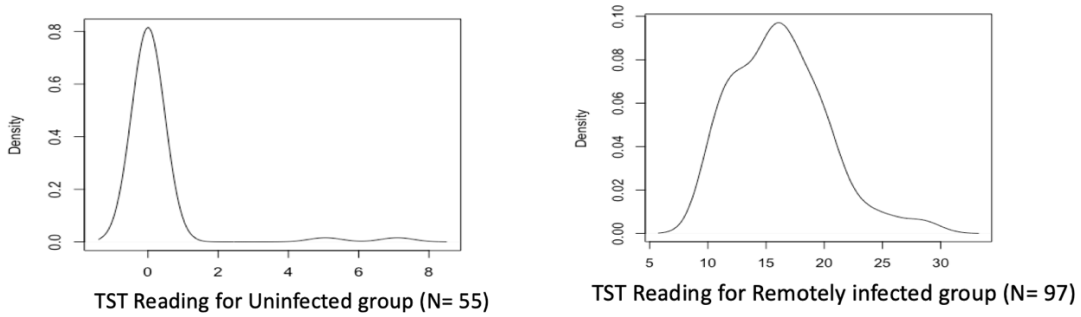


Figure 40: Graph showing final TST reading density plots for the uninfected group and remotely infected group(right)

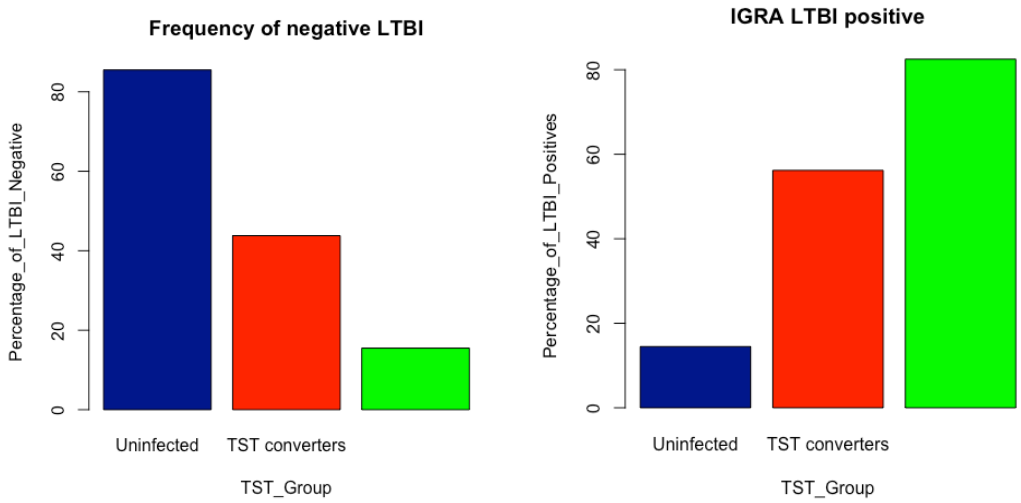


Figure 41: Graph showing proportion of LTBI negative (left) and positives (right) results by QuantiFERON_-TB Gold In-Tube assay in the three study

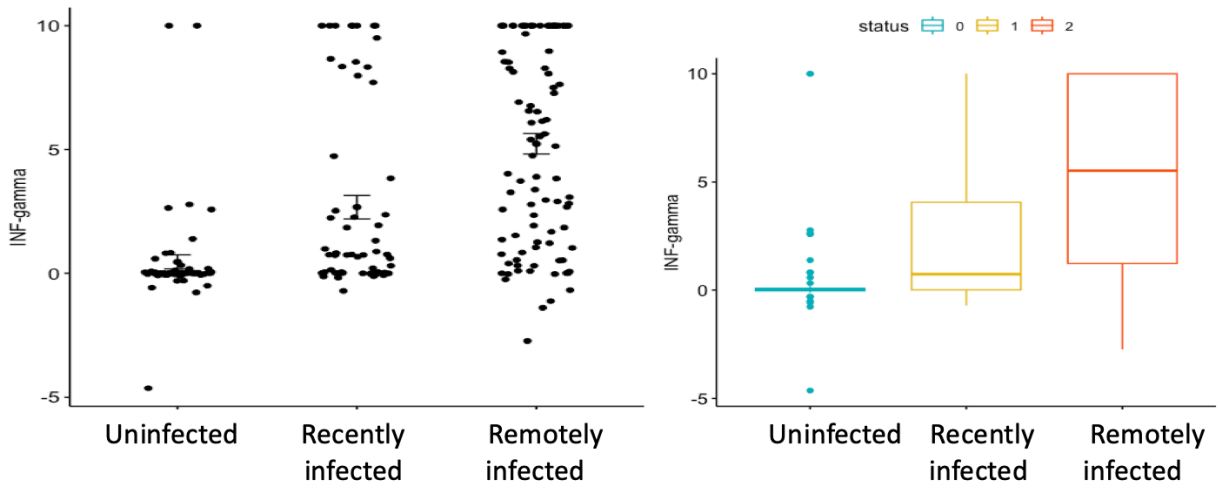


Figure 42: Plots showing concentration of interferon-γ in IU/ml for uninfected participants, recently infected and remotely infected groups.

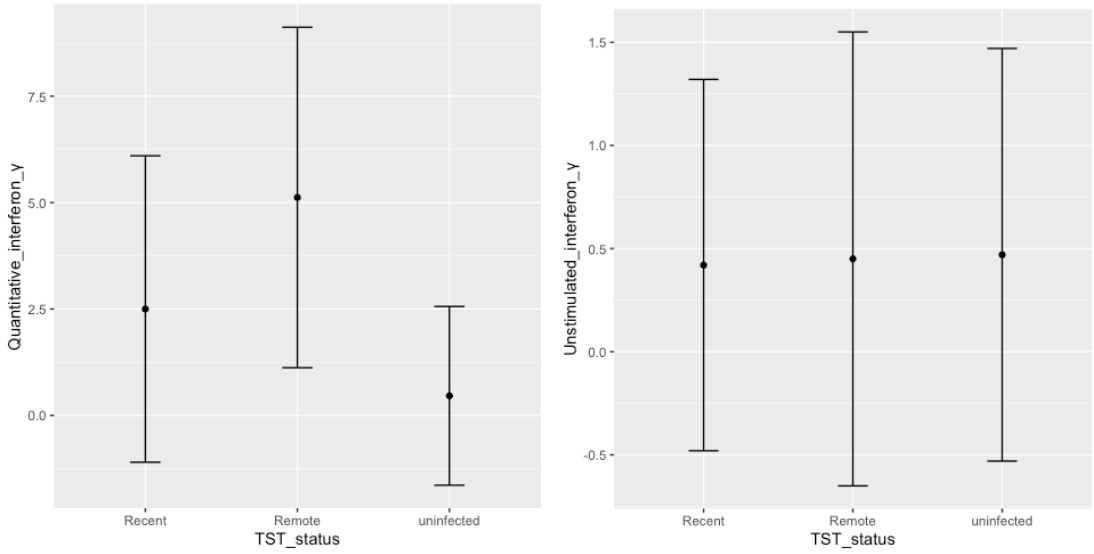


Figure 43: Scatter plots showing biomarker concentrations in Mtb-specific antigen stimulated supernatants from recently infected, remotely infected and uninfected groups (left) and from unstimulated supernatants (right)

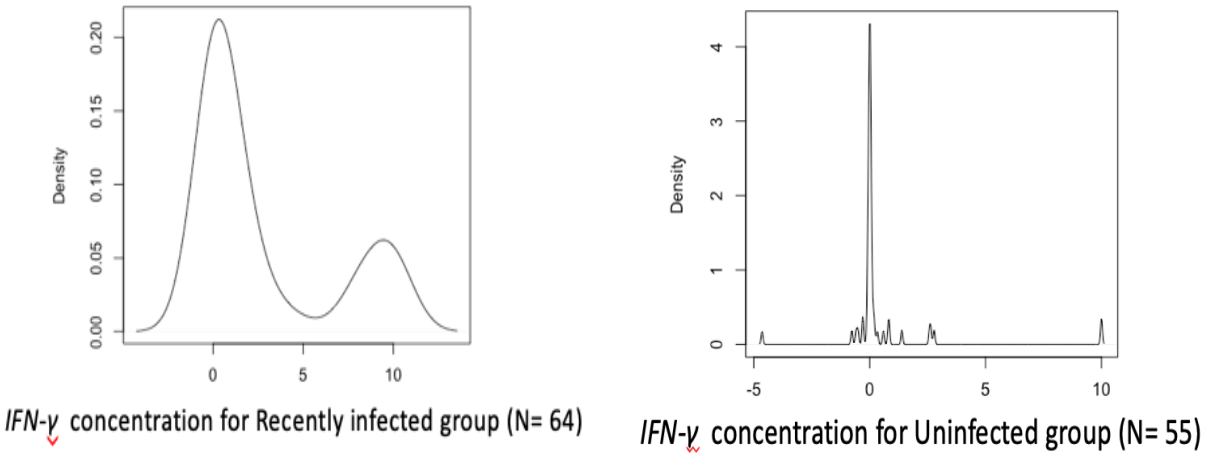


Figure 44: Graph showing concentration of interferon-γ (IFN-γ) density plots for recently infected (left) and uninfected group (right)

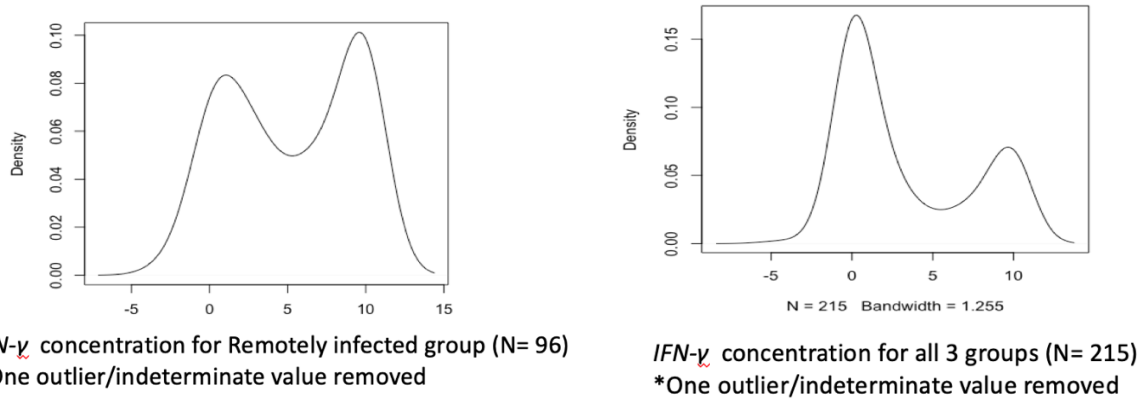


Figure 45: Graph showing concentration of interferon-γ (IFN-γ) density plots for remotely infected (left) and entire the population (right).

