CHARACTERIZATION AND PREVENTION OF TUBERCULOSIS TRANSMISSION USING SOCIAL NETWORKS OF TB PATIENTS, PATHOGEN WHOLE GENOME SEQUENCING AND MATHEMATICAL MODELING

by

RONALD GALIWANGO

(Under the Direction of Christopher Whalen)

ABSTRACT

In Aim 1 of the study, I found heterogeneity in processing of WGS data among studies and some areas of consensus especially in recent literature. SNP thresholds are the most widely used method for inferring transmission with thresholds of 12 and 5 SNPs the most widely used. Bayesian transmission modeling attempts to address their limitation and is increasingly being used in transmission studies.

In aim 2, I investigated the role of the social network of a TB case in transmission of tuberculosis using a large social network study, the Community Health and Social Networks of TB (COHSONET) study. I also determined the relationship between genetic distance and social network distance. I found that 43% of the index case pairs who had genetically linked strains of *Mycobacterium tuberculosis* had an identifiable path between them in the social network, but only 13% of these index pairs were found to have a close social distance of one step in the social network. There was no correlation between genetic distance and social network distance.

In aim 3, I investigated genetic linkage among TB patients in the COHSONET study using a threshold of 12 SNPs to identify clusters of recent transmission, and covariates associated with clustering. I found that twenty-nine (36.7%) patients of the 79 sequenced isolates formed 12 clusters. A multivariate logistic analysis showed that clustered cases were more likely to be current or past smokers.

Unlike deterministic compartmental models, network models account for heterogeneity in mixing patterns. I implemented an individual-based version (particularly a network model) of a deterministic model with two latency compartments on a dynamic network simulated from a static network (Aim 4). The model depicted expected dynamics in a viability analysis when compared with a deterministic version. The model will be used to answer research questions such as whether infections in the household are sufficient to maintain the epidemic in the community, and if not so, different scenarios explaining the observed infections in the community will be simulated.

INDEX WORDS: genome; sequencing; tuberculosis; transmission; network; model

CHARACTERIZATION AND PREVENTION OF TUBERCULOSIS TRANSMISSION USING SOCIAL NETWORKS OF INDEX TB PATIENTS, PATHOGEN WHOLE GENOME SEQUENCING AND MATHEMATICAL MODELING

by

RONALD GALIWANGO

BSc., Makerere University, Uganda, 2013

MPhil., University of Cambridge, UK, 2015

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

2016

CHARACTERIZATION AND PREVENTION OF TUBERCULOSIS TRANSMISSION USING SOCIAL NETWORKS OF TB PATIENTS, PATHOGEN WHOLE GENOME SEQUENCING AND MATHEMATICAL MODELING

by

RONALD GALIWANGO

Major Professor: Christopher Whalen

Committee: Andreas Handel
Juliet Sekandi
Liang Liu

Electronic Version Approved:

Ron Walcott
Interim Dean of the Graduate School
The University of Georgia
December 2019

DEDICATION

I dedicate this work to my parents, Mr. Joseph Katabalwa and Ms. Annet Nabatanzi whose sacrifices have brought me this far.

ACKNOWLEDGEMENTS

I wish to acknowledge Dr. John Kitayimbwa for his role in my academic and personal growth. I would also like to acknowledge MUII (Makerere University / Uganda Virus Research Institute Centre of Excellence in Infection and Immunity Research and Training programme) and its director, Prof. Alison Elliot for their role in my career development.

Special acknowledgements to the EIA (Epidemiology in Action) group, the Infectious Disease seminar (Handel group) and the Global Health family at the Global Health Institute of the University of Georgia for the encouragement and support (both academic and non-academic).

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1 : INTRODUCTION AND LITERATURE REVIEW

WHAT IS TUBERCULOSIS?

Tuberculosis (TB) is an airborne infectious disease caused by the bacillus *Mycobacterium tuberculosis* (*Mtb*). Tuberculosis typically affects the lungs (pulmonary TB) but can affect other sites (extrapulmonary TB) such as lymph nodes, bones and the central nervous system.

TB is spread when a person with infectious (pulmonary) TB expels bacteria (infectious particles) into the air (for example by coughing or sneezing) which when they survive in the air are inhaled by a susceptible individual who may become infected and has the potential to develop TB [1].

Known signs of pulmonary TB include: coughing for greater than 2 weeks (often producing sputum which may be bloody), fever, night sweats, weight loss, chest pain. Symptoms of extrapulmonary TB vary by site.

EPIDEMIOLOGY OF TB

An estimated 10 million people suffered from TB in 2018, 5.7 million of which were men, 3.2 million were women and 1.1 million children [2]. Nine-percent (9%) were people living with HIV (72% of which were in Africa). The majority of the cases (68%) were in South-East Asia (44%) and Africa (24%) where the epidemic is predominantly driven by transmission (rather than reactivation of latent infection) and high rates of HIV.

TB is the ninth leading cause of death worldwide and has maintained its position, over the past 5 years, as the leading cause from a single infectious agent, ranking above HIV/AIDS

and malaria [2]. TB also continues to be the leading cause of death among people living with HIV, accounting for nearly one in three HIV-related deaths. An estimated 1.5 million people died from TB in 2018 [2].

Humans are the sole reservoir for *Mycobacterium tuberculosis (Mtb),* the causative agent of TB. Nevertheless, animal hosts especially cattle [3–5] have been suggested though it is questionable how important they are. Therefore, person-to-person transmission is the known sole mechanism for propagating the global TB epidemic.

Close contacts of infectious TB cases are susceptible to becoming infected, and if infected, to progressing to active TB disease. With latent infection, individuals experience no adverse health effects (no symptoms and don't feel sick) and will not transmit *Mtb*, but they face an ongoing risk of developing active tuberculosis through reactivation. Overall, about 5 to 15% of infected persons who do not receive treatment for latent TB infection will develop TB disease at some time in their lives, with 5% developing active disease within 2 years of infection. However, the probability of developing TB disease is higher in children (<5 years), among people infected with HIV, in silica-exposed miners particularly those with silicosis and in people affected by risk factors such as under-nutrition, diabetes, smoking and alcohol consumption [1,6].

STATEMENT OF THE PROBLEM

In 2014, the WHO set an ambitious target to end TB by 2035 [7] which has at its core the early detection and treatment of existing cases. While diagnosis and treatment of index cases are essential for the proper management of the individual case, they may not be sufficient to control the epidemic. Like most infectious diseases, tuberculosis creates the next generation of new cases through transmission before the diagnosis is made and treatment begun in the index case.

This transmission may sustain the epidemic in the community by replacing one case with another over time [8]. Therefore, efforts to end TB will depend on our ability to halt ongoing transmission.

As long as there are unrecognized, infectious cases circulating in the community, so does the risk of infection and disease to vulnerable populations such as children and HIV seropositive persons. Control of the epidemic in the population confers protection at the individual level to these vulnerable populations who benefit from having less levels of TB circulating.

GAPS IN KNOWLEDGE

Variability in data processing and transmission inference methodology

Whole Genome Sequencing (WGS) has improved our ability to characterize transmission events by providing better resolution compared to genotyping techniques for example MIRU-VNTR (Mycobacterial Interspersed Repetitive Units - Variable Number of Tandem Repeats), Spoligotyping and RFPL (Restriction Fragment Length Polymorphism) that only use less than 0.1% of the bacterial genome. With recent improvements in Next Generation Sequencing (NGS) technologies as well as the reduction in cost and turnaround time of sequencing workflows, WGS has replaced traditional molecular typing as routine in *Mycobacterium tuberculosis*.

However, there are many computational pipelines that are used in TB transmission studies to process WGS data with each pipeline containing a series of data processing steps. The way WGS data is processed varies from one study to another and has implications in the identification of transmission events. In addition, the methods used to identify transmission events are not homogeneous among studies. Even with the SNP (Single Nucleotide Polymorphism) threshold, there are various thresholds used. This variation in WGS data processing and transmission inference methodology leads to limited comparability among transmission studies of tuberculosis. There is a need to review the individual data processing

steps, available full computational pipelines and methods used to infer transmission of *Mycobacterium tuberculosis*.

Limited understanding of local dynamics and drivers of transmission

While home-based contact investigations and infection control programs in hospitals and clinics have a successful track record as TB control activities, there is a gap in our knowledge of where, and between whom, community-based transmission of TB occurs. Household contact studies have previously highlighted the household as an important setting for transmission of *Mycobacterium tuberculosis* [9] but recent evidence suggests that household transmission accounts for a smaller percentage of the total number of TB cases (Martinez et al., 2017) indicating that majority of the cases occur outside the household (i.e., in the community).

The fact that a small proportion of TB is attributed to being a household of a TB case suggests that there are other unrecognized routes, beyond the household, via which TB is transmitted that could be sustaining the epidemic in the community. One such route could be transmission via extra-household contacts who are within the social network of a TB index case (Figure 1.2). This network may contain their workmates, same church goers, peers, persons with whom they spend a significant proportion of their time. This study will explore the role of this potential non-geographical hotspot in the transmission of TB.

Deterministic compartmental models have been used for studying the transmission of *Mycobacterium tuberculosis* [11,12]. However, these models are so simplistic in that they assume random (or homogeneous) mixing of individuals in the population meaning that all susceptible persons have equal probabilities of getting infected which is not always true. In practice, each infectious individual has a finite set of contacts to whom they can pass infection.

Individual-based models such as network models allow us to account for heterogeneity in mixing of individuals in the population. Network models have been used for the study of transmission dynamics of other infectious diseases such as HIV [13–16] but not so much for tuberculosis yet like for HIV, network structure plays a critical role in the transmission of *Mycobacterium tuberculosis*.

HOW THIS PROJECT TRIES TO FILL THE IDENTIFIED GAPS

In the first aim of this project, I performed a systematic review of individual data processing steps, available full computational pipelines and the methods used in published studies to infer (confirm or refute) direct transmission of *Mycobacterium tuberculosis* using Whole Genome Sequences from pathogen isolates. I describe the rationale behind each data processing step and discuss the strengths and limitations of each approach used for making transmission inferences.

In the second aim, I explore the role of the social network of index tuberculosis cases in the transmission of *Mycobacterium tuberculosis* by determining the proportion of putative direct transmission events that occur among index tuberculosis cases with an identifiable path in the social network. I also determine the relationship between social network distance and genetic distance.

In the third aim, I identify covariates associated with genetic clustering of index tuberculosis cases in an endemic setting in Kampala-Uganda, including social network characteristics such as degree, betweenness and centrality. The variables associated with clustering of tuberculosis patients could be maintaining the epidemic in this setting.

In the fourth aim, I developed a stochastic network model to be used to study *Mycobacterium tuberculosis* transmission. I implemented an individual-based version of a deterministic model with two latency compartments [11,12], particularly a network model.

LITERATURE REVIEW

Why systematic reviews?

Given the ever-increasing output of scientific publications, scientists can't be expected to examine in detail every single new paper relevant to their interests [17]. Timely systematic reviews try to fill this gap by providing a snapshot of the topic of interest through critical appraisal of the research studies that satisfy pre-specified eligibility criteria and a mainly qualitative synthesis of the results. They are different from meta-analyses where statistical methods are used to summarize the results of these studies.

Systematic reviews are a good starting point for researchers intending to learn about a new research topic of interest but they also give regular updates to existing researchers in the field since they give insights into the current state of the field. Due to their summarized format, systematic reviews are often widely read compared with primary research. Because data is collected using a systematic methodology, the likelihood of reproducing results is quite high.

The number of TB studies that use WGS to study transmission have increased in recent times. There is some heterogeneity in the individual data processing steps and computational pipelines used in processing pathogen WGS data. Variations in these data processing steps affect the inferences made with regards to transmission.

Previous reviews on the use of WGS in TB transmission studies

Most previous reviews compare WGS with traditional genotyping with focus on transmission inference and less on WGS data processing [18–21].

Vlad and colleagues studied the sensitivity and specificity of WGS for detection of recent transmission using conventional epidemiology as the gold standard [20,21].

The review by van der Werf and Ködmön [19] focused on use of WGS to investigate international tuberculosis outbreaks.

The review by Hatherell and colleagues [18] looked at methods used to infer transmission but included only 12 research articles that were published until 14th July 2015. More studies using WGS to study *Mycobacterium tuberculosis* transmission, employing newer methods for transmission inference and incorporating best practices for WGS data processing, have been published since then.

A typical pipeline for WGS data processing

A typical computational pipeline for processing WGS data for purposes of transmission inference begins with the raw reads, resulting from sequencing isolated DNA of a pathogen of interest, for our case *Mycobacterium tuberculosis*. The reads can either be single-end (sequenced from one end of the DNA fragment to another) or paired-end in which each end of the same DNA fragment is sequenced i.e., one sequence (e.g., the forward read) runs from one end to another and the other (the reverse) runs in the opposite direction. This helps in resolving ambiguous bases thereby improving the quality of the alignment.

Choice of WGS platform depends on length of reads (longer reads desirable), cost (low cost desirable) and sequence quality (low per base error rate desirable). Illumina, San Diego, CA, USA is the most widely used mainly due to the low per base error rate [22], though the reads are of a shorter length (a maximum of 150bp for the HiSeq and a maximum of 300bp for the MiSeq) hence cannot resolve repetitive elements. Proline-Proline-Glutamate (PPE)/ Proline-Glutamine (PE) gene families of *Mycobacterium tuberculosis* are very repetitive, so cause trouble when

sequencing on Illumina. Repetitive regions are collapsed into one hence won't be detected. For this reason, SNPs in PPE/PE genes are often removed while making inferences on *Mycobacterium tuberculosis* transmission. The law of repeats states that 'It is impossible to resolve repeats of length L unless you have reads longer than L'.

There are five main WGS technologies: *Illumina/Solexa* (Sequencing by synthesis; 100-300bp), *Ion Torrent* (Thermo Fisher Scientific; Life Technologies; Applied Biosystems Inc. (ABI: SOLiD ligation sequencing system); Ion semiconductor and sequencing by ligation; 100-400bp), *Pacific Biosciences* (Single molecule via dye labels; PacBio: longer reads: 5,000-25,000+bp: used mainly for creating reference sequences), *Nanopore* (Oxford Nanopore; Electronic nanopore sensing; 5,000 - 1,000,000+ bp: longer read length but high per read error rate) and *Roche* (454 Life Sciences; Pyrosequencing, single-molecule nanopore; 100-150bp in 2015, now up to 700+bp since launch of 454 GS FLX Titanium system in 2008) [23]. Illumina/ Solexa sequencing and ABI/solid (Applied Biosystems) support both single end and paired-end sequencing [24].

PacBio and Oxford Nanopore sequencing platforms produce longer reads (>20,000bp). Longer reads are desirable because they are better for detecting features such as repetitive elements. Illumina compensates for its short-read lengths through supporting paired-end sequencing, in which each end of the same DNA molecule is sequenced [23]. This greatly improves the quality of the alignment compared with single reads alone. In addition, Illumina's low per base error rate has made it the most widely used platform [22]. Illumina/ Solexa sequencing and ABI/solid (Applied Biosystems) support both single end and paired-end sequencing [24].

Before starting the WGS analysis, an initial quality check is performed on raw reads to ensure that it's satisfactory and decisions are made on how to improve downstream analysis by

performing a series of additional preprocessing steps. One example tool for performing initial quality checks on raw reads is FASTQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). The tool produces several statistics characterizing the raw data quality: per base sequence quality, per sequence quality scores, per base sequence content, per sequence GC content, Sequence length distribution, adapter content, Kmer content and sequence duplication level. Preprocessing involves trimming raw reads and quality filtering.

Trimming involves Identifying and removing low quality sequences or parts of sequences such as adapter fragments, nucleotide bases having less than minimum threshold quality score and known contaminants (e.g., with Kraken) from raw reads. With the benefit of longer reads in mind, trimmed reads below a minimum threshold length are discarded. Trimming has been shown to increase the quality and reliability of the analysis by reducing the false positive call rate for bases during reference-based assembly [25].

In addition, trimming reduces the amount of computational resources (RAM, disk space and execution time) needed during subsequent data processing and downstream analysis. Different trimmers produce different results and are highly dependent on the parameters used. One such parameter is the minimum quality threshold, Q. A high value of Q leads to as small size surviving dataset while a lower value of Q retains a lot of low quality regions and unnecessarily increasing computational requirements [25]. Preprocessing ends with filtering the reads by quality score and those with a pre-set percentage of bases below the minimal threshold quality score (MinimalQ) are discarded.

Now that the quality of reads is satisfactory, the next step in the pipeline is to map the quality reads to a reference genome of choice. A study by Lee and Behr showed that the choice

of reference genome, within the *Mycobacterium tuberculosis* complex, has negligible influence transmission inferences made [26]. For each sample, you map each read to the reference genome using a mapping algorithm/software of choice.

Prior to variant discovery, post-mapping Quality Control (QC) is performed to assess the quality of the mapping i.e., having completed the alignment, the first thing we want to know is how well did our reads align to the reference. This is important because some issues such as low coverage only appear after alignment. Identifying and fixing mapping issues makes downstream processing easier and more accurate. An example metric for post mapping QC is the sequencing coverage depth which is the number of reads that cover a given genome base or the average number of times a given region (e.g., a base) has been sequenced or covered by independent reads. Sequencing coverage depth determines with what confidence variant calling is done. The deeper the coverage, the more reads are mapped on each base and the higher the reliability and the accuracy of base calling.

Samples with average genomic coverage (sequencing depth) and a minimum threshold percentage of reads mapped correctly (uniquely mapped reads) less than minimum threshold values are flagged for further assessment. Poorly mapped reads include unmapped reads (reads that failed to map), duplicated mapped reads and multi-mapped reads. Duplicated mapped reads are those that accumulate at the same start position in the reference genome. They may arise due to errors in the sample or library preparation leading to multiple reads from the exact same input DNA template. Although read duplicates could represent true DNA materials, it's impossible to distinguish them from PCR artifacts which are results of uneven amplification of DNA fragments. Therefore, to reduce their harmful effect of multiplying any sequencing errors leading to artifacts in downstream analysis, duplicated mapped reads are identified and removed. Multi-

mappers (reads that don't map to a single unique position in the reference genome, also called repeats) are also removed. Reads can be filtered by mapping quality. Only reads aligned with a quality score higher than a given threshold are retained.

Now that we are confident with our assembly, the next step is variant discovery. For purposes of transmission analysis, the only variants detected are Single Nucleotide Polymorphisms (SNPs. A SNP is a change of a single nucleotide at a specific position in the genome, where each variation is present to some appreciable degree within a population (for example greater than 1% frequency). A Single Nucleotide Variant (SNV) is a variation in a single nucleotide without any limitations on frequency. Variants (SNPs) are called/detected using a given choice of SNP caller and are then annotated say for gene function (e.g., drug resistant SNPs). The number and quality of variants varies by variant caller with each variant caller detecting SNPs with a different level of sensitivity and specificity [27]. In their study, Hwang and colleagues showed that SAMtools variant caller combined with the BWA-MEM aligner had the best performance, but Freebayes with any aligner showed an equally high performance for the SNP calls [27].

The 'raw' variants are filtered so as to remain with only quality variants i.e., low quality variants are discarded. This involves steps such as removing variants in repetitive regions of the genome (due to the difficulty in sequencing such regions), base calls below the defined minimum read coverage/depth, high-density SNPs (highly clustered SNPs), SNPs below minimum allele frequency, base calls below minimum base quality and mapping quality, SNPs in drug resistant regions (figure 4). Regions of high SNP density are indicative of recombination. SNPs in resistance-mediating (drug-resistance associated) genes are removed to rule out selection pressure (mutation) due to drug resistance.

Methods used to infer transmission

The most used approach for inferring TB transmission is the use of Single Nucleotide Polymorphism (SNP) thresholds [18] because of their simplicity. However, there is always a question on which threshold to use for confirming (or refuting) transmission. Another limitation of this approach is that it has a higher dependence on the fraction of sequenced isolates. Hence transmission may be under reported in case some of the outbreak patients, for example are not sampled. Seemingly unrelated TB patients could have transmission links with un-sampled TB patients.

Alternative approaches have been suggested such as transmission modeling (e.g., R packages: Transphylo and Outbreaker, use of transmission kernels etc.). Transphylo [28] is a Bayesian transmission modeling approach that uses a time-labelled phylogeny such as the ones output by BEAST [29,30] to infer a transmission tree. It has an added advantage of inferring undetected cases and incorporating within host diversity of the pathogen.

Outbreaker [31] can infer the reproductive number of the pathogen which can tell us about how effective the infection is transmitted. Unlike Transphylo, Outbreaker doesn't consider within host diversity. In Outbreaker, potential transmission events are inferred from clustering events on the phylogeny such as clades/lineages. From the transmission tree, one can estimate the number of secondary infections generated by each case and thus of the transmission intensity (characterized by the reproduction number, R) over time [31].

Using transmission modeling, TB patients can be connected through transmission tree inference by combining epidemiologic (sample collection dates, start period of coughing), social network data (as a proxy for contact proximity) and genomic data (pathogen genome sequences).

A list of other open source tools for inferring TB transmission with WGS is found at https://github.com/molecular-epidemiology/molepi-tools. Other methods will be arrived at from the systematic review.

Where is transmission occurring?

There is a wealth of evidence to support transmission of tuberculosis in households and the household has been highlighted as an important setting for TB transmission (Martinez et al., 2017; Morrison et al., 2008). A systematic review and meta-analysis performed by Martinez and colleagues showed that exposed children in households of an index TB case are 3.79 (95% confidence interval (CI): 3.01, 4.78) times more likely to be infected than their community counterparts (Martinez et al., 2017).

Despite the high risk of TB infection among household contacts of a TB case compared to their community counterparts, there is a small proportion, only 14%, of transmission is attributable to household exposure (Martinez et al., 2017). In settings with a high tuberculosis burden, tuberculosis transmission is therefore more likely to occur outside the household such as in healthcare and congregate settings for example schools, public transportation settings, workplaces, mines, shelters and prisons [1].

Social networks and transmission of *Mycobacterium tuberculosis*

The relevance of social network structure to transmission dynamics of disease has been well studied for HIV and other sexually transmitted infections [33–36] particularly in the study of sexual networks and behavior among these networks but not so much for TB, yet like for HIV, social network structure including contact and mixing patterns of the population play a critical role in the transmission of TB. Network characteristics, such as size, composition, and density have been found to be associated with HIV risk behaviors that include sharing injection

equipment, drug use cessation, having multiple concurrent sexual partnerships, unprotected sex, and exchanging sex for money or drugs. Social network approaches have thus been developed for HIV prevention interventions to reduce risk behaviors [36].

Most social networks used in the study of infectious diseases are, first-order egocentric social networks. This means, for example for TB, an index case is identified who is asked to list their close contacts (first level contacts). The first order egocentric structure can be extended to include second level contacts i.e., the contacts of contacts.

Most network studies are cross-sectional and social networks are normally constructed by means of interviews with patients. Social network questionnaires are given to patients which they fill out with guidance from the interviewers who are part of the study team [37]. Questions asked include identifying information (such as age/date of birth, sex, ethnicity); questions on medical history (e.g., HIV status, previous TB episodes); symptom onset (e.g., start of cough, previous contact with a TB case); questions on risk factors such as smoking, drug and alcohol use; questions on place of residence; travel history; places of social aggregation; social contacts (including closest household and non-household members); and time spent with each of the listed social contacts [38].

Although the associations between certain social determinants and the occurrence of tuberculosis have been explored, the relationships among individuals have been less studied and Social Network Analysis (SNA) methods have been less used. More so, the role of the social network of a TB case in the transmission of *Mycobacterium tuberculosis* has not been explored. For example, the relationship between genetic linkage and genetic distance with social network distance has not been studied. SNA has been used retrospectively to characterize *Mtb* outbreaks,

identify risk factors for transmission, locate places of recent transmission and highlight the importance of places of social aggregation in sustaining transmission [38–43]

SNA (together with WGS) was used to study an outbreak of TB in British Columbia, Canada [38]. The social network was constructed by means of interviews with patients to determine the origins and transmission dynamics of the outbreak. The methodology was used to study how TB was transmitted making it possible to identify the individuals and characteristics that facilitated transmission. Traditional contact tracing didn't identify the source of the outbreak. SNA identified an adult with cavitary, smear positive pulmonary TB that had been asymptomatic and un-treated for at least 8 months before detection of the first case, as the source of the outbreak. SNA identified increased crack cocaine use among a high-risk social network as a socioeconomic factor that may have triggered the simultaneous expansion of the two lineages from a common ancestor that had been detected in the community before the outbreak. Use of a social network questionnaire improved contact tracing and subsequent active case finding efforts by revealing previously unreported social interactions and identifying several locations frequented by infectious patients, including two hotels, a meal center, two community centers, and a series of crack houses. Use of the social network questionnaire also identified demographic characteristics associated with an increased risk of TB transmission. Transient living arrangements, crack cocaine use and alcohol use were associated with an increased risk of TB transmission.

Traditional epidemiological methods in combination with SNA and WGS were used to investigate the transmission of TB in an educational institution following an outbreak in the South West of England [43]. SNA identified shared exposures (with a suspected/active disease case) associated with an increased odds/risk of developing active disease and Latent TB

Infection (LTBI). The community including the suspected index case was at significantly elevated risk of active disease (odds ratio 7.5, 95% CI=1.3 to 44.0).

Drivers of *Mycobacterium tuberculosis* transmission

The drivers of TB transmission differ by setting. This is because, countries (or regions) differ in the burden of prevalent tuberculosis, HIV burden, capacity of healthcare and public health systems to identify and effectively treat individuals with infectious forms of tuberculosis, and the ways in which individuals live, work, and interact i.e., social mixing patterns [1,44].

Before the emergency of WGS, traditional genotyping has been used to identify factors associated with recent transmission, using clustering of isolates based on their genotypic profiles as a measure of recent transmission [45]. In this approach, individuals with identical or similar fingerprint patterns are considered to be clustered. Patients whose isolates cluster together are considered to be part of the same recent transmission chains while those with unique (un-clustered) isolates are more likely to be cases of reactivated TB disease that was acquired in the past [45]. The covariates associated with clustering are determined by comparing the characteristics of clustered and non-clustered TB patients.

Due to its low-resolution nature, standard genotyping over estimates the proportion of isolates involved in a recent transmission chain and falsely clusters the isolates. This is why WGS (Whole Genome Sequencing), that is characterized by its high-resolution nature, has replaced traditional molecular typing as routine in *Mycobacterium tuberculosis (Mtb)* transmission studies. WGS has been shown to separate isolates that had previously been identified as part of the same transmission chain using traditional genotyping techniques leading to multiple smaller distinct clusters and less clustering [46–51]. As such, with WGS, the factors associated with clustering can be identified with a high degree of accuracy.

Recent literature has seen WGS being used to identify drivers of tuberculosis transmission [38,52,53]. In these studies, recent transmission events are mainly identified using the number of Single Nucleotide Polymorphisms (SNPs) or a Bayesian model that uses WGS data and temporal data such as sample collection dates or dates of symptom onset. These transmission events are related with epidemiological data so as identify factors associated with transmission and thus the drivers of transmission.

WGS was used to study an outbreak of TB in British Columbia, Canada that happened between May 2006 and December 2008 [38]. 36 complete *Mtb* genomes, of which 34 were from the outbreak and 4 were from historical isolates from the same region but sampled before the outbreak with matching genotypes, were sequenced on the Illumina platform (Genome Analyzer II sequencer). Transient living arrangements, crack cocaine use and alcohol were associated with an increased risk of tuberculosis transmission.

In a study in Rural Malawi, WGS of DNA for 1907 culture confirmed TB patients was used to identify transmission events and analyze risk factors associated with transmission [53]. The study analyzed risk factors associated with confirmation of transmission using logistic regression. The number of pairwise SNP differences between isolates were used to identify likely transmission events. Risk factors included: age, sex and HIV status of the index cases and the contact; isoniazid resistance and *Mtb* lineage; relationship, intensity of contact, and time interval between the case and the contact. Intensity of contact was defined as high if the contact was prolonged, indoors and no more than one day, and very high if the case had nursed the prior contact while they were ill.

In the Karonga prevention study in Malawi (Guerra-Assuncao et al., 2015), WGS was used to identify factors associated with recent transmission and transmissibility of TB. Recent

transmission was defined as a clustered isolate (within a SNP threshold of 10 SNP differences) whose most likely source was within 5 years. The seqtrack algorithm implemented in R's adegenet package [55] was used to reconstruct the outbreak and to identify the number of putative secondary cases per source case. Ordered logistic regression was used to assess risk factors for transmission and the number of transmissions. They found that, compared to lineage-4 (the commonest lineage), lineage-2 and lineage-3 strains were more likely to be clustered and in larger clusters and lineage-1 strains were less likely to be clustered and were in smaller clusters. They also found that the elderly (age 50+ years) were less likely to cluster while those living outside the district were more likely to cluster. There was no association between clustering with sex, HIV status, sputum smear status or isoniazid resistance.

Modeling *Mycobacterium tuberculosis* transmission

Deterministic compartmental models have been the go-to modeling methodology for studying transmission of *Mycobacterium tuberculosis* because they are generally easy to formulate and implement and require low computational resources in order to perform simulations. All these models have at least one compartment for latently infected individuals who move to this compartment on infection by a person with infectious TB disease.

Models with two latency compartments [11,12], one for latently infected individuals with a low-risk of progressing to active TB (the slow progressors) and another for latently infected individuals with a high-risk of progressing to active TB disease (the fast progressors) have been shown to produce better fits to observed data compared to those with a single latency compartment [12].

One limitation of deterministic compartmental models is that they are so simplistic in that they assume random (or homogeneous) mixing of individuals in the population meaning that all

susceptible persons have equal probabilities of getting infected which is not always true. In practice, each infectious individual has a finite set of contacts to whom they can pass infection. Individual-based models on the other hand such as network models allow us to account for the nature of mixing of individuals in the population. We can thus explore the effect of the underlying structure of the network on dynamics occurring on the network.

Network models are compelling for studying transmission dynamics of *Mycobacterium tuberculosis* since adequate contact is required for effective transmission to occur. It has always been known that compartmental models are too simplistic. What has been lacking are the necessary tools to implement more accurate connection structures. With the emergency of tools such as the Statnet suite of packages [56], we can explore transmission dynamics of *Mycobacterium tuberculosis* using the more realistic stochastic network models.

PROJECT GOAL

The goal of this project is to characterize extra-household transmission of *Mycobacterium tuberculosis* using social networks of TB patients so as to inform public health interventions aimed at interrupting transmission.

UNDERLYING THEORY

This project is based on two theories:

Epidemic theory

The reproductive number, R, defined as the average number of secondary infectious cases caused by one infectious individual (before they recover or die or are otherwise not able to further transmit) is useful in studying how effectively an infectious disease transmits. A special case of R, called the reproductive number, $R_{0,}$ is where the infection is introduced in a population of totally susceptible individuals such as at the beginning of an outbreak. It is a measure of the

transmission potential of a disease in a particular setting. It does by definition not change during an ongoing outbreak. The more general definition of the reproductive number, R, does change during an epidemic. R is the basic reproductive number discounted by the fraction of the host population that is susceptible (x) i.e., $R= R_0x$. Since, a population will rarely be totally susceptible to an infection in the real world, R is therefore the average number of secondary cases per infectious case in a population made up of both susceptible and non-susceptible hosts.

In its simplest form, R_0 depends on the risk of transmission per contact (β, the attack rate), the number of susceptible contacts per unit time (k, the contact rate) and the duration of infectiousness (d). (i.e., $R_0 = \beta \times k \times d$). We are still lacking in our understanding of these factors with regards to TB yet these factors are important in determining the next generation of cases.

For example, we know little about the mixing patterns of individuals in the population and consequently the contact rate. As such most compartmental models of TB transmission assume homogeneous mixing of the population, which is not always true. Social Network studies have shown that individuals preferentially mix such as with peers, agemates etc. More so, persons bed ridden with TB or any health problem tend to be less mobile. Studying social networks of TB cases will enable us to understand the mixing patterns of the population from which we can derive the contact rate.

We also don't know about the risk of transmission per contact. WGS can enable us to estimate the probability (risk) of transmission given contact between two individuals with the same strain of *Mtb*. Contact can be established via epidemiologic linkage in terms of temporal and spatial connectivity. The time of symptom onset (or time of TB diagnosis) and geographical coincidence such as residing in the same locality or frequenting a particular location (e.g., a bar, school, place of worship etc.) can be used to establish temporal and spatial connectivity,

respectively and as such to establish contact between individuals. The probability of transmission can be calculated from, for example, the genetic distance between isolates using the number of Single Nucleotide Polymorphisms (SNPs) as a distance metric given a particular model of evolution.

The duration of infectiousness is approximated by subtracting the time when an individual started developing symptoms (such as chronic cough) from the time they are 'removed' from the population when a diagnosis is made and treatment is given.

Modes of inheritance in *Mycobacterium tuberculosis*

Inheritance is a key process in the evolution of bacteria and also represents a source of genetic variation in eukaryotes. Transfer of genetic information between individuals is achieved by two mechanisms: vertical, from parent to siblings, and horizontal between individuals of the same or different species [57].

Under vertical inheritance, mutations such as Single Nucleotide Polymorphisms (SNPs) accumulate in the DNA of the daughter cell or DNA is rearranged (e.g., an inversion, insertion or deletion). The daughter cell's DNA is from the parent DNA and differences between the parent cell's DNA and the daughter cell's DNA accumulate over time. Therefore, if we know the rate of accumulation of these differences, we can infer when two bacteria had common ancestor (hence transmission). Vertical inheritance suffices in explaining the evolution of *Mtb* as well as genetic variation in isolates sequenced from different TB patients.

During analysis, the sequenced isolate is compared to a reference strain and regions of difference (such as SNPs) are ascertained and counted. Consequently, the smaller the number of SNP differences, the higher the likelihood of a recent transmission. The number of SNP differences between pairs of outbreak isolates has been used in many studies as a genetic

distance metric to infer the presence of potential transmission links between outbreak isolates [53,58–60].

Under horizontal gene transfer, new DNA is incorporated into the existing bacterial DNA leading to recombination (integration into chromosome) or establishment of a plasmid. The mechanisms for horizontal gene transfer are: transformation – transfer of naked DNA, transduction – transfer of DNA by viruses and conjugation – bacterial mating. Under horizontal gene transfer, the daughter cells' DNA is from parent cell plus other sources of DNA (coming and going). The rate of accumulation of differences between the parent cell's DNA and the daughter cell's DNA can be drastically different with for example SNPS accumulating over time together with added DNA. It's often hard to detect new DNA when aligning a DNA sequence to a reference that doesn't have it. As such inferring transmission becomes challenging due to presence of different types of mutation.

SPECIFIC AIMS

Aim 1

To perform a systematic review of the individual data processing steps, full computational pipelines and the methods used in published studies to infer (confirm or refute) direct transmission of *Mycobacterium tuberculosis* using Whole Genome Sequences from pathogen isolates.

Research questions

a) What individual data processing steps are done when processing WGS data for purposes of making inferences about transmission of *Mycobacterium tuberculosis*?
b) Are there any full computational pipelines for processing *Mycobacterium tuberculosis* pathogen WGS data that have been developed?

c) Which methods are being used in making transmission inferences?

Aim 2

To determine the role of social networks of index TB cases in the transmission of *Mycobacterium tuberculosis*.

Research question

The study seeks to answer the question on whether TB is transmitted in social networks of index tuberculosis patients.

Is TB transmitted in social networks? If yes, what's the relationship between social network structure (such as social network distance) and genetic distance?

What proportion of direct transmission events are between index TB patients with an identifiable path in the social network?

Hypotheses

a) Just like transmission in the household, the proportion of transmission that occurs via the social network of an index TB case is low.
b) The likelihood of direct TB transmission between pairs of index TB cases with the same strain of *Mtb* increases with decrease in social network distance.

Aim 3

To identify critical drivers of *Mycobacterium tuberculosis* transmission in an endemic urban setting in Kampala-Uganda.

Research question

What host, setting and pathogen factors are associated with *Mycobacterium tuberculosis* transmission?

Aim 4

To develop a stochastic network model of *Mycobacterium tuberculosis* transmission.

STRUCTURE OF THE DISSERTATION

Aims 1, 2, 3 and 4 of the study are in chapters 2, 3, 4 and 5 respectively. Each of these chapters is in a manuscript-style format with a standalone abstract, introduction, methods, results, discussion and references. Chapter 6 summarizes the major conclusions and implications from the four aims of the study. The Community Health and Social Networks of TB (COHSONET) study was approved by the Ethics committee of the University of Georgia and that of Makerere University.

REFERENCES

1. Ameni, G., Tadesse, K., Hailu, E., Deresse, Y., Medhin, G., Aseffa, A., … Berg, S. (2013). Transmission of Mycobacterium tuberculosis between Farmers and Cattle in Central Ethiopia. *PLoS ONE*, *8*(10), 1–10. https://doi.org/10.1371/journal.pone.0076891
2. 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*(August), 33180. https://doi.org/10.1038/srep33180
3. Chamie, G., Wandera, B., Marquez, C., Kato-Maeda, M., Kamya, M. R., Havlir, D. V., & Charlebois, E. D. (2015). Identifying locations of recent TB transmission in rural Uganda: A multidisciplinary approach. *Tropical Medicine and International Health*, *20*(4), 537–545. https://doi.org/10.1111/tmi.12459

4. Churchyard, G., Kim, P., Shah, N. S., Rustomjee, R., Gandhi, N., Mathema, B., … Cardenas, V. (2017). What We Know about Tuberculosis Transmission: An Overview. *Journal of Infectious Diseases*, *216*(August), S629–S635. https://doi.org/10.1093/infdis/jix362
5. Cook, V. J., Shah, L., & Gardy, J. (2012). Modern contact investigation methods for enhancing tuberculosis control in Aboriginal communities. *International Journal of Circumpolar Health*, *71*(1), 1–6. https://doi.org/10.3402/ijch.v71i0.18643
6. 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
7. Davies, P. D. O. (2006). Tuberculosis in humans and animals: Are we a threat to each other? *Journal of the Royal Society of Medicine*, *99*(10), 539–540. https://doi.org/10.1258/jrsm.99.10.539
8. Fok, A., Numata, Y., Schulzer, M., & Fitzgerald, M. J. (2008). Risk factors for clustering of tuberculosis cases: a systematic review of population-based molecular epidemiology studies, *12*(March 2007), 480–492.
9. Gardy, J. L., Johnston, J. C., Sui, S. J. H., Cook, V. J., Shah, L., Brodkin, E., … Tang, P. (2011). 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
10. 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), 1–12. https://doi.org/10.1371/journal.pone.0132840

11. Goodreau, S. M., Hamilton, D. T., Jenness, S. M., Sullivan, P. S., Valencia, R. K., Wang, L. Y., … Rosenberg, E. S. (2018). Targeting Human Immunodeficiency Virus Pre-Exposure Prophylaxis to Adolescent Sexual Minority Males in Higher Prevalence Areas of the United States: A Modeling Study. *Journal of Adolescent Health*, *62*(3), 311–319. https://doi.org/10.1016/j.jadohealth.2017.09.023

12. Guerra-Assuncao, J. A., Crampin, A. C., Houben, R. M. G. J., Mzembe, T., Mallard, K., Coll, F., … Glynn, J. R. (2015a). Large-scale whole genome sequencing of M-tuberculosis provides insights into transmission in a high prevalence area. *ELIFE*, *4*. https://doi.org/10.7554/eLife.05166

13. Guerra-Assuncao, J. A., Crampin, A. C., Houben, R. M. G. J., Mzembe, T., Mallard, K., Coll, F., … Glynn, J. R. (2015b). Large-scale whole genome sequencing of M. tuberculosis provides insights into transmission in a high prevalence area. *ELife*, *4*. https://doi.org/10.7554/eLife.05166

14. Gurjav, U., Outhred, A. C., Jelfs, P., Mccallum, N., Wang, Q., Hill-Cawthorne, G. A., … Sintchenko, V. (2016). Whole Genome Sequencing Demonstrates Limited Transmission within Identified Mycobacterium tuberculosis Clusters in New South Wales, Australia. *Australia. PLoS ONE*, *11*(10). https://doi.org/10.1371/journal.pone.0163612

15. Hamilton, D. T., Goodreau, S. M., Jenness, S. M., Sullivan, P. S., Wang, L. Y., Dunville, R. L., … Rosenberg, E. S. (2018). Potential impact of HIV preexposure prophylaxis among black and white adolescent sexual minority males. *American Journal of Public Health*, *108*, S284–S291. https://doi.org/10.2105/AJPH.2018.304471

16. Hamilton, D. T., Rosenberg, E. S., Jenness, S. M., Sullivan, P. S., Wang, L. Y., Dunville, R. L., … Goodreau, S. M. (2019). Modeling the joint effects of adolescent and adult PrEP for

sexual minority males in the United States. *PLoS ONE*, *14*(5), 1–12. https://doi.org/10.1371/journal.pone.0217315

17. Handcock, M. S., Hunter, D. R., Butts, C. T., Goodreau, S. M., & Morris, M. (2003). statnet: Software Tools for the Statistical Modeling of Network Data . *Statnet Project Http: //Statnetproject.Org/*, *24*(1), Seattle, WA. R package version 2.0, URL http://CRA.
18. 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*, 21. https://doi.org/10.1186/s12916-016-0566-x
19. Jajou, R., de Neeling, A., van Hunen, R., de Vries, G., Schimmel, H., Mulder, A., … van Soolingen, D. (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). https://doi.org/10.1371/journal.pone.0195413
20. Jajou, R., Neeling, A. De, Hunen, R. Van, Vries, G. De, Schimmel, H., Mulder, A., … Hoek, W. Van Der. (2018). Epidemiological links between tuberculosis cases identified twice as efficiently by whole genome sequencing than conventional molecular typing: A population-based study, 1–11. https://doi.org/10.1371/journal.pone.0195413
21. Jamieson, F. B., Teatero, S., Guthrie, J. L., Neemuchwala, A., Fittipaldi, N., & Mehaffy, C. (2014). Whole-genome sequencing of the Mycobacterium tuberculosis Manila sublineage results in less clustering and better resolution than mycobacterial interspersed repetitive-unit-variable-number tandem-repeat (MIRU-VNTR) typing and spoligotyping. *Journal of Clinical Microbiology*, *52*(10), 3795–3798. https://doi.org/10.1128/JCM.01726-14
22. Jenness, S. M., Maloney, K. M., Smith, D. K., Hoover, K. W., Goodreau, S. M., Rosenberg, E. S., … Sullivan, P. S. (2019). Addressing Gaps in HIV Preexposure Prophylaxis Care to

Reduce Racial Disparities in HIV Incidence in the United States. *American Journal of Epidemiology*, *188*(4), 743–752. https://doi.org/10.1093/aje/kwy230

23. Jombart, T. (2008). Adegenet: A R package for the multivariate analysis of genetic markers. *Bioinformatics*, *24*(11), 1403–1405. https://doi.org/10.1093/bioinformatics/btn129
24. Klovdahl, A. S. (1985). Social Networks and the Spread of Infectious-Diseases - The AIDS Example. *Social Science & Medicine*, *21*(11), 1203–1216. https://doi.org/10.1016/0277-9536(85)90269-2
25. Latkin, Carl A.; Davey-Rothwell, Melissa A.; Knowlton, Amy R.; Alexander, Kamila A.; Williams, Chyvette T.; Boodram, B. (2013). Social network approaches to recruitment, HIV prevention, medical care, and medication adherence. *Journal of Acquired Immune Deficiency Syndromes*, *63*(0 1), S54–S58. https://doi.org/10.1097/QAI.0b013e3182928e2a.Social
26. Logan, J. J., Jolly, A. M., & Blanford, J. I. (2016). The sociospatial network: Risk and the role of place in the transmission of infectious diseases. *PLoS ONE*, *11*(2), 1–14. https://doi.org/10.1371/journal.pone.0146915
27. Lorenzo-Díaz, F., Fernández-López, C., Lurz, R., Bravo, A., & Espinosa, M. (2017). Crosstalk between vertical and horizontal gene transfer: Plasmid replication control by a conjugative relaxase. *Nucleic Acids Research*, *45*(13), 7774–7785. https://doi.org/10.1093/nar/gkx450
28. Martinez, L., Shen, Y., Mupere, E., Kizza, A., Hill, P. C., & Whalen, C. C. (2017a). Systematic Reviews and Meta-and Pooled Analyses Transmission of Mycobacterium tuberculosis in Households and the Community: A Systematic Review and Meta-Analysis. *American Journal of Epidemiology*, *185*(12). https://doi.org/10.1093/aje/kwx025

29. 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. *American Journal of Epidemiology*, *185*(12), 1327–1339. https://doi.org/10.1093/aje/kwx025

30. Mathema, B., Andrews, J. R., Cohen, T., Borgdorff, M. W., Behr, M., Glynn, J. R., … Wood, R. (2017). Drivers of Tuberculosis Transmission. *Journal of Infectious Diseases*, *216*(April), S644–S653. https://doi.org/10.1093/infdis/jix354

31. 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. *International Journal of Tuberculosis and Lung Disease*, *7*(12 SUPPL. 3), 486–493.