*One indeterminate result outlier value not shown.

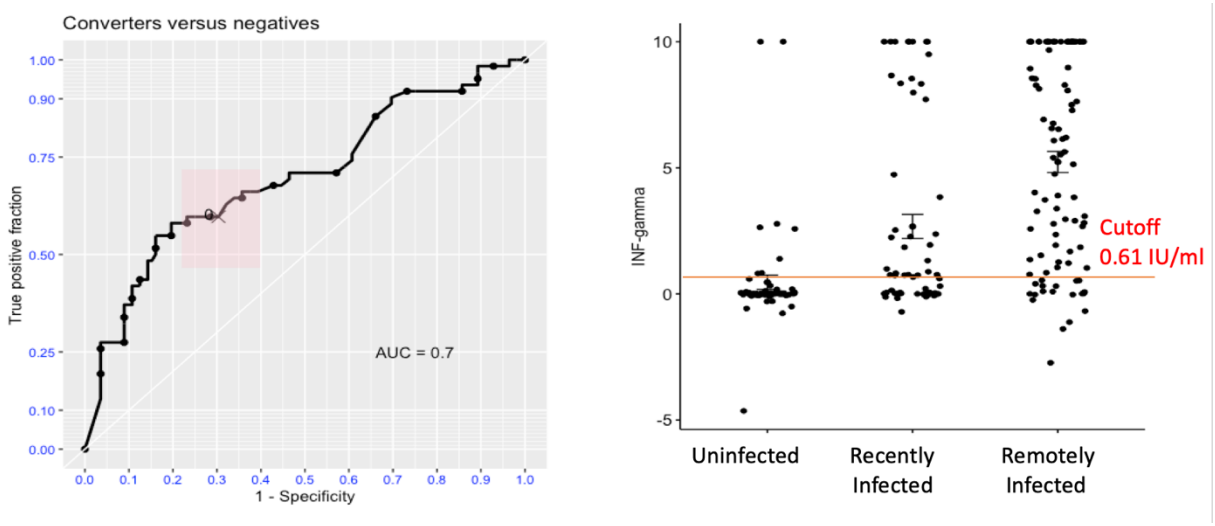


Figure 46: Graph showing potential of concentration of interferon-γ (IFN-γ) to discriminate recently infected and uninfected participants.

*The optimal cut off value is shown on the panel on the right.

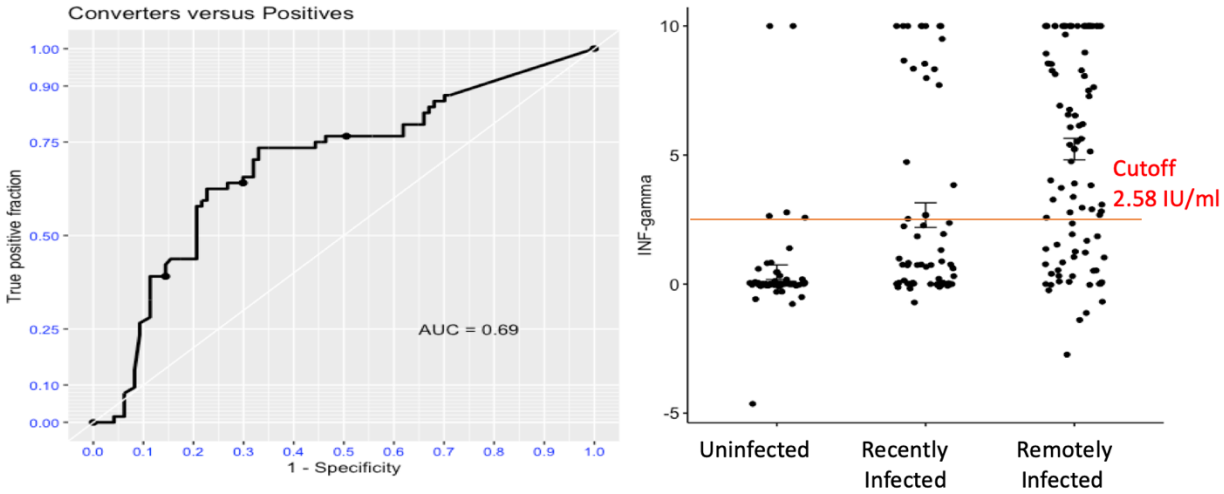


Figure 47: Graph showing potential of concentration of interferon-γ (IFN-γ) to discriminate recently infected and remotely infected participants.

*The optimal cut off value is shown on the panel on the right.

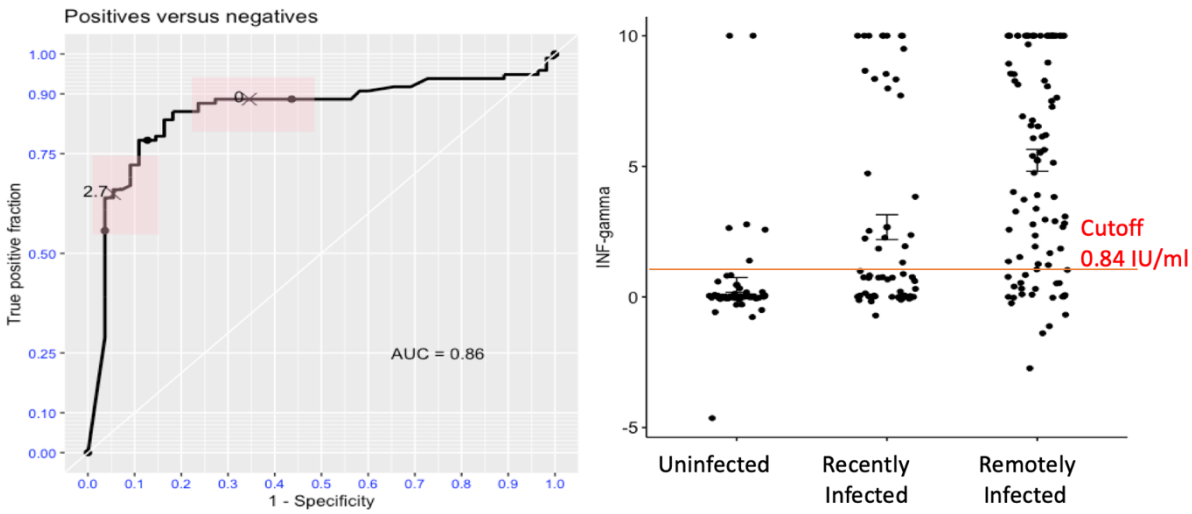


Figure 48: Graph showing potential of concentration of interferon-γ (IFN-γ) to discriminate uninfected participants and remotely infected participants.

*The optimal cut off value is shown on the panel on the right

Chapter 7

RESULTS SYNTHESIS AND CONCLUSIONS

MOTIVATION

The motivation for this study was to identify factors that are associated with transmission of tuberculosis and whether we can exploit such factors to the control of transmission in the community. The current methods used to study transmission using WGS and social network data are sophisticated and require bioinformatics expertise to deploy (Alaridah et al., 2019; Roetzer et al., 2013). In this work, we set out to deploy a basic outbreak investigation method to successfully identify who infected whom in endemic settings using data similar to that collected by TB national control programs.

Developing a new diagnostic assay or deploying the current tests to identify recent *Mtb* infection would allow for targeted treatment of those persons most likely to progress to active TB and is a priority among international TB agencies (Pai & Schito, 2015). To address this challenge, new techniques, including transcript microarrays, flow cytometry of intracellular cytokines, and multiplex micro bead-based immunoassay (Luminex assay) of cytokines, have recently been introduced and used to study TB diagnostics (Berry et al., 2010; Caccamo et al., 2010; Chegou et al., 2009; Sutherland et al., 2010). Unfortunately, the majority of these studies were conducted in cohorts of household or close contacts of TB cases. Whereas the household is an environment of intense transmission of *Mtb*, it does not account for majority of new transmission of tuberculosis yet it remains the focus of most TB transmission studies in the current literature (Crampin et al., 2006; Glynn et al., 2015; Martinez et al., 2017a; Verver et al., 2004; Whalen et al., 2011). There was need to study whether the current methods or novel

methods could differentiate recently infected (TST converted) individuals from remotely infected ones.

MAIN FINDINGS AND THEIR IMPLICATIONS

Aim 1 showed that a basic model originally proposed for outbreak investigation can successfully be deployed to know who infected whom at a defined degree of certainty in a TB endemic setting. When considered individually, WGS or social network/epidemiological approaches alone cannot account for transmission events in most studies. The implication of our findings is that we can now identify transmission links among TB cases in a Sub-Saharan Africa setting given various data sources using a basic method. In this study, we combined social network, molecular epidemiology data and clinical data even though the proposed method has potential for extension to include other routinely collected data by National TB Control Programs. These findings are a solution to the challenge not knowing which individuals are transmitting TB in the community. There are approaches that have been advanced to identify this unknown risk group using Geo-spatial data of locations and settings where TB cases spend most of their time outside of the house using the cellphone data. For instance, in a follow up study, "Mapping TB transmission in Uganda (MATTS study)" cellphones are used to track the path of recently diagnosed TB patients retrospectively or to prospectively to monitor high-risk locations, settings, and contact patterns. This approach promised more data pertinent to answering the question of who infected whom. The use and interpretation of links among cases identified with more data sources will benefit from our suggested 20% certainty score cutoff by highlighting clusters most likely to belong to the same epidemic with a shared index case. This score therefore can be used to pin point clusters of ongoing transmission to public health for effected localized control measures.