32. 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

33. Ocepek, M., Pate, M., Zolnir-Dove, M., & Poljak, M. (2005). Transmission of Mycobacterium tuberculosis from human to cattle. *Journal of Clinical Microbiology*, *43*(7), 3555–3557. https://doi.org/10.1128/JCM.43.7.3555

34. Packer, S., Green, C., Brooks-Pollock, E., Chaintarli, K., Harrison, S., & Beck, C. R. (2019). Social network analysis and whole genome sequencing in a cohort study to investigate TB transmission in an educational setting. *BMC Infectious Diseases*, *19*(1), 1–8. https://doi.org/10.1186/s12879-019-3734-8

35. 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

36. Roetzer, A., Diel, R., Kohl, T. A., Rueckert, C., Nuebel, 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

37. Rothenberg, R. B., Woodhouse, D. E., Potterat, J. J., Muth, S. Q., Darrow, W. W., & Klovdahl, A. S. (1995). Social networks in disease transmission: The Colorado Springs Study. *NIDA Research Monograph Series*, (151), 3–19. Retrieved from http://www.scopus.com/inward/record.url?eid=2-s2.0-0028991620&partnerID=tZOtx3y1

38. Stop TB Partnership. (2015). Global Plan to End TB: The Paradigm Shift, 2016-2020. https://doi.org/22 August 2016

39. Stucki, D., Ballif, M., Egger, M., Furrer, H., Altpeter, E., Battegay, M., … Fenner, L. (2016). Standard Genotyping Overestimates Transmission of Mycobacterium tuberculosis among Immigrants in a Low-Incidence Country. *Journal of Clinical Microbiology*, *54*(7), 1862–1870. https://doi.org/10.1128/JCM.00126-16

40. Walker, T. M., Ip, C. L. C. L. C., Harrell, R. H. R. H., Evans, J. T. J. T., Kapatai, G., Dedicoat, M. J. M. J., … others. (2012). Whole-genome sequencing to delineate Mycobacterium tuberculosis outbreaks: a retrospective observational study. *The Lancet Infectious Diseases*, *13*, 137–146. https://doi.org/10.1016/S1473-3099(12)70277-3

41. Whalen, C. C. (2016). The replacement principle of tuberculosis. *American Journal of Respiratory and Critical Care Medicine*, *194*(4), 400–401. https://doi.org/10.1164/rccm.201603-0439ED

42. WHO. (2014). *The End TB Strategy*.

43. WHO. (2017). *Global Tuberculosis Report*. https://doi.org/10.1001/jama.2014.11450

44. Wood, R., Racow, K., Bekker, L. G., Morrow, C., Middelkoop, K., Mark, D., & Lawn, S. D. (2012). Indoor social networks in a south african township: Potential contribution of location to tuberculosis transmission. *PLoS ONE*, *7*(6), 4–8. https://doi.org/10.1371/journal.pone.0039246

45. Wyllie, D. H., Davidson, J. A., Smith, E. G., Rathod, P., Crook, D. W., Peto, T. E. A., … Campbell, C. (2018). A Quantitative Evaluation of MIRU-VNTR Typing Against Whole-Genome Sequencing for Identifying Mycobacterium tuberculosis Transmission: A Prospective Observational Cohort Study. *EBioMedicine*. https://doi.org/10.1016/j.ebiom.2018.07.019

TABLES AND FUGRES

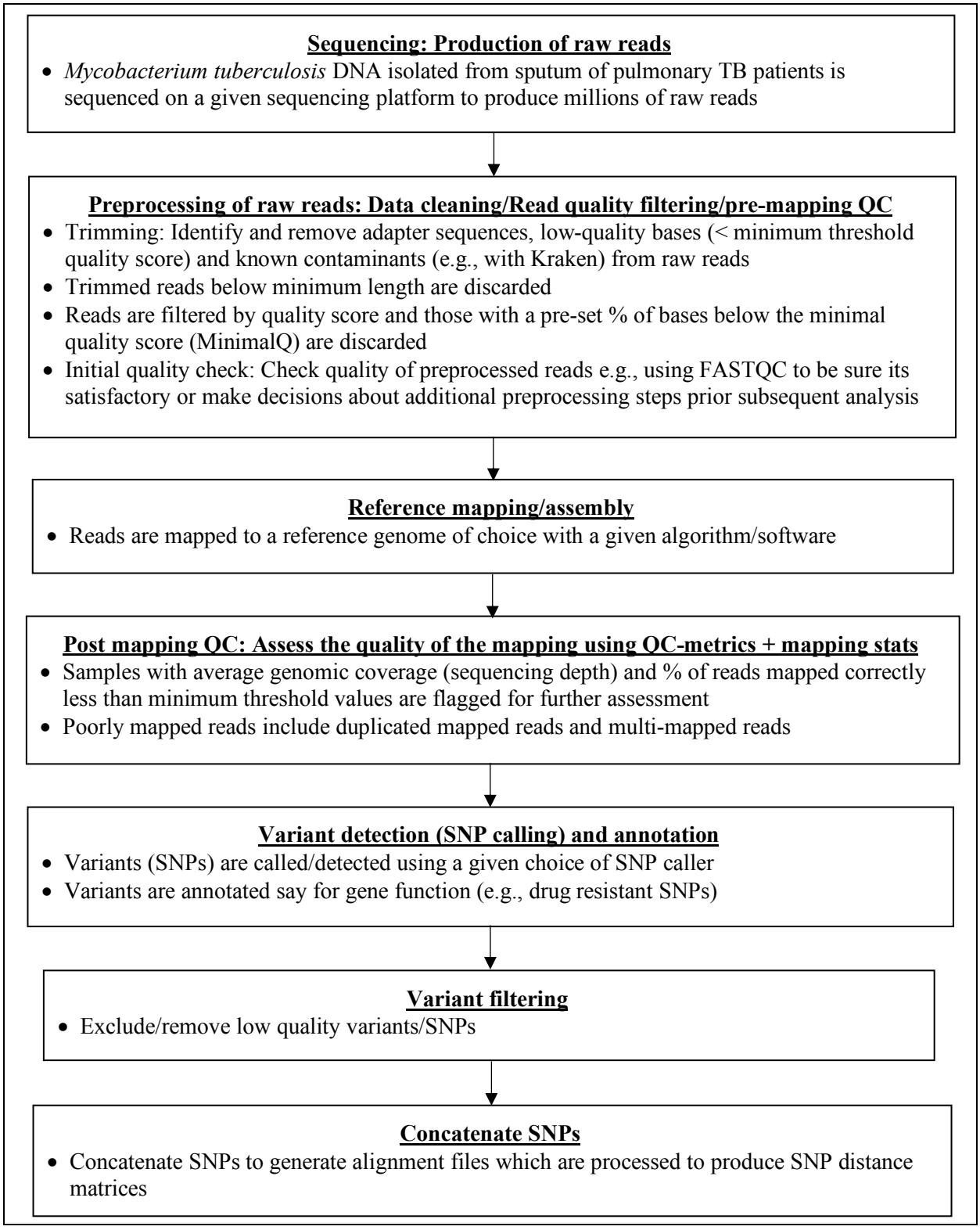


Figure 1.1: A typical reference-based pipeline WGS data processing

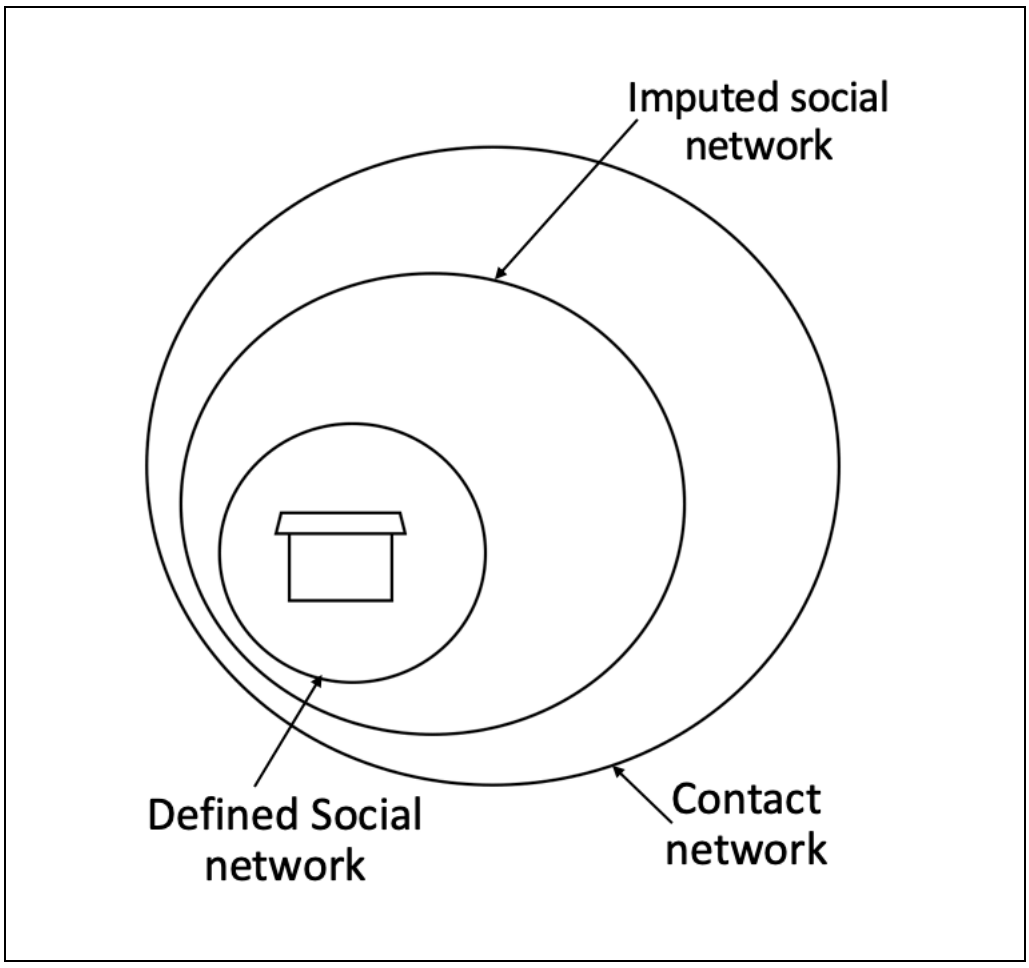


Figure 1.2: Conceptual model of extra-household transmission

CHAPTER 2 : MAKING INFERENCES ABOUT TUBERCULOSIS TRANSMISSION USING WHOLE GENOME SEQUENCING. A SYSTEMATIC REVIEW[1]

[1] R Galiwango, S Kirimunda, A Handel, J Sekandi, L Liu and C Whalen. To be submitted to PLOS Computational Biology

ABSTRACT

Background: Whole genome sequencing (WGS) has improved our ability to identify transmission events by providing better resolution compared with traditional genotyping techniques. We conducted a systematic review to describe the individual data processing steps, computational pipelines and the methods used in WGS studies to infer direct transmission of *Mycobacterium tuberculosis*.

Methods: We searched PubMed and Web of science for all published articles on the topic. The inclusion criteria were: studies that used WGS to study tuberculosis transmission. We excluded articles in which the pathogen under study was not *Mycobacterium tuberculosis*, studies that were not studying transmission, studies in which the method used to infer transmission was not stated, studies in which WGS was not used, reviews, non-English language articles, non-journal articles and articles published after 31st May 2019. We initially screened the pool of retrieved journal articles by removing duplicates. Using the predefined eligibility criteria, we screened articles based on titles, abstract and then the full text. In the end we identified articles to be included in the final review for qualitative synthesis.

Results: Out of the 709 screened articles, 85 were eligible for inclusion in the systematic review.

Data processing: Since 2010, 76 (90%) used the Illumina platform and 70 (82%) used the H37Rv reference genome. Many mapping algorithms and variant callers are used. However, majority of the studies use the BWA-EM algorithm for mapping and SAMtools for variant calling since January 2019. During variant filtering, masking high density variants as well as those in drug resistance and repetitive regions are the consensus.

Computational pipelines: We found five readily available computational pipelines: MTBseq, Bresq, SNVPhyl, NASP and the RedDog pipeline.

Methods used to infer transmission: Use of a SNP threshold is the most widely used method (76.84%) with many thresholds identified in the literature. However, consensus appears with a threshold of 12 SNPs. Other methods used were: Bayesian transmission modeling, using the structure of the phylogeny, shared drug resistance and non-resistance mutations, having an identical SNP pattern, sharing at least two of the same Single Nucleotide Polymorphisms (SNPs) compared with the reference group and overlaying a social network onto a dendrogram obtained from a pairwise SNP difference matrix.

Conclusion: There is heterogeneity in processing of WGS data among studies and some areas of consensus especially in recent literature. Standardization of data processing methodology such as with creation of standardized computational pipelines could improve comparability of transmission inference results. SNP thresholds are the most widely used method for inferring transmission because of their simplicity, with a threshold of 12 SNPs appearing to be the consensus. Bayesian transmission modeling attempts to address their limitation and is increasingly being used in transmission studies.

INTRODUCTION

Reconstructing transmission events during or after an outbreak improves our understanding of TB transmission pathways, thus increasing our ability to interrupt transmission or prevent subsequent outbreaks. Characterizing these events can improve our understanding of routes and patterns of transmission which can translate into meaningful improvements in control activities by informing targeted, evidence-based public health interventions and the allocation of scarce resources.

Whole Genome Sequencing has improved our ability to make inferences about direct transmission of infectious diseases, TB inclusive. Given the low mutation rate of the pathogen [61], a small number of SNPs are expected to separate pairs of isolates that have been involved in a recent transmission event. Such low diversity is better detected by a method that leverages the entire genome compared to traditional molecular typing techniques that only use <0.1% of the bacterial genome. With recent improvements in Next Generation Sequencing (NGS) technologies as well as the reduction in cost and turnaround time of sequencing workflows, WGS has largely replaced traditional molecular typing as routine in *Mycobacterium tuberculosis* transmission studies.

The immediate output of any WGS workflow are millions of 'raw' reads. For transmission inference purposes, the raw reads are processed via a given computational pipeline involving a series of steps, with an initial aim of producing a high-quality sequence for each study sample. The sequences are then analyzed in subsequent steps in order to make inferences on transmission. Variation in the sequencing platforms used and in subsequent data processing steps may lead to heterogeneous results and conclusions as regards to transmission inferences even when the same method is used to make inferences about transmission.

We conducted a systematic review to describe the individual data processing steps, computational pipelines and the methods used in published studies to infer (confirm or refute) direct transmission of *Mycobacterium tuberculosis* using whole genome sequences from pathogen isolates. The rationale behind each data processing step and how it affects results and conclusions relating to transmission is described as well as a discussion of the strengths and limitations of each approach used for making transmission inferences.

To our knowledge, this is the first review of this kind. Previous reviews focused on transmission inference and less on data processing and computational pipelines used [18–21]. Vlad and colleagues studied the sensitivity and specificity of WGS for detection of recent transmission using conventional epidemiology as the gold standard [20,21]. The review by van der Werf and Ködmön [19] focused on use of WGS to investigate international tuberculosis outbreaks. The review by Hatherell and colleagues [18] looked at methods used to infer transmission but included only 12 research articles that were published until 14th July 2015. More studies using WGS to study *Mycobacterium tuberculosis* transmission, employing newer methods for transmission inference and incorporating best practices for WGS data processing, have been published since then.

METHODS

The review was conducted from a pre-set protocol and where relevant, in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA). We searched PubMed and Web of science databases for all published articles on the topic. We also looked through reference lists of included articles for articles that we may have missed during the systematic search.

Inclusion and exclusion criteria

The inclusion criteria were: studies on transmission of *Mycobacterium tuberculosis* that used Whole Genome Sequences from pathogen isolates to study transmission. We excluded: studies in which the pathogen under study was not *Mycobacterium tuberculosis*, studies that were not studying transmission, studies in which the method used to infer transmission was not stated, studies in which Whole Genome Sequence data was not used, reviews, non-English language

articles, non-journal articles (poster or conference abstracts) and articles published after 31st May 2019.

Search strategy

We searched PubMed database using keywords and other search terms relating to "transmission", "mycobacterium tuberculosis" and "whole genome sequencing" for published studies that use Whole Genome Sequences from pathogen isolates to infer *Mycobacterium tuberculosis* transmission (Supplementary table 2.1). The search strategy applied to PubMed (Medline) was adapted for Web of science (Supplementary table 2.2).

Identification of studies

Titles and abstracts of collected studies were screened to remove studies not meeting the inclusion criteria that could be judged on the basis of the title and abstract alone e.g., non-transmission studies, non-*Mtb* transmission studies, studies that don't use WGS (figure 2.1). We then went ahead and retrieved the full texts for the remaining articles. Where in doubt about eligibility of an article at a given processing stage, the article was retained and assessed at the next stage in the eligibility assessment pipeline. This is was done to make sure no articles were excluded pre-maturely.

Data extraction

For each study, we extracted: Bibliographic information (journal, publication month and year, author(s), title), Study type (category of the study, if excluded; the reason for exclusion, Characteristics of the study population/setting (sampling period, country, method used to infer transmission, threshold used to rule in/out transmission, how the transmission threshold was arrived at, maximum number of SNPs between any pair of TB cases; where no threshold was used, kind of epidemiological data used for epidemiological linkage of TB cases), whether the

direction of transmission was inferred and if yes, the method used, Whole Genome Sequencing and subsequent processing steps (sequencing platform/machine used, pipeline to process the raw reads if available, read-quality control steps (quality control tool, whether reads were trimmed, software used for trimming reads, criteria for excluding samples), reference mapping and variant calling steps (mapping algorithm, reference genome, GenBank ID of the reference genome, SNP/variant caller), thresholds for variant calling (base quality score, mapping quality score, alternate allele frequency, depth/coverage), variant filtering/excluded genomic positions (definition of a mixed base, how SNP positions were verified, minority variant frequency, whether repetitive regions of the genome were excluded, positions with missing genotypes across all samples excluded, whether highly clustered SNPs removed, whether SNPs in resistance-related target genes were excluded, whether ambiguous base calls were removed/ignored, whether SNPs close to indels removed).

RESULTS

709 articles were identified after deduplication (figure 2.1). The titles of these articles were screened and 446 of them were dropped, leaving 263 articles only. The abstracts of these articles were screened and 124 of them were dropped, leaving 139 articles. Full texts articles of the 139 were accessed and assessed for eligibility. 85 full text articles met our inclusion and exclusion criteria (figure 2.1; Supporting Information: database of included articles). It is only these articles whose data was extracted and were included in the qualitative synthesis.

Publication timeline

The included articles spanned the years from 2010 to 2019 (figure 2.2A), with the peak appearing in 2018 (28 articles). More articles are expected to be published throughout 2019 and

beyond due to the reduction in sequencing costs and the increased adoption of sequencing technologies in studying *Mycobacterium tuberculosis* transmission.