Aim 2 demonstrated a significant positive association between being an ancestor MTBC genomes in transmission clusters with carrying mutations in *gyrA, ribD* and *ethR* genes which are vital for the survival and virulence of the bacteria. However ancestor genome status was inversely associated with carrying mutations in the *rrs* and *rrl* genes. This implies that if combined with strain-specific SNP typing and targeted WGS, the mutations identified in this study could be deployed to investigate *Mtb* clusters in an ongoing outbreak or in the study of previous outbreaks in retrospective research.

Aim 3 found that the optimal cutoff for interferon-γ concentration in blood for differentiating LTBI (defined as having a TST reading greater or equal to 10 mm) was 0.21 IU/ml compared to the current 0.35 IU/ml recommended by QFT manufacturers. However, the performance characteristics when the proposed cutoff value was used were not better than when the currently recommended cutoff was used. After adjusting for background interferon-γ levels, the interferon-γ blood levels in converters were intermediate between uninfected and those with established infection. Interferon-γ measurement resulted in the accurate prediction of 69% of recently infected individuals from those that were remotely infected at an optimal cutoff value of 2.58 IU/ml. This implies that interferon-γ can be deployed as a biomarker to identify recently infected individuals, a group most likely to progress to disease. Recent infection remains an important distinction in prioritizing who get Tuberculosis Preventive Therapy (TPT) in endemic settings where treating every one who is infected remain unpractical.

FUTURE DIRECTIONS

By identifying drug resistance markers associated with transmission, this work has provided a methodology to harness the full potential of the rapidly accumulating MTBC Whole Genome Sequence data to understand transmission. A natural progression of this work is to

replicate these findings in larger studies and in different regions where different MTBC lineages are predominant to confirm generalizability. Identified MTBC lineage four strain-specific drug resistance candidate gene mutations could be used to develop SNP-typing assays to rapidly and inexpensively identify participant genomes that are most likely to be transmitted in an endemic Ugandan population, a potential marker for tracking *Mtb* transmission.

The findings that interferon-γ provided satisfactory diagnostic power to differentiate individuals who are recently infected with *Mtb* from remotely infected individuals with a specificity of about 73% and sensitivity of 67% in the community need to be replicated in larger more powered studies. There is also need to study more proteins along with this biomarker to test whether the current performance characteristics can be improved.

Our results demonstrate a potential approach for the investigation of transmission away from the household in a programmatic settings. By conducting our study in a community cohort and not the household of TB cases or their contacts, this study is representative of operational research and edges the field of tuberculosis transmission research closer to a programmatic setting (Cobelens et al., 2012). The deployment of this biomarker in real time to identify hotspots of ongoing infection or individuals with recent infection remain a potential future areas of research.

CONCLUSION

The current TB control measures focusing exclusively on the passive detection and treatment of active cases remains insufficient to increase the rate at which tuberculosis is decreasing world-wide (WHO, 2019). This calls for innovative ways to tackle the problem by focusing on individuals most likely to spread the disease when sick and those most likely to progress to disease when infected within the communities. To respond to that need, this work

has demonstrated a basic model, reliant on sentinel data such as that collected by National TB control programs to identify hotspot of tuberculosis and individuals most likely to spread the disease. Secondly, this work has identified three candidate gene mutations which can be used to identify MTBC genome isolated from individuals who are most likely to transmit the disease. Lastly, this work has demonstrated that the level of interferon-γ in blood after stimulation with *Mtb*-specific antigens could serve as a biomarker to identify recently infected individuals in the community. The major findings of this work are new and warranty further research, and replication in larger study population settings.

REFERENCES

Alangaden, G. J., Kreiswirth, B. N., Aouad, A., Khetarpal, M., Igno, F. R., Moghazeh, S. L., … Lerner, S. A. (1998). Mechanism of resistance to amikacin and kanamycin in Mycobacterium tuberculosis. *Antimicrobial Agents and Chemotherapy*, *42*(5), 1295–1297. https://doi.org/10.1128/aac.42.5.1295

Alaridah, N., Hallbäck, E. T., Tångrot, J., Winqvist, N., Sturegård, E., Florén-Johansson, K., … Godaly, G. (2019). Transmission dynamics study of tuberculosis isolates with whole genome sequencing in southern Sweden. *Scientific Reports*, *9*(1). https://doi.org/10.1038/s41598-019-39971-z

Almeida, A. S., Lago, P. M., Boechat, N., Huard, R. C., Lazzarini, L. C. O., Santos, A. R., … Ho, J. L. (2009). Tuberculosis Is Associated with a Down-Modulatory Lung Immune Response That Impairs Th1-Type Immunity. *The Journal of Immunology*, *183*(1), 718–731. https://doi.org/10.4049/jimmunol.0801212

Andersen, P., Munk, M. E., Pollock, J. M., & Doherty, T. M. (2000a). Specific immune-based diagnosis of tuberculosis. *Lancet (London, England)*, *356*(9235), 1099–1104. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11009160

Andersen, P., Munk, M. E., Pollock, J. M., & Doherty, T. M. (2000b, September 23). Specific immune-based diagnosis of tuberculosis. *Lancet*. Elsevier Limited. https://doi.org/10.1016/S0140-6736(00)02742-2

Antonucci, G., Girardi, E., Raviglione, M. C., & Ippolito, G. (1995). Risk factors for tuberculosis in HIV-infected persons. A prospective cohort study. The Gruppo Italiano di Studio Tubercolosi e AIDS (GISTA). *JAMA*, *274*(2), 143–148. Retrieved from

http://www.ncbi.nlm.nih.gov/pubmed/7596002

Applied Logistic Regression - David W. Hosmer, Jr., Stanley Lemeshow, Rodney X. Sturdivant - Google Books. (n.d.). Retrieved March 26, 2020, from https://books.google.com/books?hl=en&lr=&id=64JYAwAAQBAJ&oi=fnd&pg=PR13&ots=DsjP4-4okO&sig=TmgabF74gKo1z-UfLjZci9RosCU#v=onepage&q&f=false

Asiimwe, B. B., Joloba, M. L., Ghebremichael, S., Koivula, T., Kateete, D. P., Katabazi, F. A., … Kallenius, G. (2009). DNA restriction fragment length polymorphism analysis of Mycobacterium tuberculosis isolates from HIV-seropositive and HIV-seronegative patients in Kampala, Uganda. *BMC Infectious Diseases*, *9*(1), 12. https://doi.org/10.1186/1471-2334-9-12

Auld, S. C., Shah, N. S., Mathema, B., Brown, T. S., Ismail, N., Omar, S. V., … Gandhi, N. R. (2018). Extensively drug-resistant tuberculosis in South Africa: genomic evidence supporting transmission in communities. *The European Respiratory Journal*, *52*(4). https://doi.org/10.1183/13993003.00246-2018

Azzurri, A., Sow, O. Y., Amedei, A., Bah, B., Diallo, S., Peri, G., … Del Prete, G. (2005). IFN-γ-inducible protein 10 and pentraxin 3 plasma levels are tools for monitoring inflammation and disease activity in Mycobacterium tuberculosis infection. *Microbes and Infection*, *7*(1), 1–8. https://doi.org/10.1016/j.micinf.2004.09.004

Bainomugisa, A., Lavu, E., Hiashiri, S., Majumdar, S., Honjepari, A., Moke, R., … 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*(2). https://doi.org/10.1099/mgen.0.000147

Bark, C. M., Manceur, A. M., Malone, L. L., Nsereko, M., Okware, B., Mayanja, H. K., …

Paramithiotis, E. (2017). Identification of Host Proteins Predictive of Early Stage Mycobacterium tuberculosis Infection. *EBioMedicine*, *21*, 150–157. https://doi.org/10.1016/j.ebiom.2017.06.019

Barnes, P. F., Yang, Z., Preston-Martin, S., Pogoda, J. M., Jones, B. E., Otaya, M., … Cave, M. D. (1997). Patterns of tuberculosis transmission in Central Los Angeles. *JAMA*, *278*(14), 1159–1163. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9326475

Barrick, J. E., & Lenski, R. E. (2013, December). Genome dynamics during experimental evolution. *Nature Reviews Genetics*. https://doi.org/10.1038/nrg3564

Bates, M. N., Khalakdina, A., Pai, M., Chang, L., Lessa, F., & Smith, K. R. (2007). Risk of Tuberculosis From Exposure to Tobacco Smoke. *Archives of Internal Medicine*, *167*(4), 335. https://doi.org/10.1001/archinte.167.4.335

Belanger, A. E., Besra, G. S., Ford, M. E., Mikušová, K., Belisle, J. T., Brennan, P. J., & Inamine, J. M. (1996). The embAB genes of Mycobacterium avium encode an arabinosyl transferase involved in cell wall arabinan biosynthesis that is the target for the antimycobacterial drug ethambutol. *Proceedings of the National Academy of Sciences of the United States of America*, *93*(21), 11919–11924. https://doi.org/10.1073/pnas.93.21.11919

Bennett, P. M. (2008). Plasmid encoded antibiotic resistance: acquisition and transfer of antibiotic resistance genes in bacteria. *British Journal of Pharmacology*, *153 Suppl 1*(Suppl 1), S347-57. https://doi.org/10.1038/sj.bjp.0707607

Berkel, G. M., Cobelens, F. G. J., de Vries, G., Draayer-Jansen, I. W. E., & Borgdorff, M. W. (2005). Tuberculin skin test: estimation of positive and negative predictive values from routine data. *The International Journal of Tuberculosis and Lung Disease : The Official Journal of the International Union against Tuberculosis and Lung Disease*, *9*(3), 310–316.

Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/15786896

Berry, M. P. R., Graham, C. M., McNab, F. W., Xu, Z., Bloch, S. A. A., Oni, T., … O'Garra, A. (2010). An interferon-inducible neutrophil-driven blood transcriptional signature in human tuberculosis. *Nature*, *466*(7309), 973–977. https://doi.org/10.1038/nature09247

Biraro, I. A., Kimuda, S., Egesa, M., Cose, S., Webb, E. L., Joloba, M., … Katamba, A. (2016a). The Use of Interferon Gamma Inducible Protein 10 as a Potential Biomarker in the Diagnosis of Latent Tuberculosis Infection in Uganda. *PloS One*, *11*(1), e0146098. https://doi.org/10.1371/journal.pone.0146098

Biraro, I. A., Kimuda, S., Egesa, M., Cose, S., Webb, E. L., Joloba, M., … Katamba, A. (2016b). The Use of Interferon Gamma Inducible Protein 10 as a Potential Biomarker in the Diagnosis of Latent Tuberculosis Infection in Uganda. *PloS One*, *11*(1), e0146098. https://doi.org/10.1371/journal.pone.0146098

Bjorn-Mortensen, K., Soborg, B., Koch, A., Ladefoged, K., Merker, M., Lillebaek, T., … Kohl, T. A. (2016). Tracing Mycobacterium tuberculosis transmission by whole genome sequencing in a high incidence setting: a retrospective population-based study in East Greenland. *Scientific Reports*, *6*(1), 33180. https://doi.org/10.1038/srep33180

Bobadilla, K., Sada, E., Jaime, M. E., González, Y., Ramachandra, L., Rojas, R. E., … Torres, M. (2013). Human phagosome processing of M *ycobacterium tuberculosis* antigens is modulated by interferon-γ and interleukin-10. *Immunology*, *138*(1), 34–46. https://doi.org/10.1111/imm.12010

Borrell, S., Teo, Y., Giardina, F., Streicher, E. M., Klopper, M., Feldmann, J., … Gagneux, S. (2013). Epistasis between antibiotic resistance mutations drives the evolution of extensively drug-resistant tuberculosis. *Evolution, Medicine and Public Health*, *2013*(1), 65–74.

https://doi.org/10.1093/emph/eot003

Bos, K. I., Harkins, K. M., Herbig, A., Coscolla, M., Weber, N., Comas, I., … Krause, J. (2014). Pre-Columbian mycobacterial genomes reveal seals as a source of New World human tuberculosis. *Nature*, *514*(7253), 494–497. https://doi.org/10.1038/nature13591

Bradley, P., Gordon, N. C., Walker, T. M., Dunn, L., Heys, S., Huang, B., … Iqbal, Z. (2015). Rapid antibiotic-resistance predictions from genome sequence data for Staphylococcus aureus and Mycobacterium tuberculosis. *Nature Communications*, *6*. https://doi.org/10.1038/ncomms10063

Brosch, R., Gordon, S. V., Marmiesse, M., Brodin, P., Buchrieser, C., Eiglmeier, K., … Cole, S. T. (2002). A new evolutionary scenario for the Mycobacterium tuberculosis complex. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(6), 3684–3689. https://doi.org/10.1073/pnas.052548299

Caccamo, N., Guggino, G., Joosten, S. A., Gelsomino, G., Di Carlo, P., Titone, L., … Dieli, F. (2010). Multifunctional CD4+ T cells correlate with active Mycobacterium tuberculosis infection. *European Journal of Immunology*, *40*(8), 2211–2220. https://doi.org/10.1002/eji.201040455

Castellanos, M. E., Kirimunda, S., Martinez, L., Quach, T., Woldu, H., Kakaire, R., … Whalen, C. C. (2018). Performance of the QuantiFERON ® -TB Gold In-Tube assay in tuberculin skin test converters: a prospective cohort study. *The International Journal of Tuberculosis and Lung Disease*, *22*(9), 1000–1006. https://doi.org/10.5588/ijtld.18.0073

Chegou, N. N., Black, G. F., Kidd, M., van Helden, P. D., & Walzl, G. (2009). Host markers in Quantiferon supernatants differentiate active TB from latent TB infection: Preliminary report. *BMC Pulmonary Medicine*, *9*, 21. https://doi.org/10.1186/1471-2466-9-21

Chia, S., Karim, M., Elwood, R. K., & FitzGerald, J. M. (1998). Risk of tuberculosis in dialysis patients: a population-based study. *The International Journal of Tuberculosis and Lung Disease : The Official Journal of the International Union against Tuberculosis and Lung Disease*, *2*(12), 989–991. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9869114

Chieppa, M., Bianchi, G., Doni, A., Del Prete, A., Sironi, M., Laskarin, G., … Allavena, P. (2003). Cross-linking of the mannose receptor on monocyte-derived dendritic cells activates an anti-inflammatory immunosuppressive program. *Journal of Immunology (Baltimore, Md. : 1950)*, *171*(9), 4552–4560. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/14568928

Chin, D. P., DeRiemer, K., Small, P. M., De Leon, A. P., Steinhart, R., Schecter, G. F., … Hopewell, P. C. (1998). Differences in contributing factors to tuberculosis incidence in U.S.- born and foreign-born persons. *American Journal of Respiratory and Critical Care Medicine*, *158*(6), 1797–1803. https://doi.org/10.1164/ajrccm.158.6.9804029

Clark, T. G., Mallard, K., Coll, F., Preston, M., Assefa, S., Harris, D., … McNerney, R. (2013). Elucidating Emergence and Transmission of Multidrug-Resistant Tuberculosis in Treatment Experienced Patients by Whole Genome Sequencing. *PLoS ONE*, *8*(12), e83012. https://doi.org/10.1371/journal.pone.0083012

Cobelens, F., van Kampen, S., Ochodo, E., Atun, R., & Lienhardt, C. (2012). Research on Implementation of Interventions in Tuberculosis Control in Low- and Middle-Income Countries: A Systematic Review. *PLoS Medicine*, *9*(12), e1001358. https://doi.org/10.1371/journal.pmed.1001358

Coll, F., McNerney, R., Preston, M. D., Guerra-Assunção, J. A., Warry, A., Hill-Cawthorne, G., … Clark, T. G. (2015). Rapid determination of anti-tuberculosis drug resistance from

whole-genome sequences. *Genome Medicine*, *7*(1). https://doi.org/10.1186/s13073-015-0164-0

Coll, F., Phelan, J., Hill-Cawthorne, G. A., Nair, M. B., Mallard, K., Ali, S., … Clark, T. G. (2018). Genome-wide analysis of multi- and extensively drug-resistant Mycobacterium tuberculosis. *Nature Genetics*, *50*(2), 307–316. https://doi.org/10.1038/s41588-017-0029-0

Comas, I., Borrell, S., Roetzer, A., Rose, G., Malla, B., Kato-Maeda, M., … Gagneux, S. (2012). Whole-genome sequencing of rifampicin-resistant Mycobacterium tuberculosis strains identifies compensatory mutations in RNA polymerase genes. *Nature Genetics*, *44*(1), 106–110. https://doi.org/10.1038/ng.1038

Committee on Infectious Diseases; American Academy of Pediatrics; David W. Kimberlin, MD, FAAP; Michael T. Brady, MD, FAAP; Mary Anne Jackson, M. (n.d.). Red Book® 2015 | Red Book Online | AAP Point-of-Care-Solutions. Retrieved September 14, 2018, from https://redbook.solutions.aap.org/book.aspx?bookid=1484

Cook, V. J., Sun, S. J., Tapia, J., Muth, S. Q., Argüello, D. F., Lewis, B. L., … McElroy, P. D. (2007). Transmission Network Analysis in Tuberculosis Contact Investigations. *The Journal of Infectious Diseases*, *196*(10), 1517–1527. https://doi.org/10.1086/523109

Cottam, E. M., Wadsworth, J., Shaw, A. E., Rowlands, R. J., Goatley, L., Maan, S., … Knowles, N. J. (2008). Transmission Pathways of Foot-and-Mouth Disease Virus in the United Kingdom in 2007. *PLoS Pathogens*, *4*(4), e1000050. https://doi.org/10.1371/journal.ppat.1000050

Crampin, A. C., Glynn, J. R., Traore, H., Yates, M. D., Mwaungulu, L., Mwenebabu, M., … Fine, P. E. M. (2006). Tuberculosis transmission attributable to close contacts and HIV status, Malawi. *Emerging Infectious Diseases*, *12*(5), 729–735.

https://doi.org/10.3201/eid1205.050789

Dahl, K. E., Shiratsuchi, H., Hamilton, B. D., Ellner, J. J., & Toossi, Z. (1996). Selective induction of transforming growth factor beta in human monocytes by lipoarabinomannan of Mycobacterium tuberculosis. *Infection and Immunity*, *64*(2), 399–405. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8550183

DeBarber, A. E., Mdluli, K., Bosman, M., Bekker, L. G., & Barry, C. E. (2000). Ethionamide activation and sensitivity in multidrug-resistant Mycobacterium tuberculosis. *Proceedings of the National Academy of Sciences of the United States of America*, *97*(17), 9677–9682. https://doi.org/10.1073/pnas.97.17.9677

Dejesus, M. A., Gerrick, E. R., Xu, W., Park, S. W., Long, J. E., Boutte, C. C., … Ioerger, T. R. (2017). Comprehensive essentiality analysis of the Mycobacterium tuberculosis genome via saturating transposon mutagenesis. *MBio*, *8*(1). https://doi.org/10.1128/mBio.02133-16

Delgado-Rodríguez, M., & Llorca, J. (2004). Bias. *Journal of Epidemiology and Community Health*, *58*(8), 635–641. https://doi.org/10.1136/jech.2003.008466

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V, Maguire, J. R., Hartl, C., … Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, *43*(5), 491–498. https://doi.org/10.1038/ng.806

Drummond, A. J., & Rambaut, A. (2007). BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology*, *7*(1), 214. https://doi.org/10.1186/1471-2148-7-214

Dufour, J. H., Dziejman, M., Liu, M. T., Leung, J. H., Lane, T. E., & Luster, A. D. (2002). IFN-gamma-inducible protein 10 (IP-10; CXCL10)-deficient mice reveal a role for IP-10 in effector T cell generation and trafficking. *Journal of Immunology (Baltimore, Md. : 1950)*, *168*(7), 3195–3204. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11907072

Dunn, J. J., Starke, J. R., & Revell, P. A. (2016). Laboratory Diagnosis of Mycobacterium tuberculosis Infection and Disease in Children. *Journal of Clinical Microbiology*, *54*(6), 1434–1441. https://doi.org/10.1128/JCM.03043-15

Dye, C., Williams, B. G., Espinal, M. A., & Raviglione, M. C. (2002, March 15). Erasing the world's slow stain: Strategies to beat multidrug-resistant tuberculosis. *Science*. https://doi.org/10.1126/science.1063814

Efron, B. (1986). How Biased is the Apparent Error Rate of a Prediction Rule? *Journal of the American Statistical Association*, *81*(394), 461–470. https://doi.org/10.1080/01621459.1986.10478291