Geographical locations spanning the included articles

Geographical locations refer to the countries from which the WGS data was collected. In case the study used data from more than one country, all these countries were recorded. Most studies were from European counties (48), followed by the Americas (18) (Figure 2.2B).

Data processing

Sequencing platforms used

Six articles didn't state the sequencing platform that was used. Of the remaining 79 articles (out of 85 included articles), two articles [62,63] used two platforms (Illumina and Ion Torrent) for sequencing (Table 2.1). Of the 81 sequencing platforms used in the 79 articles, Next Generation Sequencing (NGS) by Illumina was the most widely used method (93.83%) in included studies. This can be attributed to the low per base error rate of the platform and the fact that it uses paired end reads that improve the accuracy of the resultant alignment/mapping despite its shorter read length compared to for example Pacific Biosystems and Oxford Nanopore. Various Illumina sequencers were used in included studies: MiSeq, HiSeq, NextSeq, Genome Analyzer and MiniSeq (supplementary materials). Other sequencing platforms used were Ion Torrent (Thermo Fisher Scientific), Applied Biosystems (ABI, particularly the SOLiD 5500XL instrument) and the Yikon Genomics Co. (Jiangsu, China).

Preprocessing of raw reads

Despite the importance of performing an initial quality check, only six articles out of 85 included articles (7.06%) reported having done an initial quality check on raw reads. Of these, five studies used FASTQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) to do quality

control checks on raw sequence data (Table 2.1). The other article used KvarQ [64] for initial quality check. On the other hand, only 18 articles (21.18%) reported having trimmed raw reads (table 4) with the majority using Trimmomatic software [65] to perform the trimming (61.11%) (supplementary material). Other trimming tools used were: PRINSEQ [66], sickle (https://omictools.com/sickle-tool), Trim Galore (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/), Geneious software (https://www.geneious.com) and the CLC Genomics Workbench (https://www.qiagenbioinformatics.com/).

Reference mapping

There were many reference mapping algorithms/software that were used in included studies (Table 2.1). These included: BWA: Burrows Wheeler Aligner [67] which was used 35 times (43.75%), Bowtie 2 [68], SARUMAN [69], Stampy [70], SMALT (https://www.sanger.ac.uk/science/tools/smalt-0), SSAHA [71], CLC Genomics Workbench (https://www.qiagenbioinformatics.com/), Geneious software (https://www.geneious.com), Lasergene Genomics Suite (https://www.txgen.tamu.edu/lasergene-genomics-suite/), MAQ [72], Bionumerics software (http://www.applied-maths.com/bionumerics), BLAST [73], BLAT [74], Bowtie [75], Breseq pipeline [76] , in-house scripts, MTBseq pipeline [77], MUMmer package [78], RedDog pipeline (https://github.com/katholt/RedDog), Ridom SeqSphere software (https://www.ridom.de/seqsphere/), RoVar (unpubished work) and TMAP (https://github.com/iontorrent/TMAP).

Three reference genomes were used in included studies with the majority of the articles (86.25%) using the H37Rv reference genome. Other articles used the CDC1551 reference genome and the hypothetical *Mtb* ancestral genome [79].

SNP calling

Just like for mapping algorithms, a variety of SNP callers were used in included studies (Table 2.1). These included: SAMtools [80] which was used 33 times (38.37%). Others were: in-house scripts, GATK UnifiedGenotyper [81], Pilon [82], FreeBayes (https://github.com/ekg/freebayes), VarScan [83], SSAHA [71], GATK [84], Snippy (https://github.com/tseemann/snippy), Geneious software (https://www.geneious.com), CLC Genomics Workbench (https://www.qiagenbioinformatics.com/), Breseq pipeline [76], Bionumerics software (http://www.applied-maths.com/bionumerics), SMALT (https://www.sanger.ac.uk/science/tools/smalt-0), RoVar (unpubished work), RIDOM Seqsphere software (https://www.ridom.de/seqsphere/), MUMmer package [78], MTBseq pipeline [77], LoFreq [85], Lasergene Genomics Suite (https://www.txgen.tamu.edu/lasergene-genomics-suite/), GenoScreen (http://www.genoscreen.fr/en/genoscreen/147-english), CLC Assembly Cell (https://www.qiagenbioinformatics.com/) and chewBBACA [86].

Variant filtering

Only 27/85 (31.77%) of the eligible articles reported using a base quality threshold in variant (SNP) detection. Most articles used Q20 as the base quality threshold (Table 2.1). Similarly, only 11/85 (12.94%) of the eligible articles reported using a mapping quality threshold in variant (SNP) detection. Three thresholds were found i.e., Q20, Q30 and Q45 (table 3). Only 30/85 (35.29%) of the eligible articles reported using a minimum allele frequency threshold in variant (SNP) detection. 70% of the articles used a minimum allele frequency of 75%, so there appears to be consensus among researchers on the minimum allele frequency threshold.

Depth/coverage thresholds were defined in a variety of ways: either by the number of reads that support a given variant or by the fold coverage (e.g., 20x means on average each base

was sequenced 20 times) or the percentage of total reads that cover a given position or a given percentage of the mean depth of coverage (Table 2.1). Having in mind the difficulty in sequencing repetitive regions, 55/85 (64.71%) of the included articles reported having removed variants found in repetitive regions (table 3). 17/85 (20%) of the included articles reported to have removed highly clustered SNPs i.e., those found within a specified distance of each other (table 3). Many distances for sliding windows were used to define clustered SNPs (supplementary materials). 19/85 (22.35%) of the included articles reported having removed variants found in drug resistance regions (table 3). 7/85 (8.24%) of the included articles reported having removed variants that were close to indels (insertions or deletions).

Full computational pipelines for processing WGS data

Five complete pipelines were found in included studies (Table 2.5).

a) The MTBseq pipeline

The pipeline uses BWA [87] for reference mapping and SAMtools [80] for variant discovery. Quality variants are those that are supported by four reads in both the forward and reverse orientation, respectively, at 75% allele frequency, and by at least four calls with a phred quality score of at least 20 [77]. Variants are filtered for repetitive regions, drug resistance regions and the presence of other variants within a window of 12 bp within the same dataset i.e., filtering for high density variants [77].

b) The bresq pipeline

The pipeline uses Bowtie2 [68] for reference mapping, keeping track of uniquely mapped reads and multi-mapped reads (the repeats). The pipeline provides for trimming of the ends of the reads. Variants are called with frequencies between 0% and 100% with the possibility of calling mixed bases/populations.

c) The SNVPhyl pipeline

The pipeline is performed on the Galaxy platform [88] with each stage of the pipeline implemented as a separate Galaxy tool. The pipeline begins with masking repetitive regions. The reads (either single-end or paired-end) are mapped to the reference genome using SMALT (https://www.sanger.ac.uk/science/tools/smalt-0). SNVPhyl evaluates each pileup for a user-defined mean coverage and any genomes less than this threshold are flagged for further assessment [89]. The pipeline uses both SAMtools [80] and FreeBayes (https://github.com/ekg/freebayes) to call variants independently. The variants are merged into a single file, flagging mismatches between the two. Base calls below the defined minimum read coverage and minimum mean mapping quality are identified and flagged [89]. Finally, high-density SNV regions are discarded.

d) The NASP pipeline

The pipeline starts by masking off duplicated regions. Raw reads are trimmed with Trimmomatic [65]. NASP supports a variety of reference mapping algorithms including BWA [87] and Bowtie2 [68]. It also supports various SNP callers, including SAMtools [80], GATK UnifiedGenotyper [81] and VarScan [83].

e) RedDog pipeline

The pipeline is implemented in python programming language. It performs reference mapping with Bowtie2 [68] and SNP calling with SAMtools/bcftools [80].

Methods used to infer transmission

The SNP or Allelic Difference (AD) threshold (or fewer number of SNP or Allelic differences between isolates) was the most widely used method used to make recent transmission inferences. The method was used 73 times in included articles (76.84%) (Table 2.2). In second

place was use of shared drug resistance mutations i.e., used 10 times in included studies (10.53%). In this method, patients are considered to be involved in a recent transmission event if their isolates share identical drug resistance mutations.

A similar approach used was considering patients to be involved in a recent transmission event if their isolates shared non-drug resistance SNPs that are co-selected with drug resistant SNPs [90]. Five articles used the phylogeny (or structure of the phylogeny) to exclude transmission. In this approach, isolates involved in a recent transmission event must be close to each other on the phylogenetic tree of all isolates and share a common ancestor. Existence of another isolate between possible transmission pairs on the phylogenetic tree is argument against recent transmission [91].

Having an identical SNP pattern [92]- equivalent to zero SNP differences between isolates, sharing ≥2 of the same SNPs compared with the reference group [93], use of a Social network overlaid onto a dendrogram obtained from a pairwise SNP difference matrix [38] were the other methods used.

Bayesian transmission modeling has recently been suggested to infer transmission by combining WGS data with other epidemiological data such as dates of symptom onset (or sample isolation dates), contact network data and spatial data under a Bayesian framework. Three studies used TransPhylo [28], one such methodology, which is implemented as a package in both R statistical software and in MATLAB software.

SNP thresholds were arrived at either by own definition or from published studies (Table 2.3). 57.14% of the articles that used a SNP threshold derived it from the work of Walker and colleagues [94]. In this study, authors estimated the mutation rate of *Mycobacterium tuberculosis* to be 0·5 SNPs per genome per year (95% CI 0·3–0·7) in longitudinal isolates. They predicted that

the maximum number of genetic changes at 3 years would be 5 SNPs and at 10 years would be 10 SNPs. Authors found that none of the epidemiologically linked patients were separated by more than five SNPs (i.e., all links were ≤5 SNPs). 17% of epidemiologically unlinked patients were separated by >5 SNPs and 9% by > 12 SNPs. The authors used these results to construct thresholds for transmission. They expected epidemiological linkage consistent with transmission to exist between isolates differing by ≤5 SNPs, and not to exist between isolates differing by > 12 SNPs. They deemed pairs differing by 6 to 12 SNPs to be indeterminate.

Interestingly, all thresholds derived from the literature were within 12 SNPs, consistent with the work by of Walker and colleagues [94]. For those that used their own thresholds, these ranged from ≤2 to ≤50 for existence of transmission. One study defined 11–99 as uncertain and ≥100 for no transmission (Table 2.3).

For some studies, defining a SNP threshold wasn't necessary because they observed a small number of SNP differences between the isolates (supplementary material). They used these to make inferences on transmission. In these studies, the maximum number of SNP differences between any pair of isolates ranged from 0 to 20. Among epidemiologically linked cases, the maximum number of SNP differences ranged from 5 to 11. One study found a **median** of 5 SNPs between any pair of isolates. Another found a **median** of 1 SNP difference among epidemiologically linked cases.

Inferring the directionality of transmission

The directionality of transmission was inferred using temporal data, Bayesian transmission modeling with TransPhylo [28], the SeqTrack algorithm [95] and order of accumulation of SNPs (Table 2.4). When using temporal data such as sample isolation dates, dates of symptom onset, transmission is inferred forward in time i.e., the isolate with an earlier date is considered the

source and transmission is to the patient with a later isolate. When order of accumulation of SNPs is used, presence of a SNP in other isolates that are not found in a given case suggests directionality from the given case to other cases.

The SeqTrack algorithm builds a directed minimum-spanning tree, minimizing the number of SNPs between links and keeping the temporal data such as disease onset dates, sample collection dates and dates of symptom onset coherent. The algorithm seeks ancestors directly from the sampled isolates, rather than attempting to reconstruct unobserved and hypothetical ancestral transmission events [95]. The TransPhylo model has the advantage of taking into consideration the within host diversity of the pathogen and can be used for both completed and ongoing outbreaks.

DISCUSSION

Main findings

Data processing

Illumina is the most frequently used sequencing platform due to its low per base error rate (<1%). However, the platform produces shorter reads compared to Pacific Biosystems (PacBio) and Oxford Nanopore making it poor at detecting repetitive regions. It compensates for this by using paired-end reads which improve the quality of the alignment. Many mapping algorithms and variant callers are used. However, majority of the studies use the BWA-EM algorithm for mapping and SAMtools for SNP calling. Most studies use the H37Rv TB reference genome. For variant detection and filtering, a 75% minimum allele frequency and a threshold of Q20 for the base and mapping quality are the most used. Depth and coverage are defined in different ways in published literature i.e., as fold coverage, percentage of reads supporting a variant or the number of reads supporting a variant, with differing thresholds being used. During variant filtering,

masking high density variants as well as those in drug resistance and repetitive regions are the consensus.

Full computational pipelines

We found five readily available computational pipelines: MTBseq, Bresq, SNVPhyl, NASP and the RedDog pipeline.

Methods used to infer transmission

Use of a SNP threshold is the most widely used method with many thresholds identified in the literature. However, consensus appears with a threshold of 12 SNPs. Other methods used were: Bayesian transmission modeling, using the structure of the phylogeny, shared drug resistance and non-resistance mutations, having an identical SNP pattern, sharing at least two of the same Single Nucleotide Polymorphisms (SNPs) compared with the reference group and overlaying a social network onto a dendrogram obtained from a pairwise SNP difference matrix.

SNP thresholds are a simple method to use and interpret, which makes them a widely use method. However, SNP thresholds by themselves have a greater dependence on the fraction of sequenced isolates. Hence transmission may be under reported. Seemingly unclustered isolates could have transmission links with un-sequenced isolates. SNP inferred transmission events require corroboration with other epidemiological information.

Bayesian transmission models have made it possible to use both SNP data and epidemiological information simultaneously by combining them via a Bayesian framework where probabilities of transmission are computed using the epidemiological information to weight the transmission probabilities. The TransPhylo model [28], for example has the advantage of taking into account within-host diversity and can also be used for partially sampled and ongoing outbreaks.

Previous reviews

Previous reviews discuss the use of whole genome sequencing in tuberculosis studies giving a general overview of the advantages whole genome sequencing compared to traditional genotyping [20,21,96–99], limitations of whole genome sequencing [20,96,100] and how directionality of transmission is inferred [96]. Croucher and Didelot briefly discuss how direct inference is inferred but their review was not systematic [100].

The review by van der Werf and Ködmön focused on use of WGS to investigate international tuberculosis outbreaks [19]. Vlad and colleagues discuss the use of a threshold of fewer than 6 SNPs and other thresholds to identify recent transmission events [20]. They also discuss quality assurance and the need for standardization in data processing pipelines. Vlad and colleagues studied the sensitivity and specificity of WGS for detection of recent transmission using conventional epidemiology as the gold standard [20,21].

In their systematic review, Hollie-Ann Hatherell and colleagues discuss methods used to infer transmission and directionality of transmission and the implications of these methods on transmission inference. However, the review contained only 12 studies that were published until 14th July 2015. More studies using WGS to study *Mycobacterium tuberculosis* transmission, employing newer methods for transmission inference and incorporating best practices for WGS data processing, have been published since then [18].

Limitations of the study

One limitation of this study is that the information is extracted as reported. For example, researchers may have done a particular data processing step during the analysis but may have not reported it. Nevertheless, the major data processing steps should be reported because each step in the pipeline influences the inferences made.

Conclusions

We found heterogeneity in processing of WGS data among studies and some areas of consensus especially in recent literature. Standardization of data processing methodology could improve comparability of transmission inference results. The five computational pipelines found in the literature bring us closer to standardization of data processing methodology.

While preprocessing of raw reads for example by trimming of adapter sequences and performing an initial quality check prior to mapping them to a reference genome is not mandatory, it is good practice to do perform steps as they reduce the amount of computational resources (RAM, disk space and execution time) needed during subsequent data processing and downstream analysis. It is important to mask SNPs in drug resistance regions so as to rule out selection pressure due to drug resistance. Regions of high SNP density are indicative of recombination and thus masking them is paramount. Repetitive regions are masked due to the difficulty in sequencing such regions with the current technologies. Therefore, a good pipeline for processing WGS data should involve: sequencing of pathogen DNA, preprocessing of raw reads, reference mapping, assessment of the quality of the mapping, variant (SNP) calling and variant filtering.

SNP thresholds are the most widely used method for inferring transmission because of their simplicity, with a threshold of 12 SNPs (or a more stringent threshold of 5 SNPS) appearing to be the consensus. However, there is unlikely to be a single threshold for inferring transmission as the resultant number of SNPs greatly depend on the computational pipeline used to process WGS data. This is an area where we need to do more research: Further research is needed on how WGS can be effectively used to infer transmission more accurately. Bayesian transmission

modeling attempts to address the limitations of SNP thresholds and is increasingly being used in transmission studies.

This systematic review picks up the most recent technologies for WGS, better practices for processing WGS data and most recent studies of TB transmission that use pathogen WGS data and hence provides us with a better understanding of the current state of the field. Without a doubt, the technology will continue to develop and new studies will be published. For example, the premise of Nanopore sequencing (Oxford Nanopore) to produce longer reads and a portable sequencer that can be deployed in the field will revolutionize the field given the reduction in the per base error rate and cost of the sequencing machine, the two biggest limitations of this technology. Therefore, the state of the field will be evaluated regularly.

SUPPLEMENTARY MATERIALS

Epidemiological data used to corroborate WGS inferred transmission events

Epidemiological data used to corroborate WGS inferred transmission included geospatial-temporal data (shared space and time) and mobility information, exposure information, information on infectiousness of cases, previous history of TB and contact tracing data or listed contacts (supplementary data). Geospatial data included shared household, same country of origin and frequenting same community venues. Temporal data included dates for symptom onset, sample isolation dates, enrolment dates, hospital admission and discharge dates. Mobility data encompassed information on travel history such as route and means of migration, country of migration, date of exit and persons encountered en route. Exposure information included being in conversation distance with a case for a cumulative period of at least 8 hours in a closed space or documented cumulative exposure of at least 8 hours or at least 40 hours to, respectively, a sputum smear- or culture-positive but sputum smear-negative source case. In some studies,

smear positive TB cases were considered more infectious than smear negative cases [38] while in others smear-negative TB cases were deemed not infectious [101]. Only pulmonary TB patients were considered infectious.

REFERENCES

1. Afgan, E., Baker, D., Batut, B., Van Den Beek, M., Bouvier, D., Ech, M., … Blankenberg, D. (2018). The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Research*, *46*(W1), W537–W544. https://doi.org/10.1093/nar/gky379
2. Bergval, I., Coll, F., Schuitema, A., de Ronde, H., Mallard, K., Pain, A., … Anthony, R. M. (2015). A proportion of mutations fixed in the genomes of in vitro selected isogenic drug-resistant Mycobacterium tuberculosis mutants can be detected as minority variants in the parent culture. *FEMS Microbiology Letters*, *362*(2), 1–7. https://doi.org/10.1093/femsle/fnu037
3. Blom, J., Jakobi, T., Doppmeier, D., Jaenicke, S., Kalinowski, J., Stoye, J., & Goesmann, A. (2011). Exact and complete short-read alignment to microbial genomes using graphics processing unit programming. *Bioinformatics*, *27*(10), 1351–1358. https://doi.org/10.1093/bioinformatics/btr151
4. Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170
5. Bryant, J. M., Schürch, A. C., van Deutekom, H., Harris, S. R., de Beer, J. L., de Jager, V., … van Soolingen, D. (2013). Inferring patient to patient transmission of

Mycobacterium tuberculosis from whole genome sequencing data. *BMC Infectious Diseases*, *13*(1), 1–12. https://doi.org/10.1186/1471-2334-13-110

6. Casali, N., Broda, A., Harris, S. R., Parkhill, J., Brown, T., & Drobniewski, F. (2016). Whole Genome Sequence Analysis of a Large Isoniazid-Resistant Tuberculosis Outbreak in London: A Retrospective Observational Study. *PLoS Medicine*, *13*(10), e1002137. https://doi.org/10.1371/journal.pmed.1002137
7. Casali, N., Nikolayevskyy, V., Balabanova, Y., Harris, S. R., Ignatyeva, O., Kontsevaya, I., … Drobniewski, F. (2014). Evolution and transmission of drug-resistant tuberculosis in a Russian population. *Nature Publishing Group*, *46*(3). https://doi.org/10.1038/ng.2878
8. Comas, Ĩ., Chakravartti, J., Small, P. M., Galagan, J., Niemann, S., Kremer, K., … Gagneux, S. (2010). Human T cell epitopes of Mycobacterium tuberculosis are evolutionarily hyperconserved. *Nature Genetics*, *42*(6), 498–503. https://doi.org/10.1038/ng.590
9. Croucher, N. J., & Didelot, X. (2015). The application of genomics to tracing bacterial pathogen transmission. *CURRENT OPINION IN MICROBIOLOGY*, *23*, 62–67. https://doi.org/10.1016/j.mib.2014.11.004
10. Deatherage, D. E., & Barrick, J. E. (2014). Identification of mutations in laboratory-evolved microbes from next-generation sequencing data using breseq. *Methods in Molecular Biology (Clifton, N.J.)*, *1151*, 165–188. https://doi.org/10.1007/978-1-4939-0554-6_12
11. 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), 491–501. https://doi.org/10.1038/ng.806