Eldholm, V., & Balloux, F. (2016, August 1). Antimicrobial Resistance in Mycobacterium tuberculosis: The Odd One Out. *Trends in Microbiology*. Elsevier Ltd. https://doi.org/10.1016/j.tim.2016.03.007

Eldholm, V., Monteserin, J., Rieux, A., Lopez, B., Sobkowiak, B., Ritacco, V., & Balloux, F. (2015). Four decades of transmission of a multidrug-resistant Mycobacterium tuberculosis outbreak strain. *Nature Communications*, *6*(1), 7119. https://doi.org/10.1038/ncomms8119

Eppinger, M., Pearson, T., Koenig, S. S. K., Pearson, O., Hicks, N., Agrawal, S., … Keim, P. S. (2014). Genomic epidemiology of the Haitian cholera outbreak: a single introduction followed by rapid, extensive, and continued spread characterized the onset of the epidemic. *MBio*, *5*(6), e01721. https://doi.org/10.1128/mBio.01721-14

Escombe, A. R., Moore, D. A. J., Gilman, R. H., Pan, W., Navincopa, M., Ticona, E., … Evans, C. A. (2008). The Infectiousness of Tuberculosis Patients Coinfected with HIV. *PLoS Medicine*, *5*(9), e188. https://doi.org/10.1371/journal.pmed.0050188

Farber, J. M. (1997). Mig and IP-10: CXC chemokines that target lymphocytes. *Journal of*

Leukocyte Biology, *61*(3), 246–257. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9060447

Farhat, M., Greenaway, C., Pai, M., & Menzies, D. (2006). False-positive tuberculin skin tests: what is the absolute effect of BCG and non-tuberculous mycobacteria? *The International Journal of Tuberculosis and Lung Disease : The Official Journal of the International Union against Tuberculosis and Lung Disease*, *10*(11), 1192–1204. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/17131776

Ferrero, E., Biswas, P., Vettoretto, K., Ferrarini, M., Uguccioni, M., Piali, L., … Pardi, R. (2003). Macrophages exposed to Mycobacterium tuberculosis release chemokines able to recruit selected leucocyte subpopulations: focus on gammadelta cells. *Immunology*, *108*(3), 365–374. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12603603

Feuerriegel, S., Schleusener, V., Beckert, P., Kohl, T. A., Miotto, P., Cirillo, D. M., … Fellenberg, K. (2015). PhyResSE: A web tool delineating Mycobacterium tuberculosis antibiotic resistance and lineage from whole-genome sequencing data. *Journal of Clinical Microbiology*, *53*(6), 1908–1914. https://doi.org/10.1128/JCM.00025-15

Fitzpatrick, L. K., Hardacker, J. A., Heirendt, W., Agerton, T., Streicher, A., Melnyk, H., … Onorato, I. (2001). A Preventable Outbreak of Tuberculosis Investigated through an Intricate Social Network. *Clinical Infectious Diseases*, *33*(11), 1801–1806. https://doi.org/10.1086/323671

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). https://doi.org/10.1038/s41598-018-30363-3

Gagneux, S. (2018, April 1). Ecology and evolution of Mycobacterium tuberculosis. *Nature Reviews Microbiology*. Nature Publishing Group. https://doi.org/10.1038/nrmicro.2018.8

Gagneux, S., DeRiemer, K., Van, T., Kato-Maeda, M., De Jong, B. C., Narayanan, S., … Small, P. M. (2006). Variable host-pathogen compatibility in Mycobacterium tuberculosis. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(8), 2869–2873. https://doi.org/10.1073/pnas.0511240103

Galvani, A. P., & May, R. M. (2005, November 17). Epidemiology: Dimensions of superspreading. *Nature*. Nature Publishing Group. https://doi.org/10.1038/438293a

García Jacobo, R. E., Serrano, C. J., Enciso Moreno, J. A., Gaspar Ramírez, O., Trujillo Ochoa, J. L., Uresti Rivera, E. E., … García Hernández, M. H. (2014). Analysis of Th1, Th17 and regulatory T cells in tuberculosis case contacts. *Cellular Immunology*, *289*(1–2), 167–173. https://doi.org/10.1016/j.cellimm.2014.03.010

Gardy, J. L., Johnston, J. C., Sui, S. J. H., Cook, V. J., Shah, L., Brodkin, E., … Tang, P. (2011a). Whole-Genome Sequencing and Social-Network Analysis of a Tuberculosis Outbreak. *New England Journal of Medicine*, *364*(8), 730–739. https://doi.org/10.1056/NEJMoa1003176

Gardy, J. L., Johnston, J. C., Sui, S. J. H., Cook, V. J., Shah, L., Brodkin, E., … Tang, P. (2011b). Whole-Genome Sequencing and Social-Network Analysis of a Tuberculosis Outbreak. *New England Journal of Medicine*, *364*(8), 730–739. https://doi.org/10.1056/NEJMoa1003176

Garg, A., Barnes, P. F., Roy, S., Quiroga, M. F., Wu, S., García, V. E., … Vankayalapati, R. (2008). Mannose-capped lipoarabinomannan- and prostaglandin E2-dependent expansion of regulatory T cells in humanMycobacterium tuberculosisinfection. *European Journal of*

Immunology, *38*(2), 459–469. https://doi.org/10.1002/eji.200737268

Glass, K., Mercer, G. N., Nishiura, H., McBryde, E. S., & Becker, N. G. (2011). Estimating reproduction numbers for adults and children from case data. *Journal of the Royal Society, Interface*, *8*(62), 1248–1259. https://doi.org/10.1098/rsif.2010.0679

Glynn, J. R., Guerra-Assunção, J. A., Houben, R. M. G. J., Sichali, L., Mzembe, T., Mwaungulu, L. K., … Clark, T. G. (2015). Whole Genome Sequencing Shows a Low Proportion of Tuberculosis Disease Is Attributable to Known Close Contacts in Rural Malawi. *PLOS ONE*, *10*(7), e0132840. https://doi.org/10.1371/journal.pone.0132840

Grenfell, B. T., Pybus, O. G., Gog, J. R., Wood, J. L. N., Daly, J. M., Mumford, J. A., & Holmes, E. C. (2004, January 16). Unifying the Epidemiological and Evolutionary Dynamics of Pathogens. *Science*. https://doi.org/10.1126/science.1090727

Guthrie, J. L., Kong, C., Roth, D., Jorgensen, D., Rodrigues, M., Hoang, L., … Gardy, J. L. (2018). Molecular Epidemiology of Tuberculosis in British Columbia, Canada: A 10-Year Retrospective Study. *Clinical Infectious Diseases : An Official Publication of the Infectious Diseases Society of America*, *66*(6), 849–856. https://doi.org/10.1093/cid/cix906

Hatherell, H. A., Colijn, C., Stagg, H. R., Jackson, C., Winter, J. R., & Abubakar, I. (2016). Interpreting whole genome sequencing for investigating tuberculosis transmission: A systematic review. *BMC Medicine*, *14*(1). https://doi.org/10.1186/s12916-016-0566-x

Hirsch, C. S., Hussain, R., Toossi, Z., Dawood, G., Shahid, F., & Ellner, J. J. (1996). Cross-modulation by transforming growth factor beta in human tuberculosis: suppression of antigen-driven blastogenesis and interferon gamma production. *Proceedings of the National Academy of Sciences of the United States of America*, *93*(8), 3193–3198. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8622912

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*(10), e1002152. https://doi.org/10.1371/journal.pmed.1002152

Hussain, R., Talat, N., Shahid, F., & Dawood, G. (2009). Biomarker Changes Associated with Tuberculin Skin Test (TST) Conversion: A Two-Year Longitudinal Follow-Up Study in Exposed Household Contacts. *PLoS ONE*, *4*(10), e7444. https://doi.org/10.1371/journal.pone.0007444

Iwai, H., Kato-Miyazawa, M., Kirikae, T., & Miyoshi-Akiyama, T. (2015, December 1). CASTB (the comprehensive analysis server for the Mycobacterium tuberculosis complex): A publicly accessible web server for epidemiological analyses, drug-resistance prediction and phylogenetic comparison of clinical isolates. *Tuberculosis*. Churchill Livingstone. https://doi.org/10.1016/j.tube.2015.09.002

Jajou, R., Neeling, A. de, Hunen, R. van, Vries, G. de, Schimmel, H., Mulder, A., … Soolingen, D. van. (2018). Epidemiological links between tuberculosis cases identified twice as efficiently by whole genome sequencing than conventional molecular typing: A population-based study. *PLOS ONE*, *13*(4), e0195413. https://doi.org/10.1371/journal.pone.0195413

Jeon, C. Y., & Murray, M. B. (2008). Diabetes Mellitus Increases the Risk of Active Tuberculosis: A Systematic Review of 13 Observational Studies. *PLoS Medicine*, *5*(7), e152. https://doi.org/10.1371/journal.pmed.0050152

Johansen, S. K., Maus, C. E., Plikaytis, B. B., & Douthwaite, S. (2006). Capreomycin Binds across the Ribosomal Subunit Interface Using tlyA-Encoded 2′-O-Methylations in 16S and 23S rRNAs. *Molecular Cell*, *23*(2), 173–182. https://doi.org/10.1016/j.molcel.2006.05.044

Jombart, T., Eggo, R. M., Dodd, P. J., & Balloux, F. (2011). Reconstructing disease outbreaks

from genetic data: A graph approach. *Heredity*, *106*(2), 383–390. https://doi.org/10.1038/hdy.2010.78

Kang, Y. A., Lee, H. W., Hwang, S. S., Um, S. W., Han, S. K., Shim, Y. S., & Yim, J. J. (2007). Usefulness of whole-blood interferon-γ assay and interferon-γ enzyme-linked immunospot assay in the diagnosis of active pulmonary tuberculosis. *Chest*, *132*(3), 959–965. https://doi.org/10.1378/chest.06-2805

Kaplan, G., Luster, A. D., Hancock, G., & Cohn, Z. A. (1987). The expression of a gamma interferon-induced protein (IP-10) in delayed immune responses in human skin. *The Journal of Experimental Medicine*, *166*(4), 1098–1108. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/2443597

Kim, K., Perera, R., Tan, D. B. A., Fernandez, S., Seddiki, N., Waring, J., & French, M. A. (2014). Circulating mycobacterial-reactive CD4+ T cells with an immunosuppressive phenotype are higher in active tuberculosis than latent tuberculosis infection. *Tuberculosis*, *94*(5), 494–501. https://doi.org/10.1016/j.tube.2014.07.002

Kline, S. E., Hedemark, L. L., & Davies, S. F. (1995). Outbreak of tuberculosis among regular patrons of a neighborhood bar. *New England Journal of Medicine*, *333*(4), 222–227. https://doi.org/10.1056/NEJM199507273330404

Klovdahl, A. S., Graviss, E. A., Yaganehdoost, A., Ross, M. W., Wanger, A., Adams, G. J., & Musser, J. M. (2001). Networks and tuberculosis: an undetected community outbreak involving public places. *Social Science & Medicine (1982)*, *52*(5), 681–694. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11218173

Landry, J., & Menzies, D. (2008). Preventive chemotherapy. Where has it got us? Where to go next? *International Journal of Tuberculosis and Lung Disease*, *12*(12), 1352–1364.