12. Didelot, X., Fraser, C., Gardy, J., Colijn, C., & Malik, H. (2017). Genomic infectious disease epidemiology in partially sampled and ongoing outbreaks. *Molecular Biology and Evolution*, *34*(4), 997–1007. https://doi.org/10.1093/molbev/msw275
13. Dixit, A., Freschi, L., Vargas, R., Calderon, R., Sacchettini, J., Drobniewski, F., … Farhat, M. R. (2019). Whole genome sequencing identifies bacterial factors affecting transmission of multidrug-resistant tuberculosis in a high-prevalence setting. *Scientific Reports*, *9*(1), 5602. https://doi.org/10.1038/s41598-019-41967-8
14. Fine, P. E. M., Crampin, A. C., Houben, R. M. G. J., Mzembe, T., Mallard, K., Coll, F., … Glynn, J. R. (2015). Large-scale whole genome sequencing of M. tuberculosis provides insights into transmission in a high prevalence area. *ELife*, *2015*(4), 1–17. https://doi.org/10.7554/eLife.05166
15. Folkvardsen, D. B., Norman, A., Andersen, A. B., Michael Rasmussen, E., Jelsbak, L., Lillebaek, T., … Lillebaek, T. (2017). Genomic Epidemiology of a Major Mycobacterium tuberculosis Outbreak: Retrospective Cohort Study in a Low-Incidence Setting Using Sparse Time-Series Sampling. *JOURNAL OF INFECTIOUS DISEASES*, *216*(3), 366–374. https://doi.org/10.1093/infdis/jix298
16. Galagan, J. E. (2014). Genomic insights into tuberculosis. *Nature Reviews. Genetics*, *15*(5), 307–320. https://doi.org/10.1038/nrg3664
17. Gardy, J. L., Johnston, J. C., Sui, S. J. H., Cook, V. J., Shah, L., Brodkin, E., … Tang, P. (2011). 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

18. Guerra-Assuncao, J. A., Crampin, A. C., Houben, R. M. G. J., Mzembe, T., Mallard, K., Coll, F., … Glynn, J. R. (2015). Large-scale whole genome sequencing of M. tuberculosis provides insights into transmission in a high prevalence area. *ELife*, *4*. https://doi.org/10.7554/eLife.05166
19. Guerra-assunção, J. A., Houben, R. M. G. J., Crampin, A. C., Mzembe, T., Mallard, K., Coll, F., … Glynn, J. R. (2015). Recurrence due to Relapse or Reinfection With Mycobacterium tuberculosis : A Whole-Genome Sequencing Approach in a Large , Population- Based Cohort With a High HIV Infection Prevalence and Active Follow-up, *211*. https://doi.org/10.1093/infdis/jiu574
20. Guthrie, J. L., Delli Pizzi, A., Roth, D., Kong, C., Jorgensen, D., Rodrigues, M., … Gardy, J. L. (2018). Genotyping and Whole-Genome Sequencing to Identify Tuberculosis Transmission to Pediatric Patients in British Columbia, Canada, 2005-2014. *The Journal of Infectious Diseases*, *218*(7), 1155–1163. https://doi.org/10.1093/infdis/jiy278
21. 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*, 21. https://doi.org/10.1186/s12916-016-0566-x
22. 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

23. Kent, W. J. (2002). BLAT - The BLAST-like alignment tool. *Genome Research*, *12*(4), 656–664. https://doi.org/10.1101/gr.229202. Article published online before March 2002

24. Koboldt, D. C., Chen, K., Wylie, T., Larson, D. E., McLellan, M. D., Mardis, E. R., … Ding, L. (2009). VarScan: Variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics*, *25*(17), 2283–2285. https://doi.org/10.1093/bioinformatics/btp373

25. Kohl, T. A., Utpatel, C., Schleusener, V., De Filippo, M. R., Beckert, P., Cirillo, D. M., & Niemann, S. (2018). MTBseq: a comprehensive pipeline for whole genome sequence analysis of Mycobacterium tuberculosis complex isolates. *PeerJ*, *6*, e5895. https://doi.org/10.7717/peerj.5895

26. Langmead, B. (2011). Alignment with Bowtie, 1–24. https://doi.org/10.1002/0471250953.bi1107s32.Aligning

27. Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357–359. https://doi.org/10.1038/nmeth.1923

28. Le, V. T. M., & Diep, B. A. (2013). Selected insights from application of whole-genome sequencing for outbreak investigations. *Current Opinion in Critical Care*, *19*(5), 432–439. https://doi.org/10.1097/MCC.0b013e3283636b8c

29. Lee, R. S., Radomski, N., Proulx, J.-F., Manry, J., McIntosh, F., Desjardins, F., … Behr, M. A. (2015). Reemergence and amplification of tuberculosis in the Canadian arctic. *The Journal of Infectious Diseases*, *211*(12), 1905–1914. https://doi.org/10.1093/infdis/jiv011

30. 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

31. Li, H., & Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, *26*(5), 589–595. https://doi.org/10.1093/bioinformatics/btp698

32. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

33. Li, H., Ruan, J., & Durbin, R. (2008). Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Research*, *18*(11), 1851–1858. https://doi.org/10.1101/gr.078212.108

34. Lunter, G., & Goodson, M. (2011). Stampy: A statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Research*, *21*(6), 936–939. https://doi.org/10.1101/gr.111120.110

35. Marçais, G., Delcher, A. L., Phillippy, A. M., Coston, R., Salzberg, S. L., & Zimin, A. (2018). MUMmer4: A fast and versatile genome alignment system. *PLoS Computational Biology*, *14*(1), 1–14. https://doi.org/10.1371/journal.pcbi.1005944

36. McGinnis, S., & Madden, T. L. (2004). BLAST: At the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Research*, *32*(WEB SERVER ISS.), 20–25. https://doi.org/10.1093/nar/gkh435

37. McKenna, Aaron, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, David Altshuler, Stacey Gabriel, M. D., DePristo, and M. A., McKenna, A., Hanna, M., Banks, E., Sivachenko, A., … DePristo, M. A. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-

generation DNA sequencing data. *Genome Research*, *20*(9), 254–260. https://doi.org/10.1101/gr.107524.110.20

38. Mehaffy, C., Guthrie, J. L., Alexander, D. C., Stuart, R., Rea, E., & Jamieson, F. B. (2014). Marked microevolution of a unique Mycobacterium tuberculosis strain in 17 years of ongoing transmission in a high-risk population. *PloS One*, *9*(11), e112928. https://doi.org/10.1371/journal.pone.0112928
39. Menardo, F., Duchene, S., Brites, D., & Gagneux, S. (2019). The molecular clock of Mycobacterium tuberculosis. *BioRxiv*, 532390. https://doi.org/10.1101/532390
40. Nikolayevskyy, V, Niemann, S., Anthony, R., van Soolingen, D., Tagliani, E., Kodmon, C., … Cirillo, D. M. (2019). Role and value of whole genome sequencing in studying tuberculosis transmission. *Clinical Microbiology and Infection : The Official Publication of the European Society of Clinical Microbiology and Infectious Diseases*. https://doi.org/10.1016/j.cmi.2019.03.022
41. Nikolayevskyy, Vlad, Kranzer, K., Niemann, S., & Drobniewski, F. (2016). Whole genome sequencing of Mycobacterium tuberculosis for detection of recent transmission and tracing outbreaks: A systematic review. *Tuberculosis*, *98*, 77–85. https://doi.org/10.1016/j.tube.2016.02.009
42. Ning, Z., Cox, A. J., & Mullikin, J. C. (2001). SSAHA: a fast search method for large DNA databases. - Abstract - UK PubMed Central, 1725–1729. https://doi.org/10.1101/gr.194201.1
43. Packer, S., Green, C., Brooks-pollock, E., Chaintarli, K., Harrison, S., & Beck, C. R. (2019). Social network analysis and whole genome sequencing in a cohort study to investigate TB transmission in an educational setting, 1–8.

44. Petkau, A., Mabon, P., Sieffert, C., Knox, N. C., Cabral, J., Iskander, M., … Van Domselaar, G. (2017). SNVPhyl: a single nucleotide variant phylogenomics pipeline for microbial genomic epidemiology. *Microbial Genomics*, *3*(6), e000116. https://doi.org/10.1099/mgen.0.000116

45. Popovici, O., Monk, P., Chemtob, D., Chiotan, D., Freidlin, P. J., Groenheit, R., … Van Der Werf, M. J. (2018). Cross-border outbreak of extensively drug-resistant tuberculosis linked to a university in Romania. *EPIDEMIOLOGY AND INFECTION*, *146*(7), 824–831. https://doi.org/10.1017/S095026881800047X

46. Sahl, J. W., Lemmer, D., Travis, J., Schupp, J. M., Gillece, J. D., Aziz, M., … Keim, P. (2016). NASP: an accurate, rapid method for the identification of SNPs in WGS datasets that supports flexible input and output formats. *Microbial Genomics*, *2*(8), e000074. https://doi.org/10.1099/mgen.0.000074

47. Schmieder, R., & Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, *27*(6), 863–864. https://doi.org/10.1093/bioinformatics/btr026

48. Seraphin, M. N., Didelot, X., Nolan, D. J., May, J. R., Khan, M. S. R., Murray, E. R., … Lauzardo, M. (2018). Genomic Investigation of a Mycobacterium tuberculosis Outbreak Involving Prison and Community Cases in Florida, United States. *The American Journal of Tropical Medicine and Hygiene*, *99*(4), 867–874. https://doi.org/10.4269/ajtmh.17-0700

49. Silva, M., Machado, M. P., Silva, D. N., Rossi, M., Moran-Gilad, J., Santos, S., … Carriço, J. A. (2018). chewBBACA: A complete suite for gene-by-gene schema creation and strain identification. *Microbial Genomics*, *4*(3), 1–7. https://doi.org/10.1099/mgen.0.000166

50. 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*. https://doi.org/10.1186/1471-2164-15-881

51. Takiff, H. E., & Feo, O. (2015). Clinical value of whole-genome sequencing of Mycobacterium tuberculosis. *The Lancet Infectious Diseases*, *15*(9), 1077–1090. https://doi.org/10.1016/S1473-3099(15)00071-7

52. Tyler, A. D., Randell, E., Baikie, M., Antonation, K., Janella, D., Christianson, S., … Sharma, M. K. (2017). Application of whole genome sequence analysis to the study of Mycobacterium tuberculosis in Nunavut, Canada. *PLOS ONE*, *12*(10), e0185656. https://doi.org/10.1371/journal.pone.0185656

53. van der Werf, M. J., & Ködmön, C. (2019). Whole-Genome Sequencing as Tool for Investigating International Tuberculosis Outbreaks: A Systematic Review. *Frontiers in Public Health*, *7*(April), 1–9. https://doi.org/10.3389/fpubh.2019.00087

54. Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., … Earl, A. M. (2014). Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS ONE*, *9*(11). https://doi.org/10.1371/journal.pone.0112963

55. Walker, T M, Monk, P., Smith, E. G., & Peto, T. E. A. (2013). Contact investigations for outbreaks of Mycobacterium tuberculosis: advances through whole genome sequencing. *Clinical Microbiology and Infection*, *19*, 796–802. https://doi.org/10.1111/1469-0691.12183

56. Walker, Timothy M, Clp, C. L., Harrell, R. H., Evans, J. T., Kapatai, G., Dedicoat, M. J., … Peto, T. E. A. (2013). Whole-genome sequencing to delineate Mycobacterium

tuberculosis outbreaks: a retrospective observational study. *LANCET INFECTIOUS DISEASES*, *13*(2), 137–146. https://doi.org/10.1016/S1473-3099(12)70277-3

57. Walker, Timothy M, Ip, C. L. C., Harrell, R. H., Evans, J. T., Kapatai, G., Dedicoat, M. J., … Peto, T. E. A. (2013). Whole-genome sequencing to delineate Mycobacterium tuberculosis outbreaks: a retrospective observational study. *The Lancet Infectious Diseases*, *13*, 137–146. https://doi.org/10.1016/S1473-3099(12)70277-3
58. Walker, Timothy 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
59. Walker, Timothy M, Merker, M., Knoblauch, A. M., Helbling, P., Schoch, O. D., van der Werf, M. J., … Consortium, M.-T. C. (2018). A cluster of multidrug-resistant Mycobacterium tuberculosis among patients arriving in Europe from the Horn of Africa: a molecular epidemiological study. *LANCET INFECTIOUS DISEASES*, *18*(4), 431–440. https://doi.org/10.1016/S1473-3099(18)30004-5
60. Wilm, A., Aw, P. P. K., Bertrand, D., Yeo, G. H. T., Ong, S. H., Wong, C. H., … Nagarajan, N. (2012). LoFreq: A sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. *Nucleic Acids Research*, *40*(22), 11189–11201. https://doi.org/10.1093/nar/gks918
61. Yang, C., Lu, L., Warren, J. L., Wu, J., Jiang, Q., Zuo, T., … Cohen, T. (2018). Internal migration and transmission dynamics of tuberculosis in Shanghai, China: an

epidemiological, spatial, genomic analysis. *LANCET INFECTIOUS DISEASES*, *18*(7), 788–795. https://doi.org/10.1016/S1473-3099(18)30218-4

62. Yang, C., Luo, T., Shen, X., Wu, J., Gan, M., Xu, P., … Gao, Q. (2017). Transmission of multidrug-resistant Mycobacterium tuberculosis in Shanghai, China: a retrospective observational study using whole-genome sequencing and epidemiological investigation. *www.thelancet.com/infection*, *17*. https://doi.org/10.1016/S1473-3099(16)30418-2

TABLES AND FIGURES

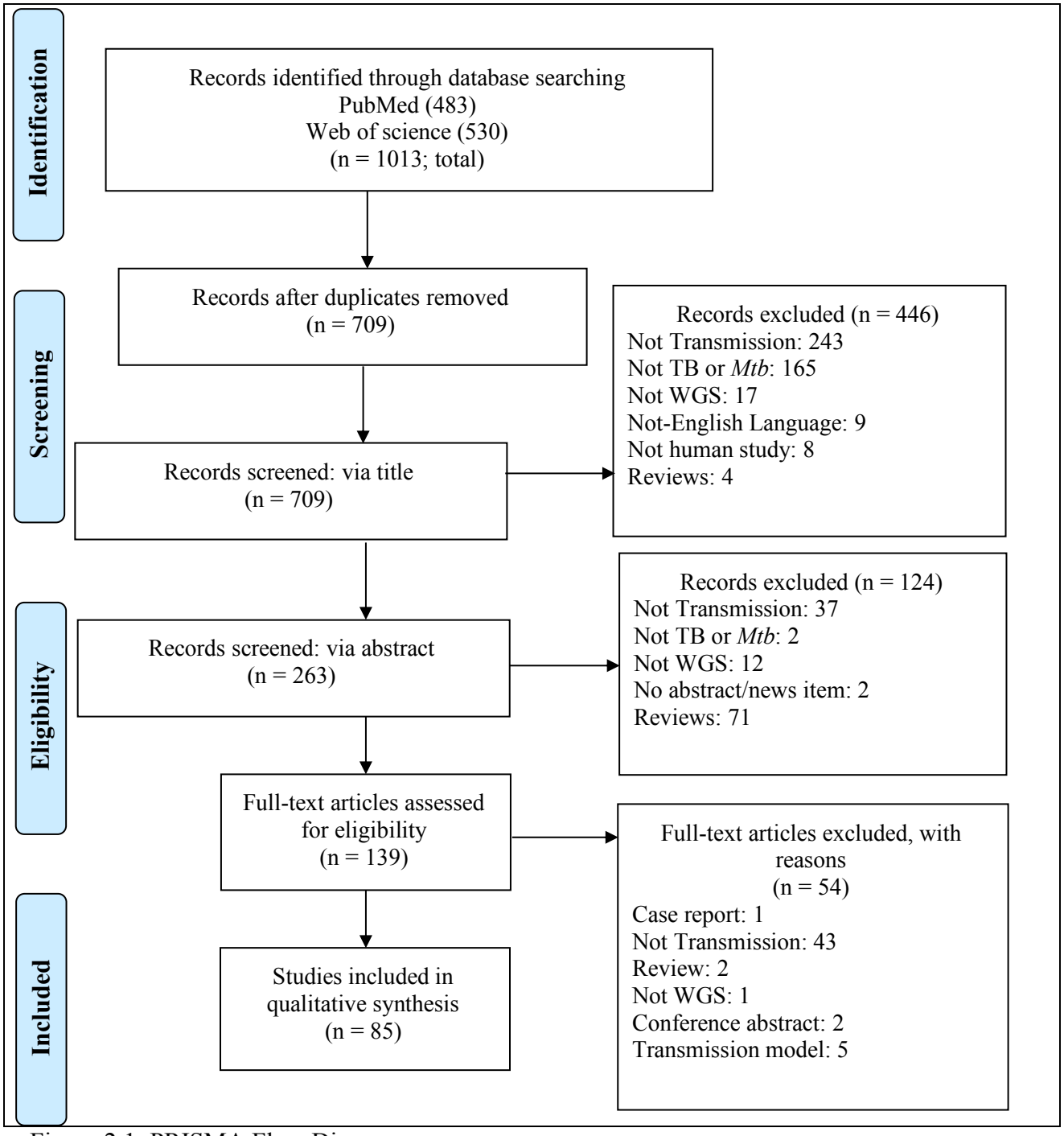


Figure 2.1: PRISMA Flow Diagram

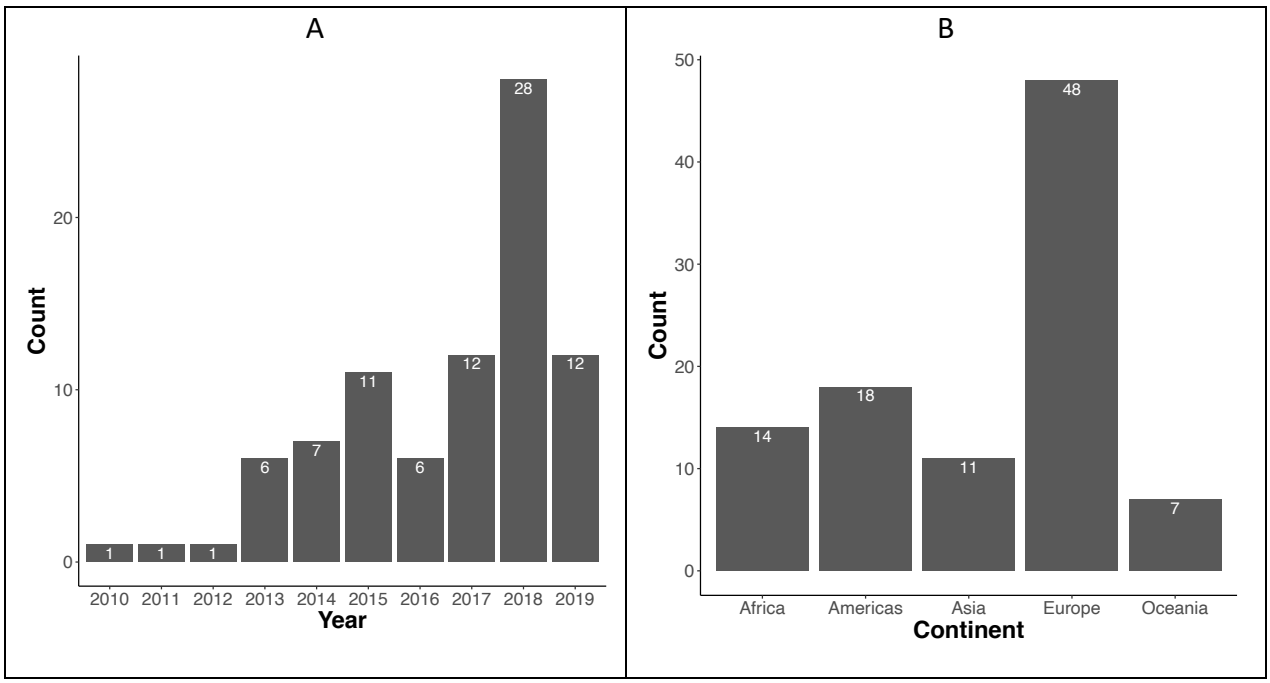


Figure 2.2: A - Total publications per year. B - Geographical locations of included articles

Table 2.1: Pipeline characteristics of included articles.

Total studies included is N=85. Not all studies report all steps.

Pipeline step	Characteristic	N (%)
Sequencing of DNA to produce raw reads	**Sequencing platform**	N=81
	Illumina	76 (93.83)
	Ion Torrent (Thermo Fisher Scientific)	3 (3.70)
	Applied Biosystems (ABI/solid)	1 (1.23)
	Yikon Genomics Co. (Jiangsu, China)	1 (1.23)
	Sequencing platform not stated	6
Preprocessing of raw reads	**Initial quality check of prior to mapping**	5 (7.06)
	Quality trimming (Yes/No)	18 (21.18)
Reference mapping	**Reference genome**	N=78
	H37Rv	70 (89.74)
	CDC1551	3 (3.85)
	hypothetical *Mtb* ancestral genome [79]	5 (6.41)
	Reference genome not stated	7
	Mapping algorithm/software	N=80
	BWA	35 (43.75)
	Bowtie2	7 (8.75)
	SARUMAN, Stampy	6* (7.50*)
	SMALT, SSAHA	3* (3.75*)

	CLC Genomics Workbench, Geneious software, Lasergene Genomics Suite, MAQ	2* (2.50*)
	Bionumerics software, BLAST, BLAT, Bowtie, Breseq pipeline, in-house scripts, MTBseq pipeline, MUMmer package, RedDog pipeline, Ridom SeqSphere software, RoVar, TMAP	1* (1.25*)
	Mapping algorithm not stated	11
Post mapping quality control	**Assess the quality of the mapping (post-assembly analysis)**	
	Exclude multi-mapped reads or those with less than minimum average genomic coverage	8 (9.41)
Variant detection (SNP calling)	**SNP caller**	N=86
	SAMtools	33 (38.37)
	in-house scripts	9 (10.47)
	GATK UnifiedGenotyper	3 (6.98)
	Pilon, FreeBayes	4 (4.65)
	VarScan, SSAHA, GATK	3* (3.49*)
	Snippy, Geneious software, CLC Genomics Workbench, Breseq pipeline, Bionumerics software	2* (2.33*)
	SMALT, RoVar, RIDOM Seqsphere software, MUMmer package, MTBseq pipeline, LoFreq, Lasergene Genomics Suite, GenoScreen, CLC Assembly Cell, chewBBACA	1* (1.16*)
	SNP caller not stated	11
	Detection thresholds	
	Allele frequency (%)	
	75	21 (65.61)
	80	5 (15.63)
	85	1 (3.13)
	90	3 (9.38)
	95	2 (6.25)
	Allele frequency threshold not stated	53
	Depth of coverage	
	Number of reads (range = 2 to 10), fold coverage (range = 4x to 20x), % of reads (range = 50% to 90%), read depth > % of average read depth (10%: 2, 25%: 1)	
	Depth of coverage not stated	29
Variant filtering (Discard low quality SNPs)	**Variant filtering** Base quality threshold (Q20: 14, Q27: 1, Q30: 7, Q40: 1, Q45: 1, Q50: 30) Mapping quality threshold (Q20: 4, Q30: 5, Q45: 2) Exclude SNPs in repetitive regions of the genome Exclude SNPs in drug resistance regions	 27 (31.77) 11 (12.94) 55 (64.71) 19 (22.35)

	Exclude high density SNPs Exclude SNPs that are close to indels	17 (20.00) 7 (8.24)

Table 2.2: Methods used to infer transmission that were used in included studies

Method	Count (%)
Sharing drug resistance mutations	10 (10.53)
Identical SNP pattern	1 (1.05)
Shared non-drug resistance SNPs that are co-selected with drug resistant SNPs	1 (1.05)
Phylogeny/structure of the phylogeny	5 (5.26)
Sharing ≥2 of the same SNPs compared with the reference group	1 (1.05)
SNP/AD threshold (number of pairwise SNP/Allelic differences between isolates)	73 (76.84)
TransPhylo	3 (3.16)
Social network overlaid onto a dendrogram obtained from a pairwise SNP difference matrix	1 (1.05)

Table 2.3: How SNP thresholds were arrived at

SNP threshold method	Summary description	Number of studies that used the method (%)
Casali 2016 [102]	The maximum number of SNPs between any pair of isolates was nine SNPs.	1 (1.59)
Guerra 2015 [52]	On the basis of the distribution of SNP distances between all possible pairs of samples, the authors chose cut-offs at 5 and 10 SNPs for distinguishing links. However, to construct the transmission network, they included links of up to 10 SNPs difference.	2 (3.17)
Nikolayevskyy 2016 [21]	Systematic review of 12 published studies. Authors found that applying a more stringent criteria for epidemiological linkage (<6 SNPs instead of <12 SNPs criteria) only marginally increased the proportion of genomically unconfirmed links (9.4%). As such, a cut-off value of <6 SNPs between isolates was suggested as a predictor for recent transmission.	1 (1.59)
Walker 2013 [94]	The estimated mutation rate was 0·5 SNPs per genome per year (95% CI 0·3–0·7) in longitudinal isolates. The authors predicted that the maximum number of genetic changes at 3 years would be 5 SNPs and at 10 years would be 10 SNPs. Authors found that none of the epidemiologically linked patients were separated by more than five SNPs (i.e., all links were ≤5 SNPs). 17% of epidemiologically unlinked patients were separated by >5 SNPs and 9% by > 12 SNPs. The authors used these results to construct thresholds for	36 (57.14)

	transmission. They expected epidemiological linkage consistent with transmission to exist between isolates differing by ≤5 SNPs, and not to exist between isolates differing by > 12 SNPs. They deemed pairs differing by 6 to 12 SNPs to be indeterminate.	
Yang 2017 [103]	No patients with epidemiological links had strains that were separated by > 12 SNPs. Therefore, the authors defined a genomic cluster in this study as a group of strains that differed by ≤12 SNPs.	2 (3.17)
own	Thresholds ranged from ≤2 to ≤50 for existence of transmission. One study defined 11–99 as uncertain and ≥100 for no transmission. Walker 2013 [94] defined ≤5 as epidemiological linkage consistent with transmission; >12 no existence of epidemiological linkage consistent with transmission and 6-12 indeterminate.	21 (33.33)

Table 2.4: How was the directionality of transmission inferred?

Method	Number of studies that used the method (%)
Temporal data	9 (69.23)
Bayesian Transmission modeling with TransPhylo model	2 (15.38)
SeqTrack algorithm	1 (7.69)
Order of accumulation of SNPs	1 (7.69)

Table 2.5: Full computational pipelines

Pipeline	Mapping software	Variant caller	Definition of a quality variant	Filtered variants	Notes
MTBseq	BWA	SAMtools	-supported by 4 reads, 75% allele freq, ≥Q20	within 12bp window	
bresq	Bowtie2	-User defined allele frequency			-keeps track of uniquely & multi-mapped reads -Provides for trimming -Can call mixed bases
SNVPhyl	SMALT	SAMtools & FreeBayes	-User defined read coverage and mean	Repetitive, high-density	-Runs on Galaxy platform -User-defined mean

			mapping quality		coverage
NASP	BWA or Bowtie2	SAMtools or GATK			Trimmomatic for read trimming
RedDog	Bowtie2	SAMtools			

Supplementary tables and figures

Supplementary table 2.1: Search strategy for PubMed (31st May 2019)

Order of search	Search terms	Number of results
#1 transmission	"transmission"[Subheading] OR "transmission"[All Fields] OR spread[All Fields] OR "disease outbreaks"[MeSH Terms] OR ("disease"[All Fields] AND "outbreaks"[All Fields]) OR "disease outbreaks"[All Fields] OR "outbreak"[All Fields] OR "epidemiology"[Subheading] OR "epidemiology"[All Fields] OR "epidemiology"[MeSH Terms] OR "epidemics"[All Fields] OR "epidemics"[MeSH Terms] OR "epidemic"[All Fields] OR "pandemics"[MeSH Terms] OR "pandemics"[All Fields] OR "pandemic"[All Fields] OR endemic[All Fields] OR "Disease Transmission, Infectious"[Mesh]	2,914,048
#2 mycobacterium tuberculosis	"mycobacterium tuberculosis"[MeSH Terms] OR ("mycobacterium"[All Fields] AND "tuberculosis"[All Fields]) OR "mycobacterium tuberculosis"[All Fields] OR "tuberculosis"[MeSH Terms] OR "tuberculosis"[All Fields] OR TB [All Fields]	269,400
#3 Whole Genome Sequencing	"whole genome sequencing"[MeSH Terms] OR ("whole"[All Fields] AND "genome"[All Fields] AND "sequencing"[All Fields]) OR "whole genome sequencing"[All Fields] OR NGS[All Fields] OR WGS[All Fields] OR ((complete[All Fields] AND ("genome"[MeSH Terms] OR "genome"[All Fields])) OR (whole[All Fields] AND ("genome"[MeSH Terms] OR "genome"[All Fields])) OR (full[All Fields] AND ("genome"[MeSH Terms] OR "genome"[All Fields])) OR (entire[All Fields] AND ("genome"[MeSH Terms] OR "genome"[All Fields])) OR (next[All Fields] AND generation [All Fields])) AND ("sequence"[All Fields] or "sequences"[All Fields] or "sequencing"[All Fields])	143,369
#4	#1 AND #2 AND #3	483

Supplementary table 2.2: Search strategy for Web of science (31st May 2019)

Order of search	Search terms	Number of results
#1 transmission	TS = (transmi* OR spread* OR outbreak* OR epidemiolog* OR epidemic* OR pandemic* OR endemic)	2,430,881
#2 mycobacterium tuberculosis	TS = ("mycobacterium tuberculosis" OR tuberculosis OR TB)	203,919
#3 Whole Genome Sequencing	TS = (("full genome" OR "whole genome" OR "complete genome" OR "entire genome" OR "next generation") NEAR/3 sequenc*) OR WGS OR NGS	73,082
#4	#1 AND #2 AND #3	530

CHAPTER 3 : WHOLE GENOME SEQUENCING AND A LARGE SOCIAL NETWORK STUDY REVEAL THAT TUBERCULOSIS IS MAINLY TRANSMITTED TO CONTACTS OUTSIDE THE SOCIAL NETWORK OF A TB PATIENT[2]

[2] R Galiwango, P Miller, E Yassine, A Handel, J Sekandi, L Liu, R Kakaire, S Zalwango, N Kiwanuka and C Whalen. To be submitted to Journal of Infectious Diseases (JID)

ABSTRACT

Introduction: Tuberculosis (TB) remains a major global health problem with 10 million people suffering from TB disease and several million dying every year. The household of a TB index case has been previously identified as an important setting of transmission for *Mycobacterium tuberculosis*. However, household transmission accounts for a small proportion of the total number of observed new cases, implying that there are other routes of transmission beyond the household that maintain the epidemic in the community. The aim of this analysis was to explore one potential such route, i.e., transmission from the index case to contacts outside the household that are within the social network of a TB case.

Methods: We conducted a large cross-sectional network study, the Community Health and Social Networks of TB (COHSONET) study in Kampala Uganda. Between 2012 and 2016. Two hundred and forty-seven (247) index participants (123 case and 124 controls) and their first-level and second-level contacts were recruited. Whole genome sequencing was done at the CDC for 89 isolates. First, we created an aggregated social network by merging individual second-level egocentric social networks of the 247 indexes. Second, we used an empiric criterion of transmission of 12 SNPs to identify genetically linked patients. Third, we computed the number of genetic links at the respective social network distances between index pairs with an identifiable path between them in the social network. We also computed the proportion of genetic links that were found between patients with no identifiable path. Fourth, we determined the relationship between genetic distance and social network distance.

Results: We found that 43% of the index case pairs who had genetically linked strains of *Mycobacterium tuberculosis* had an identifiable path between them in the social network, but only 13% of these index pairs were found to have a close social distance of one step in the social

network. No genetic links were identified at social network distances from 2 to 6. There were genetically linked pairs at social distances of 7, 8, 10 and 11 with 2, 1, 3 and 1 genetically linked pairs found respectively, corresponding to 9%, 4%, 13% and 4% of the of total number of identified genetic links. There was no correlation between genetic distance and social network distance.

Conclusion: It appears that transmission often happens outside of the defined social network of an individual case. Further exploration of other mechanisms of extra-household transmission of *Mycobacterium tuberculosis* is required. Social network distance could be a poor measure of proximity compared to geographical distance in relation to tuberculosis transmission.

INTRODUCTION

Despite being curable, tuberculosis (TB) remains a major global health problem. It is estimated that over 10 million people suffer from TB every year, the majority of the cases occurring in South-East Asia (45%) and Africa (25%) where the epidemic is predominantly driven by transmission (rather than reactivation of latent infection) and high rates of HIV [6]. TB is the ninth leading cause of death worldwide and has maintained its position, over the past 5 years, as the leading cause from a single infectious agent, ranking above HIV/AIDS and malaria [6]. TB also continues to be the leading cause of death among people living with HIV, accounting for nearly one in three HIV-related deaths [104].

The household of a TB case has been previously identified as an important setting for transmission of *Mycobacterium tuberculosis* [9]. However, as systematic review of children exposed and unexposed to a household member with tuberculosis that included 26 studies found a population attributable fraction of household exposure of 14.1% (95% CI: 11.6, 16.3) [105]. This implies that there are other routes of transmission beyond the household that maintain the

epidemic in the community. This study explores one potential such route, i.e., transmission from the index case to contacts outside the household that are within the social network of the TB case. The social network has two components: a household component which is very geographically defined and membership more completely listed, and the extra-household component which is more geographically diffuse and less complete.

Mycobacterium tuberculosis and other respiratory pathogens, are mainly transmitted when an infectious individual expels pathogens into the air and susceptible persons in close proximity breathe them in. Following a complicated cascade of immunologic events, infection becomes established usually taking 4 – 6 weeks. The fact that close proximity is relevant for transmission to occur makes social network methods a compelling proposal for the study of the spread of such pathogens. The social network imposes a structure or framework on the extra-household members and their relations via which disease spreads. This linkage between individuals via a network enhances our ability to identify and prioritize contacts for evaluation and can hence guide public health intervention. We hypothesized that the extra-household members of the social network would comprise a large proportion of the index case contact network.

Whole genome sequencing overcomes limitations of traditional molecular typing techniques like MIRU-VNTR (Mycobacterial Interspersed Repetitive Units - Variable Number of Tandem Repeats), Spoligotyping and RFPL (Restriction Fragment Length Polymorphism) that lack sufficient discriminatory power to resolve transmission events. With recent improvements in Next Generation Sequencing (NGS) technologies as well as the reduction in cost and turnaround time of sequencing workflows, Whole Genome Sequencing (WGS) has replaced traditional molecular typing as routine in *Mycobacterium tuberculosis* transmission studies [38,60,106–108].

Social network analysis and WGS have been useful in studying TB transmission [38,43]. However, the networks used in these studies are egocentric meaning that index tuberculosis patients are enrolled into the study and asked to list their contacts who are normally not enrolled. Our study extends these studies by evaluating not only the contacts but the broader social network of the index case. Social network analysis and WGS have been useful in mainly low-prevalence areas and not in endemic areas of TB. We conducted the Community Health and Social Networks of TB (COHSONET) to understand transmission in the community by combining traditional epidemiology (i.e. contact tracing), social networks analysis, and WGS. To our knowledge, this is the first time these methodologies have been used in a study of *Mycobacterium tuberculosis* transmission in Africa.

METHODS

Study population

The Community Health and Social Networks of TB (COHSONET) study was a cross-sectional, community-based survey of index TB cases and their social networks. For comparison purposes, the study included a sample of controls and their contacts, without TB disease, who were frequency matched with the index cases by age group, sex, time and residence (parish). The study was conducted in the Rubaga division of Kampala, Uganda between 2012 and 2016. Rubaga is an urban area that comprises 13 parishes and 135 zones (similar to census tracks) that are political units headed by a Local Council (LC). The total population of Rubaga is approximately 383,216 people based on a census performed in 2014. Rubaga is served by one main hospital, 5 public clinics, and numerous private clinics.

The eligibility criteria were: index smear-positive tuberculosis cases, 15 years or older, who resided in Rubaga Division and presented to one of the clinics operated by the National

Tuberculosis Control Program. This age restriction was put in place because persons 15 years and older are more likely to have larger and more interactive social networks in the community than younger persons. We restricted inclusion to smear-positive cases of tuberculosis because they are infectious and most likely to transmit to their contacts. For both index cases and controls, the procedures for enrollment were identical. Household contacts and social network contacts of all ages for the index cases and controls were also eligible for the study.

Ascertainment of each index's social network

The social network of an index (case or control) was defined as members of their household and all individuals living outside their household with whom they had close contact, defined as being within talking distance for more than 4 hours during one or more contact episodes. Thus, each index's social network was ascertained in a two-step process. In the first step, index individuals listed members of their households and all individuals living outside their household with whom they had close contact, defined as being within talking distance for more than 4 hours during one or more contact episodes. These first-level contacts were then traced and evaluated for signs of latent TB infection or active disease. In the second step, the first-level contacts were asked to list their household and extra-household contacts (i.e., second-level contacts of the index participants).

Unless there were concerns for active TB, the field nurses did not trace the second-degree contacts. This sampling methodology was an extended form of egocentric sampling, which we will refer to as "second-level egocentric sampling" in the remaining sections. In addition to what is done in classic egocentric network sampling, second-level egocentric sampling includes an additional layer of contacts (the contacts of contacts).

Data collection and study measurements

Three sputum samples were collected (spot, early morning, night) from persons suspected to have TB (i.e., those with symptoms such as chronic cough). The samples were tested for *Mycobacterium tuberculosis* using microscopy and culture consistent with national guidelines. Demographic (e.g., age, sex, location of household), clinical (e.g., symptoms, risk factors, Karnofsky Performance Scale Index) and social network information (first and second level contacts and relations between them) was obtained through patient interviews using standardized questionnaires administered by trained personnel. The Karnofsky Performance Scale Index was used to assess the heath of the patients. It runs from 0 to 100 where 100 is "perfect" health and 0 is death.

Mycobacterial whole genome sequencing and processing of resultant raw FASTQ files

Mycobacterium tuberculosis chromosomal DNA was extracted from fresh cultures using standard procedures. All isolates were stored frozen in 7H9 broth at -80C for future reference. The extracted DNA was shipped to CDC, Atlanta USA for Whole Genome Sequencing (WGS). Isolates were submitted in batches for sequencing at the CDC. So far, 89 of the 123 index TB cases have been sequenced. Of the 89 submitted isolates, 79 passed set quality standards after sequencing of the whole genomes. Isolated genomic DNA of individual strains was sequenced on the Illumina platform. Resulting FASTQ paired-end reads were processed using the CDC analysis pipeline for studying transmission of *Mycobacterium tuberculosis*.

In brief, the quality of the paired end reads was checked using FAST QC software (http://www.bioinformatics.babraham.ac.uk/projects/fastqc) and trimmed using Trimmomatic [65]. The reads were mapped to the H37Rv reference genome (GenBank accession number NC_000962.3) using the BWA-MEM algorithm [67]. GATK UnifiedGenotyper [81] was used to call

single-nucleotide polymorphisms (SNPs). All SNPs in repeat regions particularly the *PE/PPE* gene families, SNPs found close to the end of a mapped read (within 5bps), SNPs in regions with < 75% coverage or <5x depth of coverage, ambiguous bases and SNPs < 12 bases apart in the genome were excluded.

The resultant SNPs were concatenated and SNP difference tables were generated using Geneious software (https://www.geneious.com), and downstream analysis was performed with R statistical software (https://www.r-project.org/).

Combining individual egocentric social networks to create an aggregated social network

A social network of indexes (cases and controls) and their first level contacts and second level contacts (the contacts of first level contacts) was constructed using R statistical software (https://www.r-project.org/) with the use of the R package Statnet [56]. To start with, an egocentric social network was built around each index (case or control) by creating a link (an edge) between the index and all 'first level' contacts (both household and extra-household contacts) they listed on the census form. After this, links were created between the 'first level' contacts and the 'second level' contacts (the contacts of first level contacts) that they listed (if any) on the census form (figure 3.1A).

Using relational information from the relational dataset that defined relations between listed contacts, links between 'first level contacts' were created where they existed (as described by the index who listed these 'first level contacts'). Similarly, links between 'second level contacts' were added (as described by the traced 'first level contact' who listed these 'second level contacts') (figure 3.1B).

The resultant unconnected individual egocentric social networks were linked to form an aggregated social network by finding persons who appeared in more than one network i.e., the

duplicates. Duplicates were identified using an advanced machine learning and statistical approach implemented in the Dedupe software (https://dedupe.io/). The matching was performed by local content experts who were knowledgeable in local names and their sex affiliation. This ensured quality matching of records. Study participants were assigned unique IDs using information on their names, sex and age.

In a sensitivity analysis (Supplementary Materials: Matching records using Fuzzy string matching), duplicates were merged with approximate (Fuzzy) string matching of concatenated first and last name of query matches using R statistical software (www.r-project.org).

Ethical considerations and data availability

The COHSONET study was approved by both the Institutional Review Board at Makerere University and one at the University of Georgia. De-identified social network data and whole genome sequence data is available upon request to the corresponding author.

Data analysis

For continuous variables, we used the two-sample t-test to examine the difference between the characteristics of index TB patients whose isolates were sequenced compared with those whose isolates were not sequenced. For categorical variables, we used the chi-square test (or Fisher's exact test were at least one cell count was <5) to examine the difference between the characteristics of index TB patients whose isolates were sequenced compared with those whose isolates were not sequenced.

Generally, the lower the number of SNP differences between isolates, the greater the similarity between the strains and the higher the likelihood of a direct transmission between the patients. A threshold of 12 SNP differences between their strains was used to identify genetically

linked TB patients [21,94,103]. The number of genetic links identified at SNP differences from 0 to 20 for the threshold was computed.

In a sensitivity analysis, we used Transphylo [28] to infer genetically related strains based on Maximum Clade Credibility trees for each lineage produced by BEAST [109]. Transphylo is a Bayesian model for inferring transmission trees from time-labelled phylogenies while accounting for within-host diversity of the pathogen and unsampled cases. We used dates of symptom onset, particularly cough, as the tip dates (Symptom start date = Diagnosis date – duration of cough). Transphylo was run for 1 million iterations with the first 10% discarded as burn-in. We used a shape parameter of 1.3 and a rate parameter of 0.3 for the parameters of the gamma distribution for the generation time [28].

We identified patients who were linked in the social network by computing the length of the shortest path between each pair of patients (the geodesic distance). We called this social the network distance. We computed the social network distance between tuberculosis patients with an identifiable path between them in the social network and the number of patient pairs without an identifiable path between them in the social network. We also determined the number of genetic links that were identified between patients at different social network distances.

To determine the relationship between social network distance and genetic distance, we plotted a scatter graph of genetic distance (the number of pairwise Single Nucleotide Polymorphisms (SNPs) between patient isolates) and social network distance (length of the shortest path between each pair) and computed the correlation between genetic distance and social network distance, for patient pairs with an identifiable path in the social network. This analysis was done using the social network built using Dedupe.io software and one where a sensitivity analysis was performed using Fuzzy string matching.