Leinonen, R., Sugawara, H., & Shumway, M. (2011). The sequence read archive. *Nucleic Acids Research*, *39*(SUPPL. 1). https://doi.org/10.1093/nar/gkq1019

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, *25*(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Lim, H. J., Okwera, A., Mayanja-Kizza, H., Ellner, J. J., Mugerwa, R. D., & Whalen, C. C. (2006). Effect of tuberculosis preventive therapy on HIV disease progression and survival in HIV-infected adults. *HIV Clinical Trials*, *7*(4), 172–183. https://doi.org/10.1310/hct0704-172

Lin, P. L., & Flynn, J. L. (2010). Understanding Latent Tuberculosis: A Moving Target. *The Journal of Immunology*, *185*(1), 15–22. https://doi.org/10.4049/jimmunol.0903856

Lloyd-Smith, J. O., Schreiber, S. J., Kopp, P. E., & Getz, W. M. (2005). Superspreading and the effect of individual variation on disease emergence. *Nature*, *438*(7066), 355–359. https://doi.org/10.1038/nature04153

Lonnroth, K., Williams, B. G., Cegielski, P., & Dye, C. (2010). A consistent log-linear relationship between tuberculosis incidence and body mass index. *International Journal of Epidemiology*, *39*(1), 149–155. https://doi.org/10.1093/ije/dyp308

Lukoye, D., Adatu, F., Musisi, K., Kasule, G. W., Were, W., Odeke, R., … Joloba, M. L. (2013). Anti-Tuberculosis Drug Resistance among New and Previously Treated Sputum Smear-Positive Tuberculosis Patients in Uganda: Results of the First National Survey. *PLoS ONE*, *8*(8), e70763. https://doi.org/10.1371/journal.pone.0070763

Luster, A. D., Unkeless, J. C., & Ravetch, J. V. (n.d.). Gamma-interferon transcriptionally regulates an early-response gene containing homology to platelet proteins. *Nature*,

315(6021), 672–676. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/3925348

Manson, A. L., Cohen, K. A., Abeel, T., Desjardins, C. A., Armstrong, D. T., Barry, C. E., … Earl, A. M. (2017). Genomic analysis of globally diverse Mycobacterium tuberculosis strains provides insights into the emergence and spread of multidrug resistance. *Nature Genetics*, *49*(3), 395–402. https://doi.org/10.1038/ng.3767

Marais, B. J., Gie, R. P., Schaaf, H. S., Hesseling, A. C., Obihara, C. C., Starke, J. J., … Beyers, N. (2004). The natural history of childhood intra-thoracic tuberculosis: a critical review of literature from the pre-chemotherapy era. *The International Journal of Tuberculosis and Lung Disease : The Official Journal of the International Union against Tuberculosis and Lung Disease*, *8*(4), 392–402. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/15141729

Martinez, L., Shen, Y., Mupere, E., Kizza, A., Hill, P. C., & Whalen, C. C. (2017a). Transmission of Mycobacterium Tuberculosis in Households and the Community: A Systematic Review and Meta-Analysis. *American Journal of Epidemiology*, *185*(12), 1327–1339. https://doi.org/10.1093/aje/kwx025

Martinez, L., Shen, Y., Mupere, E., Kizza, A., Hill, P. C., & Whalen, C. C. (2017b). Transmission of Mycobacterium Tuberculosis in Households and the Community: A Systematic Review and Meta-Analysis. In *American Journal of Epidemiology* (Vol. 185, pp. 1327–1339). Oxford University Press. https://doi.org/10.1093/aje/kwx025

McElroy, P. D., Rothenberg, R. B., Varghese, R., Woodruff, R., Minns, G. O., Muth, S. Q., … Ridzon, R. (2003). A network-informed approach to investigating a tuberculosis outbreak: implications for enhancing contact investigations. *The International Journal of Tuberculosis and Lung Disease : The Official Journal of the International Union against Tuberculosis*

and Lung Disease, *7*(12 Suppl 3), S486-93. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/14677842

Melsew, Y. A., Gambhir, M., Cheng, A. C., McBryde, E. S., Denholm, J. T., Tay, E. L., & Trauer, J. M. (2019). The role of super-spreading events in Mycobacterium tuberculosis transmission: Evidence from contact tracing. *BMC Infectious Diseases*, *19*(1). https://doi.org/10.1186/s12879-019-3870-1

Menzies, N. A., Wolf, E., Connors, D., Bellerose, M., Sbarra, A. N., Cohen, T., … Salomon, J. A. (2018). Progression from latent infection to active disease in dynamic tuberculosis transmission models: a systematic review of the validity of modelling assumptions. *The Lancet. Infectious Diseases*, *18*(8), e228–e238. https://doi.org/10.1016/S1473-3099(18)30134-8

Merker, J. D., Wenger, A. M., Sneddon, T., Grove, M., Zappala, Z., Fresard, L., … Ashley, E. A. (2018). Long-read genome sequencing identifies causal structural variation in a Mendelian disease. *Genetics in Medicine*, *20*(1), 159–163. https://doi.org/10.1038/gim.2017.86

Meurer, W. J., & Tolles, J. (2017). Logistic Regression Diagnostics. *JAMA*, *317*(10), 1068. https://doi.org/10.1001/jama.2016.20441

Murray, G. G. R., Wang, F., Harrison, E. M., Paterson, G. K., Mather, A. E., Harris, S. R., … Welch, J. J. (2016). The effect of genetic structure on molecular dating and tests for temporal signal. *Methods in Ecology and Evolution*, *7*(1), 80–89. https://doi.org/10.1111/2041-210X.12466

Murray, M., & Alland, D. (2002). Methodological problems in the molecular epidemiology of tuberculosis. *American Journal of Epidemiology*, *155*(6), 565–571. https://doi.org/10.1093/aje/155.6.565

Mutreja, A., Kim, D. W., Thomson, N. R., Connor, T. R., Lee, J. H., Kariuki, S., … Dougan, G. (2011). Evidence for several waves of global transmission in the seventh cholera pandemic. *Nature*, *477*(7365), 462–465. https://doi.org/10.1038/nature10392

Newman, M. (2010). *Networks*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199206650.001.0001

Orme, I. M., & Cooper, A. M. (1999). Cytokine/chemokine cascades in immunity to tuberculosis. *Immunology Today*, *20*(7), 307–312. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10379048

Ou, C. Y., Ciesielski, C. A., Myers, G., Bandea, C. I., Luo, C. C., Korber, B. T., … Economou, A. N. (1992). Molecular epidemiology of HIV transmission in a dental practice. *Science (New York, N.Y.)*, *256*(5060), 1165–1171. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/1589796

Pai, M., Denkinger, C. M., Kik, S. V., Rangaka, M. X., Zwerling, A., Oxlade, O., … Banaei, N. (2014). Gamma interferon release assays for detection of Mycobacterium tuberculosis infection. *Clinical Microbiology Reviews*, *27*(1), 3–20. https://doi.org/10.1128/CMR.00034-13

Pai, M., & Schito, M. (2015). Tuberculosis Diagnostics in 2015: Landscape, Priorities, Needs, and Prospects. *The Journal of Infectious Diseases*, *211*(suppl_2), S21–S28. https://doi.org/10.1093/infdis/jiu803

Poon, A. F. Y. (2016). Impacts and shortcomings of genetic clustering methods for infectious disease outbreaks. *Virus Evolution*, *2*(2). https://doi.org/10.1093/VE/VEW031

Qiagen. (n.d.). QuantiFERON-technology - QIAGEN. Retrieved August 27, 2018, from https://www.qiagen.com/kr/resources/technologies/qft_technology-spotlightpages/

Ragonnet, R., Trauer, J. M., Scott, N., Meehan, M. T., Denholm, J. T., & McBryde, E. S. (2017). Optimally capturing latency dynamics in models of tuberculosis transmission. *Epidemics*, *21*, 39–47. https://doi.org/10.1016/j.epidem.2017.06.002

Revez, J., Espinosa, L., Albiger, B., Leitmeyer, K. C., Struelens, M. J., & ECDC National Microbiology Focal Points and Experts Group, E. N. M. F. P. and E. (2017). Survey on the Use of Whole-Genome Sequencing for Infectious Diseases Surveillance: Rapid Expansion of European National Capacities, 2015-2016. *Frontiers in Public Health*, *5*, 347. https://doi.org/10.3389/fpubh.2017.00347

RILEY, R. L., MILLS, C. C., O'GRADY, F., SULTAN, L. U., WITTSTADT, F., & SHIVPURI, D. N. (1962). Infectiousness of air from a tuberculosis ward. Ultraviolet irradiation of infected air: comparative infectiousness of different patients. *The American Review of Respiratory Disease*, *85*, 511–525. https://doi.org/10.1164/arrd.1962.85.4.511

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J. C., & Müller, M. (2011). pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, *12*(1), 77. https://doi.org/10.1186/1471-2105-12-77

Roetzer, A., Diel, R., Kohl, T. A., Rückert, C., Nübel, U., Blom, 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*(2), e1001387. https://doi.org/10.1371/journal.pmed.1001387

Ruhwald, M., Aabye, M. G., & Ravn, P. (2012). IP-10 release assays in the diagnosis of tuberculosis infection: current status and future directions. *Expert Review of Molecular Diagnostics*, *12*(2), 175–187. https://doi.org/10.1586/erm.11.97

Ruhwald, M., Bjerregaard-Andersen, M., Rabna, P., Kofoed, K., Eugen-Olsen, J., & Ravn, P.