RESULTS

Description of the study population

Between 2012 and 2016, the study enrolled 123 index TB cases and 124 controls. Eleven of the index TB cases were identified from active case finding during field evaluations of index contacts. Eighty-four (68.29%) of the index cases were male while the rest were female (Table 3.1). Their median age was 29 years (range=15 to 63) while 20 (16.26%) of the cases were HIV positive. Four were smear negative while 119 were smear positive TB cases. Two of the sequenced isolates were of lineage 1, fourteen were of lineage 3 while 63 (79.75%) were of lineage 4. There was no statistically significant difference between the characteristics of index TB patients whose isolates were sequenced compared with those whose isolates were not sequenced.

Genetic links between tuberculosis patients

Twenty-three genetic links were identified (Figure 3.3) at a threshold of 12 SNP differences for defining genetic linkage [21,94,103]. One pair of genetic links was between lineage 1 isolates while six were lineage 3 and sixteen were lineage 4 (Table 3.2). The two lineage 1 isolates had zero SNP differences between them. Most lineage 3 isolates were >200 SNP differences between them and another lineage 3 isolate (Supplementary figure 3.2A). There appears to be two distributions for the number of SNP differences between the lineage 4 isolates (Supplementary figure 3.2B).

Fourteen genetic links between the TB patients were identified at a more stringent threshold of 5 SNP differences (Figure 3.3 and Supplementary Table 3.1). One pair of genetic links was between lineage 1 isolates while six were lineage 3 and seven were lineage 4 (Supplementary Table 3.1).

TransPhylo inferred six genetically linked pairs (Supplementary figure 1.3) two of which were lineage 3 and four were lineage 4. TransPhylo inferred that the rest of the index TB cases were linked to unsampled cases. All lineage 3 genetically linked pairs had each one SNP difference between the pairs. Two of the four lineage 4 genetically linked pairs had zero SNP differences between each of the pairs while the other two had one SNP difference between each of the pairs.

The aggregated social network

The aggregated social network (Figure 3.4) had 11,739 nodes including 247 index participants (i.e.,123 index TB cases and 124 index controls), 1,965 first level contacts (of which 930 were listed by the cases and 1035 by the controls) and 9,527 second level contacts (i.e., the contacts of first level contacts). 54.91% of the total number of nodes were male and the rest were female.

The network had 70,161 edges with a density of 0.001 (proportion of observed ties), a mean degree of 12, a median degree of 10 (min=1, max=148) and a clustering coefficient of 0.57 (probability that two contacts of a node are also connected to each other: transitivity). Therefore, on average, each individual in the network was connected with 10 other individuals. With regards to clustering, this is a moderately clustered network considering that the clustering coefficient ranges from 0 to 1, with values close to 0 representing low clustering and those close to 1 representing a highly clustered network.

The network had 47 component networks. The largest component had 9,885 nodes of which 85 were index cases and 102 were index controls. It had 59,797 edges, a density of 0.001 and a clustering coefficient of 0.604.

Multi-case component social networks

Four of the component networks had more than one index TB case (here and after referred to as the multi-case networks: Figure 3.5). Twelve of the components had no index TB case. Thirty-one components each had only one index TB case. Two components each contained two index TB cases. One component had 3 index TB case in it while one component contained the remaining 85 index TB cases.

On the other hand, 24 of the 47 component networks contained no index control, 22 had each 1 index control and one component contained 102 index controls.

Most pairs of patients were at a distance of 4 to 13 from one another in the social network (figure 6). Six pairs were at a close social distance = 1 from one another in the social network. Sixteen pairs were at a social distance = 2 while 8 were at a social distance = 3. There exist pairs that were linked at social distances as high as 14 to 23 (Figure 3.6).

Relating genetic linkage with social network linkage

Among the 6 pairs of patients who were at a close social distance = 1, three pairs had genetically similar strains (Figure 3.7). These correspond to 13% of the 23 genetic links that were identified between the patients in the study. No genetic links were identified at social network distances 2 to 6. Other genetically linked pairs were at a social distance 7, 8, 10 and 11 with 2, 1, 3 and 1 genetically linked pairs respectively. These correspond to 9%, 4%, 13% and 4% of the of total number of identified genetic links. Of the 23 genetic links that were identified between the patients, 13 (57%) were between patients with no identifiable path between them in the social network.

In a sensitivity analysis where the aggregated social network was built using Fuzzy string matching, the number of genetic links between patients with no identifiable path between them

in the social network reduced from 13 to 6 (Supplementary figure 3.1). The number of genetic links between pairs at a close social distance = 1 remained the same (i.e., three). This network shows genetic links between pairs at social network distances 4, 5 and 6 unlike the network built using Dedupe.io software.

When TransPhylo was used to infer genetically linked strains, two of the six identified genetically linked pairs were for index TB patients with no identifiable path between them in the social network. Two genetically linked pairs were at a social distance of one step in the network. One pair consisted isolates that were at a social distance of 7 steps while the other pair were at a social distance of 10 steps (Supplementary figure 2.3).

Correlation of genetic distance with social network distance

There was no correlation between genetic distance and social network distance (correlation coefficient=0.01, p=0.668) (Figure 3.8). A sensitivity analysis with a network built using Fuzzy string matching gave a correlation of −0.05 (p = 0.049) (Supplementary Figure 3.13).

DISCUSSION

In this study we investigated the role of social networks of tuberculosis patients in endemic transmission of *Mycobacterium tuberculosis*. We found that 43% of the index case pairs who had genetically linked strains of *Mycobacterium tuberculosis* using an empiric criterion of transmission of 12 SNPs had an identifiable path between them in the social network, but only 13% of these index pairs were found to have a close social distance of one step in the social network. It therefore appears that transmission often happens outside of the defined social network of an individual case. A sensitivity analysis showed that the definition of a genetic link is important (Supplementary figure 3.12).

In a sensitivity analysis of network construction methodology, identification of duplicate records using Fuzzy string matching gave the same result for the number and percentage of genetic links between TB patients at a close social distance of one step in the social network i.e., 3 (13%). However, the number of genetic links between pairs with an identifiable path in the social network increased from 10 (43%) to 17 (74%). Even with this analysis, most transmission happens outside the defined social network of an individual case (social network distance >2, corresponding to 61% of the identified genetic links).

Identification of duplicate records based on only the first name and last name resulted into more identified matches (duplicates) and consequently more linkage between individual second-level egocentric social networks compared with when an advanced machine learning approach based on the first name, last name, title, other name, sex and age. Using more characteristics of individuals leads to less false matches and consequently less linkage between the individual second-level egocentric social networks.

Most social networks used in the study of infectious diseases are, first-level egocentric social networks. This means, for example for TB, an index case is identified who is asked to list their close contacts (first level contacts). First-level egocentric sampling has been shown to produce biased global statistical properties compared to the underlying census network [110,111]. Our sampling methodology was an extended form of egocentric sampling which we referred to as second-level egocentric sampling. In addition to what is done in classic egocentric network sampling, second-level egocentric sampling includes an additional layer of contacts (the contacts of contacts).

We have explored the social network model as an extension of the contact tracing procedures. Contact procedures are designed just to identify the individuals who have had

adequate contact for *Mycobacterium tuberculosis* transmission. The social network approach is more systematic, broader based, looking at the social roles among individuals listed by the index case to understand the substrate for transmission beyond mere contact investigation. Since we included controls we had the opportunity to understand the additional risk from contact with the index case.

Our findings are consistent with those of previous studies that found that most transmissions were between epidemiologically-unlinked patients. In a study of extensively drug resistant tuberculosis in South Africa, whole genome sequencing revealed that 79% of patients with neither person-to-person nor hospital-based links (the epidemiologically unlinked patients) were within 10 SNPs of at least one other study participant [112]. A study in Malawi found that known contacts only explained 9.4% of transmissions, and that even for those with a prior contact with smear positive tuberculosis in their family, there was a >50% chance that they acquired their TB elsewhere [113]. However, our study extends these studies by evaluating not only the contacts but the broader social network of the index case.

We know from household contact studies that household transmission accounts for <20% of the observed cases [105]. However, even with the new framework that we introduced, we still seem to be missing most transmission events as evidenced by the fact that genetically similar strains occur in contacts who are only distantly connected in the network. This is strong evidence from molecular epidemiology that transmission is occurring within a contact network but outside the social network. Therefore, there must be some other mechanism that brings these people together such as space-time coincidences [114]. In limited resource settings like Uganda, such coincidences may occur when TB patients are seeking care. For example, Sekandi and colleagues [115] found four 'degrees of separation' between the onset of symptoms in a TB patient

and a final diagnosis. Moreover 34% of the total time spent in seeking care prior to TB diagnosis was with non-TB providers. We know that health care settings are places of high transmission, so this suggests that some of the contacts people could have been made in these high-risk settings.

We found no correlation between genetic distance and social network distance. For a disease that requires adequate contact for effective transmission to occur, our hypothesis before the study was that patients at close social network distance are more likely to have genetically similar strains but this wasn't the case. Previous studies have investigated the relationship between genetic distance and geographic distance [51,116–118] and found that patients living at close proximity were more likely to have genetically similar strains. Geographical distance could be a better measure of proximity than social network distance.

One potential limitation of this study is that we did not enroll all consecutive TB patients during the study period. It is also possible that some nodes and edges were miss-specified during the search for duplicates. However, use of local content experts when matching records who were knowledge in local names and their sex affiliation decreased the likelihood of this occurring. Despite these limitations, this study represents the largest most comprehensive social network study of tuberculosis in Africa.

Conclusion

In conclusion, our study has shown that most transmissions happen outside of the defined social network of an individual TB case. Further exploration of other mechanisms of extra-household transmission of *Mycobacterium tuberculosis* is required. One way of doing this is by studying mobility of tuberculosis patients several months prior to diagnosis so as to identify community venues and geographical locations in the community where transmission occurs. We can also

reconstruct community networks of index TB cases by identifying geographical locations spanned by each TB case using cellphone meta data.

SUPPLEMENTARY MATERIALS

Matching records using Dedupe.io software

The resultant second-level egocentric social networks for each index (case or control) were linked to form an aggregated social network by finding persons who appeared in more than one network i.e., the duplicates. Duplicates were identified using an advanced, active machine learning and statistical approach, implemented in the Dedupe.io software (https://dedupe.io/). The software learns the best way to identify similar records in the dataset and uses this training to perform the deduplication.

The records were compared using the first name, last name, sex, title and age. During the training step (the machine learning step of the software), the software provides a random sample of potential duplicates that are either accepted as duplicates by the user or the software is trained that they are records of different individuals. The training process was done by local content experts who are knowledgeable in local names, their social-cultural and sex affiliation. At a minimum, the software requires 10 negative and 10 positive responses for the training but the more the responses the better the de-duplication results will be. We trained the algorithm with 50 positive responses (confirmed duplicates) and 50 negative responses (different individuals).

After the training, the software was run to identify duplicate records. The duplicates identified by the software were verified during the verification step of the software to make sure the software did a good job at matching. After reviewing the identified clusters of potential duplicates, the software provides potential clusters to merge and records to add to clusters. These proposals were reviewed and records were added to clusters if they were the same individual as

those in the cluster. Similarly, clusters of the same individual were merged. This active machine learning procedure is one of the strengths of this approach.

In the final step of the record matching process of the software, the clusters were polished to separate falsely clustered records. At the end of the matching, the software supplies an ID to each record in the database with similar records having the same ID.

We performed another verification process outside the software by comparing the results of the matching process with a database of individuals identified to be the same by study personnel during enumeration and evaluation of indexes (case and controls) and their contacts in the field. The software did well with 95% of the records and the remaining 5% were because of minor deviations in the way the names were written on study forms. These were reviewed and manually added to their respective clusters. We also looked at all the identified clusters of duplicates and verified them accordingly.

Matching records using Fuzzy string matching

Matching was performed at two stages during network building. First, resultant second-level egocentric social networks for each index (case or control) were cleaned by merging duplicate individuals. A conservative matching parameter of 2 differences, representing either two insertions, two substitutions, two deletions (or a pairwise combination of these) between concatenated first and last name for a given query match was used. A conservative matching parameter was used at this stage since duplicate names in a second-level egocentric social network are more likely to be the same individual compared with when duplicates are searched for in the full social network.

Second, the resultant unconnected individual second-level egocentric social networks were linked by finding persons that show up more than once in the network i.e., the duplicates.

In this case, a more stringent matching parameter of one difference between concatenated strings of query names was used to ensure correct matching. A less stringent matching parameter (≥2) resulted into a reduction in the number of indexes (case/control) yet these are confirmed to be unique.

The aggregated social network built using Fuzzy string matching

The network had 10,610 nodes of which 54.91% were male and the rest were female. It had 73,295 edges with a density of 0.0013, a mean degree of 13.8, a median degree of 12 (min=0, max=215) and a clustering coefficient of 0.499.

The network had 12 component networks. The largest component had 10,436 nodes of which 111 were index cases and 124 were index controls. It had 72,010 edges, a density of 0.0013 and a clustering coefficient of 0.495. The remaining 11 components consisted one component with 2 cases and 10 components with each 1 index case (2 multi-case networks). On the other hand, one component had all 124 index controls, while 11 components had no index control (1 multi-control network).

Constructing phylogenetic trees using BEAST software

Model selection using the Maximum Likelihood method was performed with the MEGA7 software to determine a model of nucleotide substitution to use in tree building. 24 different nucleotide substitution models were tested. The General Time Reversible model (GTR) with uniform evolutionary rates among sites and no evolutionary invariable sites had the lowest BIC score (Bayesian Information Criterion).

A coalescent model with constant population size (Kingman 1982) was used for the tree prior. Other priors have been shown to give a tree with same topology. A uniform prior for the

mean of the lognormal distribution of the clock rate and an exponential prior for its standard deviation were used as derived from the literature.

The assumption of a strict/constant molecular clock model across the tree was tested using MEGA7. The null hypothesis of equal evolutionary rate throughout the tree was rejected at a 5% significance level ($p<0.001$) hence a non-correlated relaxed lognormal clock was used.

BEAST was run for 100 million iterations, with the parameter state recorded every 10,000 iterations and the first 10% discarded as burn-in. A maximum clade credibility tree was built for each lineage (lineage 3 and 4) to summarize the posterior sample of trees. No phylogeny was built for lineage 1 since the 2 isolates had zero SNPs differences between them.

Notes

Acknowledgements: We acknowledge the US Centers for Disease Control (CDC) for performing the Whole Genome Sequencing of the pathogen isolates

Financial support: My doctoral training was supported by the Fogarty International Center of the National Institutes of Health (Award Number D43TW010045). The COHSONET study was supported by the National Institute of Allergy and Infectious Diseases of the National Institutes of Health under Award Number R01AI093856.

REFERENCES

1. Almquist, Z. W. (2012). Random errors in egocentric networks. *Social Networks*, *34*(4), 493–505. https://doi.org/10.1016/j.socnet.2012.03.002
2. 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. *EUROPEAN RESPIRATORY JOURNAL*, *52*(4). https://doi.org/10.1183/13993003.00246-2018

3. Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170
4. Borgatti, S. P., Carley, K. M., & Krackhardt, D. (2006). On the robustness of centrality measures under conditions of imperfect data. *Social Networks*, *28*(2), 124–136. https://doi.org/10.1016/j.socnet.2005.05.001
5. Cavany, S. M., Vynnycky, E., Sumner, T., Macdonald, N., Thomas, H. L., White, J., … Anderson, C. (2018). Transmission events revealed in tuberculosis contact investigations in London. *Scientific Reports*, *8*(1), 6676. https://doi.org/10.1038/s41598-018-25149-6
6. Crandall, D. J., Backstrom, L., Cosley, D., Suri, S., Huttenlocher, D., Kleinberg, J., & Graham, R. L. (2010). Inferring social ties from geographic coincidences. *Proceedings of the National Academy of Sciences*, *107*(52), 22436–22441. https://doi.org/10.1073/pnas.1006155107
7. Didelot, X., Fraser, C., Gardy, J., Colijn, C., & Malik, H. (2017). Genomic infectious disease epidemiology in partially sampled and ongoing outbreaks. *Molecular Biology and Evolution*, *34*(4), 997–1007. https://doi.org/10.1093/molbev/msw275
8. Drummond, A. J., & Rambaut, A. (2007). BEAST: Bayesian evolutionary analysis by sampling trees, *8*, 1–8. https://doi.org/10.1186/1471-2148-7-214
9. Gardy, J. L., Johnston, J. C., Sui, S. J. H., Cook, V. J., Shah, L., Brodkin, E., … Tang, P. (2011). 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

10. Glynn, J. R., Guerra-Assunção, J. A., Houben, R. M. G. J. 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), 1–12. https://doi.org/10.1371/journal.pone.0132840
11. Handcock, M. S., Hunter, D. R., Butts, C. T., Goodreau, S. M., & Morris, M. (2003). statnet: Software Tools for the Statistical Modeling of Network Data. *Statnet Project Http: //Statnetproject.Org/*, *24*(1), Seattle, WA. R package version 2.0, URL http://CRA.
12. Jajou, R., Neeling, A. De, Hunen, R. Van, Vries, G. De, Schimmel, H., Mulder, A., … Hoek, W. Van Der. (2018). Epidemiological links between tuberculosis cases identified twice as efficiently by whole genome sequencing than conventional molecular typing: A population-based study, 1–11. https://doi.org/10.1371/journal.pone.0195413
13. Li, H., & Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, *26*(5), 589–595. https://doi.org/10.1093/bioinformatics/btp698
14. Martinez, L., Shen, Y., Mupere, E., Kizza, A., Hill, P. C., & Whalen, C. C. (2017). 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
15. McKenna, Aaron, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, David Altshuler, Stacey Gabriel, M. D., DePristo, and M. A., McKenna, A., Hanna, M., Banks, E., Sivachenko, A., … DePristo, M. A. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-

generation DNA sequencing data. *Genome Research*, *20*(9), 254–260. https://doi.org/10.1101/gr.107524.110.20

16. Morrison, J., Pai, M., & Hopewell, P. C. (2008). Tuberculosis and latent tuberculosis infection in close contacts of people with pulmonary tuberculosis in low-income and middle-income countries: a systematic review and meta-analysis. *The Lancet Infectious Diseases*, *8*(6), 359–368. https://doi.org/10.1016/S1473-3099(08)70071-9
17. Nikolayevskyy, V., Kranzer, K., Niemann, S., & Drobniewski, F. (2016). Whole genome sequencing of Mycobacterium tuberculosis for detection of recent transmission and tracing outbreaks: A systematic review. *Tuberculosis*, *98*, 77–85. https://doi.org/10.1016/j.tube.2016.02.009
18. Packer, S., Green, C., Brooks-Pollock, E., Chaintarli, K., Harrison, S., & Beck, C. R. (2019). Social network analysis and whole genome sequencing in a cohort study to investigate TB transmission in an educational setting. *BMC Infectious Diseases*, *19*(1), 1–8. https://doi.org/10.1186/s12879-019-3734-8
19. Roetzer, A., Diel, R., Kohl, T. A., Rü Ckert, C., Nü Bel, 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*. https://doi.org/10.1371/journal.pmed.1001387
20. Stop TB Partnership. (2015). Global Plan to End TB: The Paradigm Shift, 2016-2020. https://doi.org/22 August 2016
21. Stucki, D., Ballif, M., Egger, M., Furrer, H., Altpeter, E., Battegay, M., … Fenner, L. (2016). Standard genotyping overestimates transmission of *Mycobacterium tuberculosis*

among immigrants in a low incidence country. *Journal of Clinical Microbiology*, *54*(May), JCM.00126-16. https://doi.org/10.1128/JCM.00126-16

22. Walker, T. M., Ip, C. L. C., Harrell, R. H., Evans, J. T., Kapatai, G., Dedicoat, M. J., … Peto, T. E. A. (2013). Whole-genome sequencing to delineate Mycobacterium tuberculosis outbreaks: a retrospective observational study. *The Lancet Infectious Diseases*, *13*, 137–146. https://doi.org/10.1016/S1473-3099(12)70277-3
23. WHO. (2017). *Global Tuberculosis Report*. https://doi.org/10.1001/jama.2014.11450
24. Yang, C., Luo, T., Shen, X., Wu, J., Gan, M., Xu, P., … Gao, Q. (2017). Transmission of multidrug-resistant Mycobacterium tuberculosis in Shanghai, China: a retrospective observational study using whole-genome sequencing and epidemiological investigation. *www.thelancet.com/infection*, *17*. https://doi.org/10.1016/S1473-3099(16)30418-2

TABLES AND FIGURES

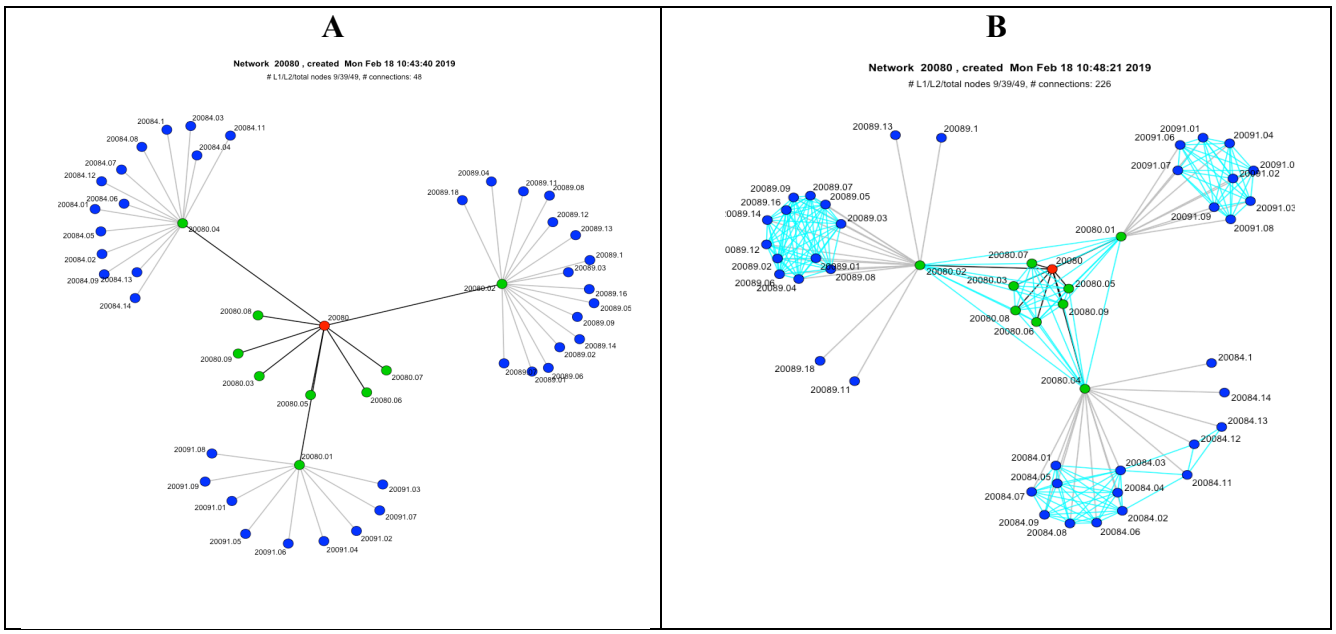


Figure 3.1: Second-level egocentric social network building

A: Second-level egocentric social network of index 20080. The index is shown in red while the 'first level contacts' are shown in green and the 'second level contacts' in blue. B: Second-level egocentric social network of index 20080 with relational links between 'first level contacts' or 'second level contacts' where they existed (as described by the index who listed these 'first level contacts' or 'the first level contact who listed these second level' contacts respectively).