(2007). CXCL10/IP-10 release is induced by incubation of whole blood from tuberculosis patients with ESAT-6, CFP10 and TB7.7. *Microbes and Infection*, *9*(7), 806–812. https://doi.org/10.1016/j.micinf.2007.02.021

Sander, P., Springer, B., Prammananan, T., Sturmfels, A., Kappler, M., Pletschette, M., & Böttger, E. C. (2002). Fitness cost of chromosomal drug resistance-conferring mutations. *Antimicrobial Agents and Chemotherapy*, *46*(5), 1204–1211. https://doi.org/10.1128/aac.46.5.1204-1211.2002

Sassetti, C. M., Boyd, D. H., & Rubin, E. J. (2003). Genes required for mycobacterial growth defined by high density mutagenesis. *Molecular Microbiology*, *48*(1), 77–84. https://doi.org/10.1046/j.1365-2958.2003.03425.x

Schisterman, E. F., Perkins, N. J., Liu, A., & Bondell, H. (2005). Optimal cut-point and its corresponding Youden index to discriminate individuals using pooled blood samples. *Epidemiology*, *16*(1), 73–81. https://doi.org/10.1097/01.ede.0000147512.81966.ba

Schleusener, V., Köser, C. U., Beckert, P., Niemann, S., & Feuerriegel, S. (2017). Mycobacterium tuberculosis resistance prediction and lineage classification from genome sequencing: Comparison of automated analysis tools. *Scientific Reports*, *7*. https://doi.org/10.1038/srep46327

Sekandi, J. N., Zalwango, S., Martinez, L., Handel, A., Kakaire, R., Nkwata, A. K., … Whalen, C. C. (2015). Four Degrees of Separation: Social Contacts and Health Providers Influence the Steps to Final Diagnosis of Active Tuberculosis Patients in Urban Uganda. *BMC Infectious Diseases*, *15*(1), 361. https://doi.org/10.1186/s12879-015-1084-8

Selwyn, P. A., Hartel, D., Lewis, V. A., Schoenbaum, E. E., Vermund, S. H., Klein, R. S., … Friedland, G. H. (1989). A Prospective Study of the Risk of Tuberculosis among

Intravenous Drug Users with Human Immunodeficiency Virus Infection. *New England Journal of Medicine*, *320*(9), 545–550. https://doi.org/10.1056/NEJM198903023200901

Shah, N. S., Auld, S. C., Brust, J. C. M., Mathema, B., Ismail, N., Moodley, P., … Gandhi, N. R. (2017). Transmission of Extensively Drug-Resistant Tuberculosis in South Africa. *The New England Journal of Medicine*, *376*(3), 243–253. https://doi.org/10.1056/NEJMoa1604544

Shah, N. S., Auld, S. C., Brust, J. C. M., Mathema, B., Ismail, N., Moodley, P., … Gandhi, N. R. (2017). Transmission of extensively drug-resistant tuberculosis in South Africa. *New England Journal of Medicine*, *376*(3), 243–253. https://doi.org/10.1056/NEJMoa1604544

Smith, I. (2003). Mycobacterium tuberculosis pathogenesis and molecular determinants of virulence. *Clinical Microbiology Reviews*, *16*(3), 463–496. https://doi.org/10.1128/CMR.16.3.463-496.2003

Snider, D. E., Kelly, G. D., Cauthen, G. M., Thompson, N. J., & Kilburn, J. O. (1985). Infection and disease among contacts of tuberculosis cases with drug-resistant and drug-susceptible bacilli. *American Review of Respiratory Disease*, *132*(1), 125–132. https://doi.org/10.1164/arrd.1985.132.1.125

Sober, E., & Steel, M. (2004). The Contest Between Parsimony and Likelihood. *Systematic Biology*, *53*(4), 644–653. https://doi.org/10.1080/10635150490468657

Sorensen, A. L., Nagai, S., Houen, G., Andersen, P., & Andersen, A. B. (1995). Purification and characterization of a low-molecular-mass T-cell antigen secreted by Mycobacterium tuberculosis. *Infection and Immunity*, *63*(5), 1710–1717. https://doi.org/10.1128/iai.63.5.1710-1717.1995

Ssengooba, W., de Jong, B. C., Joloba, M. L., Cobelens, F. G., & Meehan, C. J. (2016). Whole genome sequencing reveals mycobacterial microevolution among concurrent isolates from

sputum and blood in HIV infected TB patients. *BMC Infectious Diseases*, *16*, 371. https://doi.org/10.1186/s12879-016-1737-2

Ssengooba, W., Meehan, C. J., Lukoye, D., Kasule, G. W., Musisi, K., Joloba, M. L., … de Jong, B. C. (2016). Whole genome sequencing to complement tuberculosis drug resistance surveys in Uganda. *Infection, Genetics and Evolution*, *40*, 8–16. https://doi.org/10.1016/j.meegid.2016.02.019

Steiner, A., Stucki, D., Coscolla, M., Borrell, S., & Gagneux, S. (2014). KvarQ: targeted and direct variant calling from fastq reads of bacterial genomes. *BMC Genomics*, *15*(1), 881. https://doi.org/10.1186/1471-2164-15-881

Steingart, K. R., Henry, M., Laal, S., Hopewell, P. C., Ramsay, A., Menzies, D., … Pai, M. (2007). Commercial Serological Antibody Detection Tests for the Diagnosis of Pulmonary Tuberculosis: A Systematic Review. *PLoS Medicine*, *4*(6), e202. https://doi.org/10.1371/journal.pmed.0040202

Sterling, T. R., Thompson, D., Stanley, R. L., McElroy, P. D., Madison, A., Moore, K., … Bur, S. (2000). A multi-state outbreak of tuberculosis among members of a highly mobile social network: implications for tuberculosis elimination. *The International Journal of Tuberculosis and Lung Disease : The Official Journal of the International Union against Tuberculosis and Lung Disease*, *4*(11), 1066–1073. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11092720

Stimson, J., Gardy, J., Mathema, B., Crudu, V., Cohen, T., & Colijn, C. (2019). Beyond the SNP Threshold: Identifying Outbreak Clusters Using Inferred Transmissions. *Molecular Biology and Evolution*, *36*(3), 587–603. https://doi.org/10.1093/molbev/msy242

Strimbu, K., & Tavel, J. A. (2010). What are biomarkers? *Current Opinion in HIV and AIDS*,

5(6), 463–466. https://doi.org/10.1097/COH.0b013e32833ed177

Strych, U., Davlieva, M., Longtin, J. P., Murphy, E. L., Im, H., Benedik, M. J., & Krause, K. L. (2007). Purification and preliminary crystallization of alanine racemase from Streptococcus pneumoniae. *BMC Microbiology*, *7*. https://doi.org/10.1186/1471-2180-7-40

Stucki, D., Malla, B., Hostettler, S., Huna, T., Feldmann, J., Yeboah-Manu, D., … Gagneux, S. (2012). Two new rapid SNP-typing methods for classifying mycobacterium tuberculosis complex into the main phylogenetic lineages. *PLoS ONE*, *7*(7). https://doi.org/10.1371/journal.pone.0041253

Suliman, S., Thompson, E., Sutherland, J., Weiner 3rd, J., Ota, M. O. C., Shankar, S., … groups, and the G.-74 and A. cohort study. (2018). Four-gene Pan-African Blood Signature Predicts Progression to Tuberculosis. *American Journal of Respiratory and Critical Care Medicine*, rccm.201711-2340OC. https://doi.org/10.1164/rccm.201711-2340OC

Sullivan, K. M., Dean, A., & Minn, M. S. (2009). OpenEpi: A web-based epidemiologic and statistical calculator for public health. *Public Health Reports*. Association of Schools of Public Health. https://doi.org/10.1177/003335490912400320

Sutherland, J. S., Adetifa, I. M., Hill, P. C., Adegbola, R. A., & Ota, M. O. C. (2009). Pattern and diversity of cytokine production differentiates between *Mycobacterium tuberculosis* infection and disease. *European Journal of Immunology*, *39*(3), 723–729. https://doi.org/10.1002/eji.200838693

Sutherland, J. S., de Jong, B. C., Jeffries, D. J., Adetifa, I. M., & Ota, M. O. C. (2010). Production of TNF-α, IL-12(p40) and IL-17 can discriminate between active TB disease and latent infection in a West African cohort. *PLoS ONE*, *5*(8), e12365. https://doi.org/10.1371/journal.pone.0012365

Tang, P., & Johnston, J. (2017). Treatment of Latent Tuberculosis Infection. *Current Treatment Options in Infectious Diseases*, *9*(4), 371–379. https://doi.org/10.1007/s40506-017-0135-7

Targeted Tuberculin Testing and Treatment of Latent Tuberculosis Infection. (2000). *American Journal of Respiratory and Critical Care Medicine*, *161*(supplement_3), S221–S247. https://doi.org/10.1164/ajrccm.161.supplement_3.ats600

Templeton, A. R. (1998). Human Races: A Genetic and Evolutionary Perspective. *American Anthropologist*, *100*(3), 632–650. https://doi.org/10.1525/aa.1998.100.3.632

Teunis, P., Heijne, J. C. M., Sukhrie, F., van Eijkeren, J., Koopmans, M., & Kretzschmar, M. (2013a). Infectious disease transmission as a forensic problem: who infected whom? *Journal of the Royal Society, Interface*, *10*(81), 20120955. https://doi.org/10.1098/rsif.2012.0955

Teunis, P., Heijne, J. C. M., Sukhrie, F., van Eijkeren, J., Koopmans, M., & Kretzschmar, M. (2013b). Infectious disease transmission as a forensic problem: who infected whom? *Journal of the Royal Society, Interface*, *10*(81), 20120955. https://doi.org/10.1098/rsif.2012.0955

van Geuns, H. A., Meijer, J., & Styblo, K. (1975). The yield from mass tuberculin testing of unvaccinated children and adolescents. *Bulletin of the International Union against Tuberculosis*, *50*(1), 82–89. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/1218290

Van Soolingen, D. (2001). Molecular epidemiology of tuberculosis and other mycobacterial infections: Main methodologies and achievements. *Journal of Internal Medicine*. https://doi.org/10.1046/j.1365-2796.2001.00772.x

van Zyl-Smit, R. N., Zwerling, A., Dheda, K., & Pai, M. (2009). Within-Subject Variability of Interferon-g Assay Results for Tuberculosis and Boosting Effect of Tuberculin Skin

Testing: A Systematic Review. *PLoS ONE*, *4*(12), e8517. https://doi.org/10.1371/journal.pone.0008517

Vernikos, G., Medini, D., Riley, D. R., & Tettelin, H. (2015, February 1). Ten years of pan-genome analyses. *Current Opinion in Microbiology*. Elsevier Ltd. https://doi.org/10.1016/j.mib.2014.11.016

Verver, S., Warren, R. M., Munch, Z., Richardson, M., Van Der Spuy, G. D., Borgdorff, M. W., … Van Helden, P. D. (2004). Proportion of tuberculosis transmission that takes place in households in a high-incidence area. *Lancet*, *363*(9404), 212–214. https://doi.org/10.1016/S0140-6736(03)15332-9

Vynnycky, E., & Fine, P. E. M. (2000). Lifetime risks, incubation period, and serial interval of tuberculosis. *American Journal of Epidemiology*, *152*(3), 247–263. https://doi.org/10.1093/aje/152.3.247

Wachtman, L. M., Miller, A. D., Xia, D., Curran, E. H., & Mansfield, K. G. (2011). Colonization with nontuberculous mycobacteria is associated with positive tuberculin skin test reactions in the common marmoset (Callithrix jacchus). *Comparative Medicine*, *61*(3), 278–284. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/21819699

Walker, T. M., Ip, C. L. C., Harrell, R. H., Evans, J. T., Kapatai, G., Dedicoat, M. J., … Peto, T. E. A. (2013a). Whole-genome sequencing to delineate Mycobacterium tuberculosis outbreaks: A retrospective observational study. *The Lancet Infectious Diseases*, *13*(2), 137–146. https://doi.org/10.1016/S1473-3099(12)70277-3

Walker, T. M., Ip, C. L., Harrell, R. H., Evans, J. T., Kapatai, G., Dedicoat, M. J., … Peto, T. E. (2013b). Whole-genome sequencing to delineate Mycobacterium tuberculosis outbreaks: a retrospective observational study. *The Lancet Infectious Diseases*, *13*(2), 137–146.

https://doi.org/10.1016/S1473-3099(12)70277-3

Walker, T. M., Kohl, T. A., Omar, S. V., Hedge, J., Del Ojo Elias, C., Bradley, P., … Munang, M. (2015). Whole-genome sequencing for prediction of Mycobacterium tuberculosis drug susceptibility and resistance: A retrospective cohort study. *The Lancet Infectious Diseases*, *15*(10), 1193–1202. https://doi.org/10.1016/S1473-3099(15)00062-6

Walker, T. M., Kohl, T. A., Omar, S. V, Hedge, J., Del Ojo Elias, C., Bradley, P., … Modernizing Medical Microbiology (MMM) Informatics Group. (2015). Whole-genome sequencing for prediction of Mycobacterium tuberculosis drug susceptibility and resistance: a retrospective cohort study. *The Lancet. Infectious Diseases*, *15*(10), 1193–1202. https://doi.org/10.1016/S1473-3099(15)00062-6

Walker, T. M., Lalor, M. K., Broda, A., Ortega, L. S., Morgan, M., Parker, L., … Conlon, C. P. (2014). Assessment of Mycobacterium tuberculosis transmission in Oxfordshire, UK, 2007-12, with whole pathogen genome sequences: An observational study. *The Lancet Respiratory Medicine*, *2*(4), 285–292. https://doi.org/10.1016/S2213-2600(14)70027-X

Wallis, R. S., Pai, M., Menzies, D., Doherty, T. M., Walzl, G., Perkins, M. D., & Zumla, A. (2010, May 29). Biomarkers and diagnostics for tuberculosis: progress, needs, and translation into practice. *The Lancet*. https://doi.org/10.1016/S0140-6736(10)60359-5

Wampande, E. M., Mupere, E., Jaganath, D., Nsereko, M., Mayanja, H. K., Eisenach, K., … Joloba, M. L. (2015). Distribution and transmission of Mycobacterium tuberculosis complex lineages among children in peri-urban Kampala, Uganda. *BMC Pediatrics*, *15*(1). https://doi.org/10.1186/s12887-015-0455-z

Watkins, R. E., Brennan, R., & Plant, A. J. (2000). Tuberculin reactivity and the risk of tuberculosis: a review. *The International Journal of Tuberculosis and Lung Disease : The*

Official Journal of the International Union against Tuberculosis and Lung Disease, *4*(10), 895–903. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11055755

Wayengera, M., Kateete, D. P., Asiimwe, B., & Joloba, M. L. (2018). Mycobacterium tuberculosis thymidylate kinase antigen assays for designating incipient, high-risk latent M. tb infection. *BMC Infectious Diseases 2018 18:1*, *18*(1), 133. https://doi.org/10.1186/s12879-018-3007-y

Wayengera, M., Mwebaza, I., Welishe, J., Bayiyana, A., Kateete, D. P., Wampande, E., … Joloba, M. L. (2017). Immuno-diagnosis of Mycobacterium tuberculosis in sputum, and reduction of timelines for its positive cultures to within 3 h by pathogen-specific thymidylate kinase expression assays. *BMC Research Notes*, *10*(1), 368. https://doi.org/10.1186/s13104-017-2649-y

Weis, S. E., Pogoda, J. M., Yang, Z., Cave, M. D., Wallace, C., Kelley, M., & Barnes, P. F. (2002). Transmission Dynamics of Tuberculosis in Tarrant County, Texas. *American Journal of Respiratory and Critical Care Medicine*, *166*(1), 36–42. https://doi.org/10.1164/rccm.2109089

Whalen, C. C. (2005). Diagnosis of Latent Tuberculosis Infection. *JAMA*, *293*(22), 2785. https://doi.org/10.1001/jama.293.22.2785

Whalen, C. C. (2016). The Replacement Principle of Tuberculosis. Why Prevention Matters. *American Journal of Respiratory and Critical Care Medicine*, *194*(4), 400–401. https://doi.org/10.1164/rccm.201603-0439ED

Whalen, C. C., Johnson, J. L., Okwera, A., Hom, D. L., Huebner, R., Mugyenyi, P., … Pekovic, V. (1997). A Trial of Three Regimens to Prevent Tuberculosis in Ugandan Adults Infected with the Human Immunodeficiency Virus. *New England Journal of Medicine*, *337*(12),

801–808. https://doi.org/10.1056/NEJM199709183371201

Whalen, C. C., Zalwango, S., Chiunda, A., Malone, L., Eisenach, K., Joloba, M., … Mugerwa, R. (2011). Secondary Attack Rate of Tuberculosis in Urban Households in Kampala, Uganda. *PLoS ONE*, *6*(2), e16137. https://doi.org/10.1371/journal.pone.0016137

WHO. (2019). WHO | Global tuberculosis report 2019. *WHO*.

WHO | Global tuberculosis report 2017. (2017). *WHO*. Retrieved from http://www.who.int/tb/publications/global_report/en/

Wilkinson, D. (2000). Drugs for preventing tuberculosis in HIV infected persons. In *Cochrane Database of Systematic Reviews*. Chichester, UK: John Wiley & Sons, Ltd. https://doi.org/10.1002/14651858.CD000171

Woese, C. R. (2000). Interpreting the universal phylogenetic tree. *Proceedings of the National Academy of Sciences of the United States of America*, *97*(15), 8392–8396. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10900003

Won, E.-J., Choi, J.-H., Cho, Y.-N., Jin, H.-M., Kee, H. J., Park, Y.-W., … Kee, S.-J. (2017). Biomarkers for discrimination between latent tuberculosis infection and active tuberculosis disease. *Journal of Infection*, *74*(3), 281–293. https://doi.org/10.1016/J.JINF.2016.11.010

World Health Organization. (n.d.). *Antimicrobial resistance : global report on surveillance*.

World Health Organization. (2015a). Implementing the End TB Strategy: The Essentials. *Who*, 1–130. https://doi.org/10.1017/CBO9781107415324.004

World Health Organization. (2015b). WHO End TB Strategy.

Yang, T., Zhong, J., Zhang, J., Li, C., Yu, X., Xiao, J., … Chen, F. (2018). Pan-genomic study of Mycobacterium tuberculosis reflecting the primary/secondary genes, generality/individuality, and the interconversion through copy number variations. *Frontiers*

in Microbiology, *9*(AUG). https://doi.org/10.3389/fmicb.2018.01886

Yates, T. A., Tanser, F., & Abubakar, I. (2016, January 1). Plan Beta for tuberculosis: It's time to think seriously about poorly ventilated congregate settings. *International Journal of Tuberculosis and Lung Disease*. International Union against Tubercul. and Lung Dis. https://doi.org/10.5588/ijtld.15.0494

Yates, Tom A, Khan, P. Y., Knight, G. M., Taylor, J. G., McHugh, T. D., Lipman, M., … Abubakar, I. (2016). The transmission of Mycobacterium tuberculosis in high burden settings. *The Lancet Infectious Diseases*, *16*(2), 227–238. https://doi.org/10.1016/S1473-3099(15)00499-5

Ypma, R. J. F., Altes, H. K., Van Soolingen, D., Wallinga, J., & Van Ballegooijen, W. M. (2013). A sign of superspreading in tuberculosis: Highly skewed distribution of genotypic cluster sizes. *Epidemiology*, *24*(3), 395–400. https://doi.org/10.1097/EDE.0b013e3182878e19

Yuen, C. M., Kammerer, J. S., Marks, K., Navin, T. R., & France, A. M. (2016). Recent Transmission of Tuberculosis — United States, 2011–2014. *PLOS ONE*, *11*(4), e0153728. https://doi.org/10.1371/journal.pone.0153728

Zappala, Z., Fresard, L., Waggott, D., Hou, Y., Smith, K. S., Montgomery, S. B., … Euan, A. (2018). Long-read WGS identifies causal SV in a Mendelian disease. *Genet Med*, *20*(1), 159–163. https://doi.org/10.1038/gim.2017.86.Long-read