Table 3.1: Characteristics of index tuberculosis patients

Characteristic	All (N=123)		Sequenced isolates (N1=79)		Not sequenced (N2=44)		p-value
	n	%	n1	%	n2	%	
Sex							
Male	84	68.29	54	68.35	30	68.18	1
Female	39	31.71	25	31.65	14	31.82	
Median age (range)	28(15,63)		29(17-59)		26.5(15,63)		0.6417
HIV status							0.7089
Positive	20	16.26	12	15.19	8	18.18	
Negative	98	79.67	66	83.54	32	72.73	
Missing	5	4.07	1	1.27	4	9.09	
Median BMI (range)	32.57(22.15, 52.08)		32.57(23.41, 52.08)		32.44(22.15, 41.26)		0.274
Alcohol use							0.5574
Yes	49	39.84	34	43.04	15	34.09	
No	72	58.53	45	56.96	27	61.36	
Missing	2	1.63			2	4.55	

Smoking							0.9507
Current smoker	14	11.38	10	12.66	4	9.09	
Former smoker	18	14.63	12	15.19	6	13.64	
Never smoked	89	72.36	57	72.16	32	72.73	
Missing	2	1.63			2	4.54	
Previous TB disease							0.1403
Yes	18	14.63	15	18.99	3	6.82	
No	103	83.74	64	81.01	39	88.64	
Missing	2	1.63			2	4.54	
Lineage							
1			2	2.53			
3			14	17.72			
4			63	79.75			
Smear status							0.6167
Negative	4	3.25	2	2.53	2	4.55	
Positive	119	96.75	77	97.47	42	95.45	
Median cough duration in months (range)	2(0.46,24)		2(0.46,24)		2.5(0.69,12)		0.0852
BCG scar present							0.5629
Present	92	74.80	58	73.42	34	77.27	
Absent	26	21.14	18	22.78	8	18.18	
Uncertain	3	2.44	3	3.80			
Missing	2	1.63			2	4.55	
Median Karnofsky score (range)	90 (40-100)		90 (40-100)		85(50,90)		0.7096

**The p-value is for comparison between characteristics of patients whose isolates were sequenced compared with those whose isolates were not sequenced. Missing: Data filled wasn't filled on the questionnaire.*

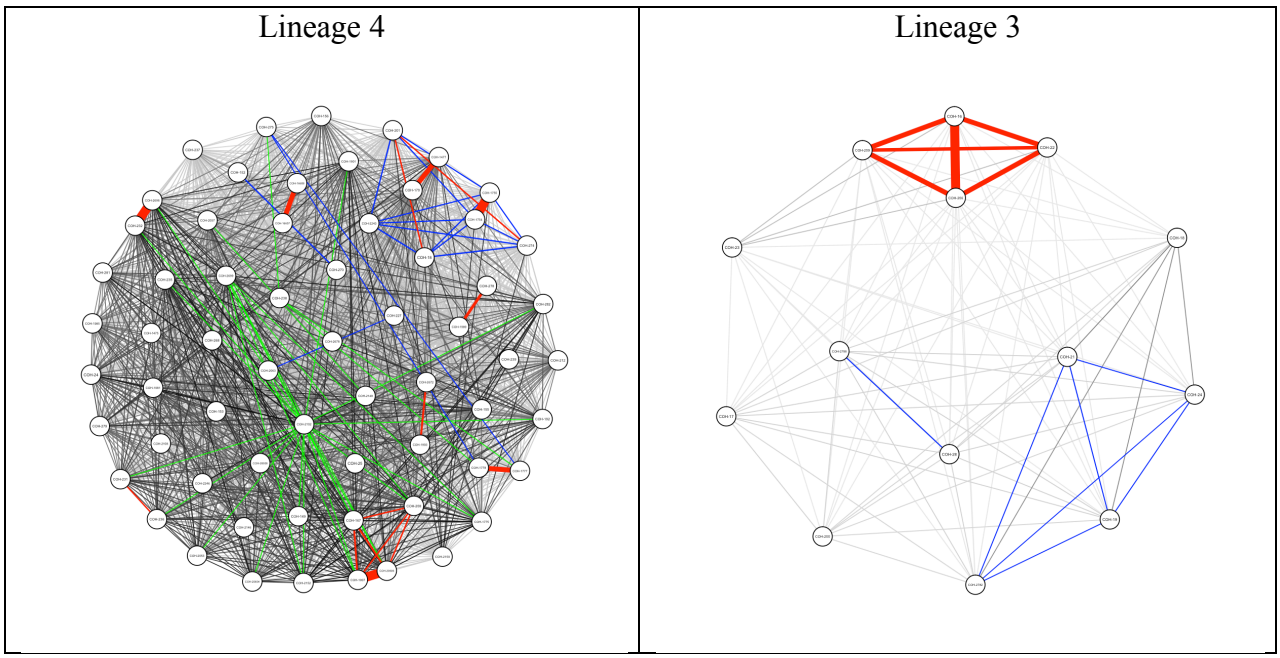


Figure 3.2: Pairwise SNP difference matrices visualized as networks

**(colored by number of SNPs: 0-12 red, 13-50 blue, 51-100 green, >100 black)*

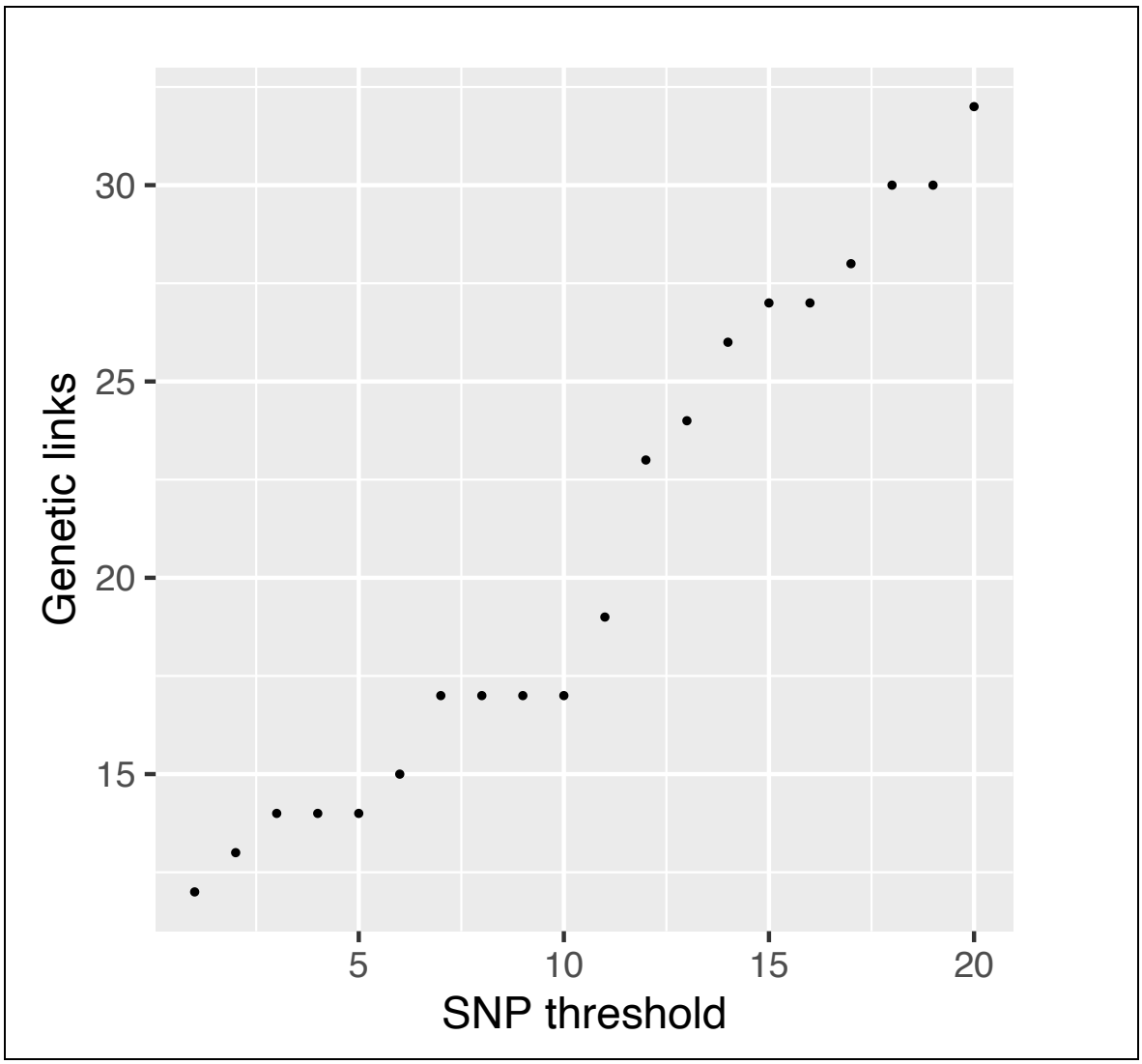


Figure 3.3: Number of genetic links per SNP threshold

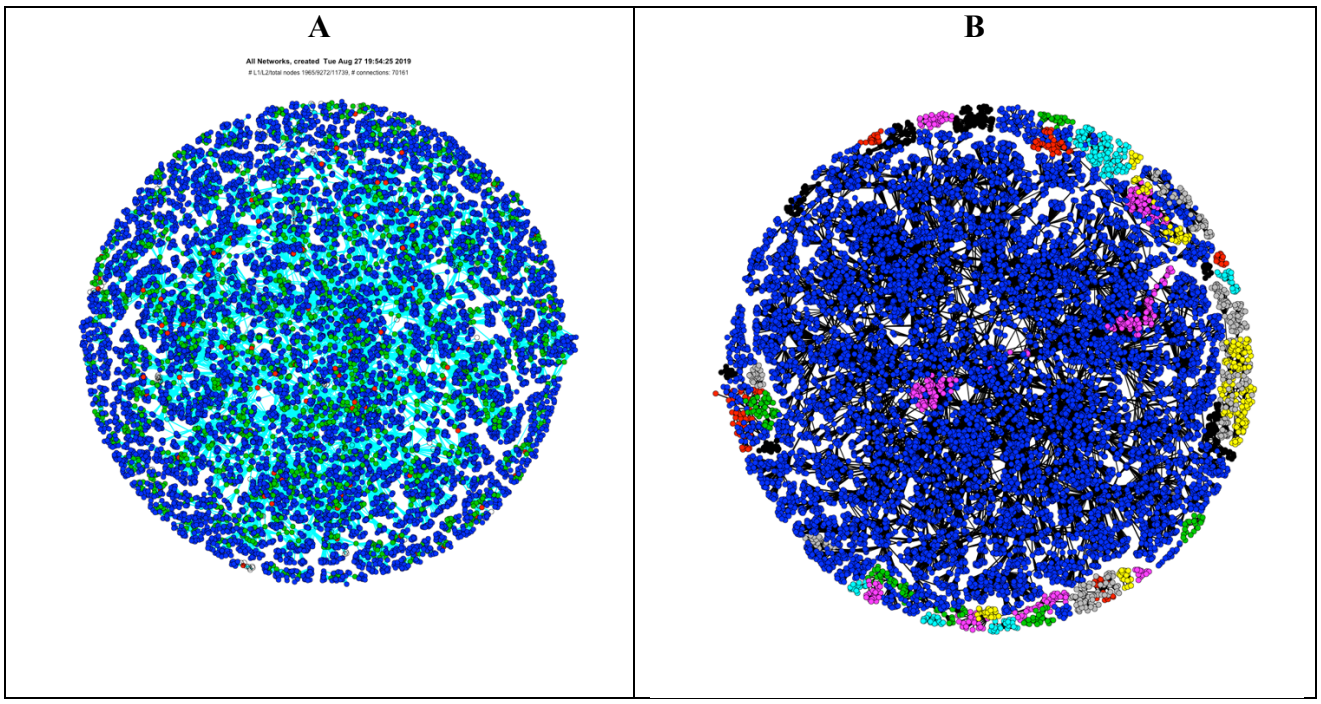


Figure 3.4: Aggregated social network and the largest component

A: Aggregated social network. B: Aggregated social network with colors showing the different component networks that make up the full network. In blue is the biggest component.

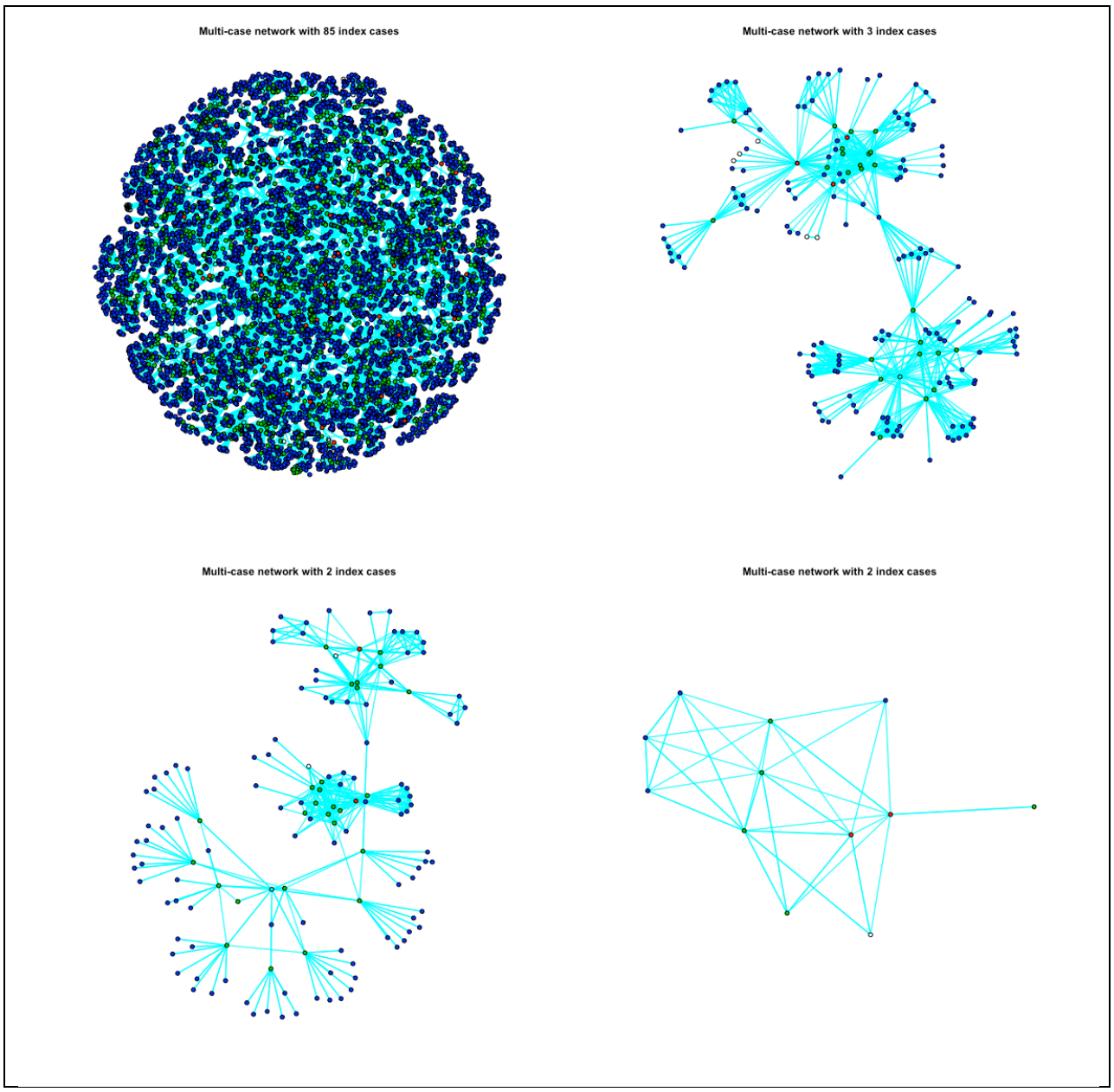


Figure 3.5: Multi-case networks

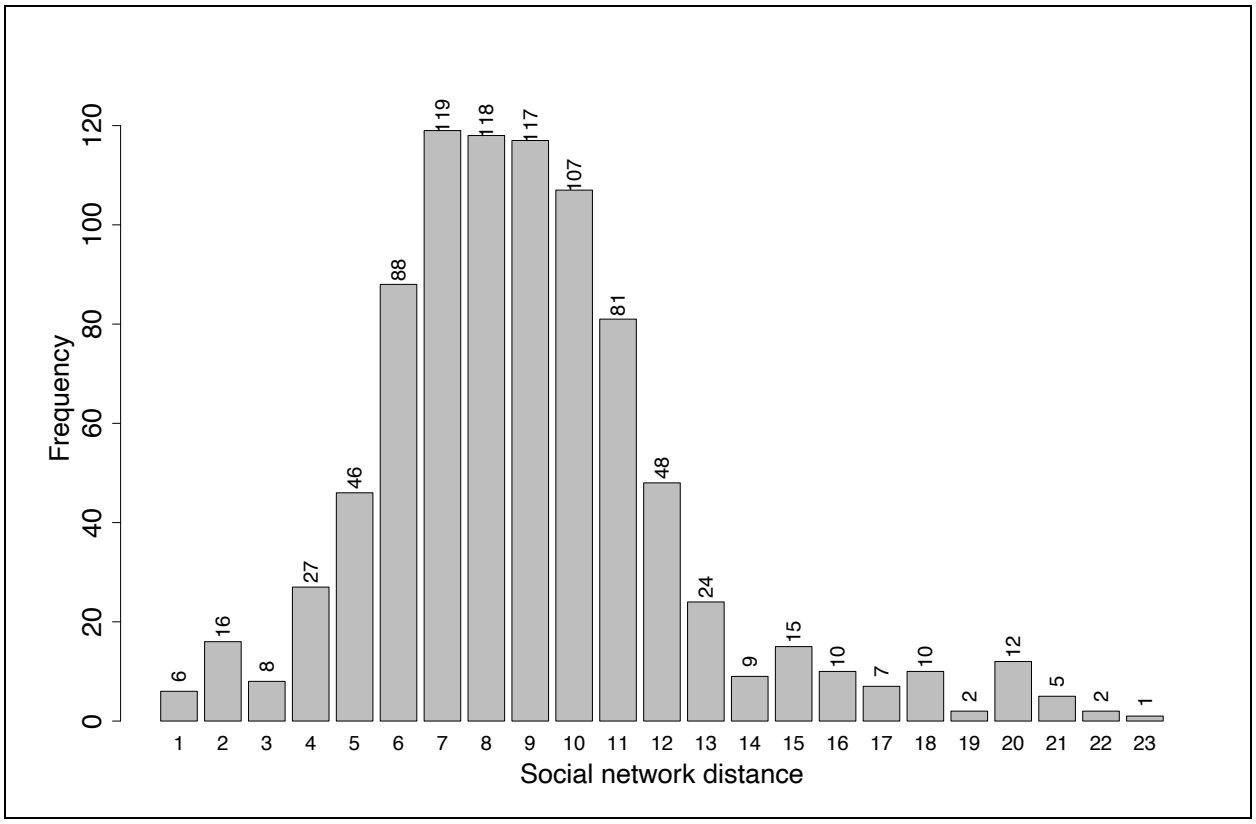


Figure 3.6: Distribution of pairwise social network distance between tuberculosis patients with an identifiable path between them in the social network

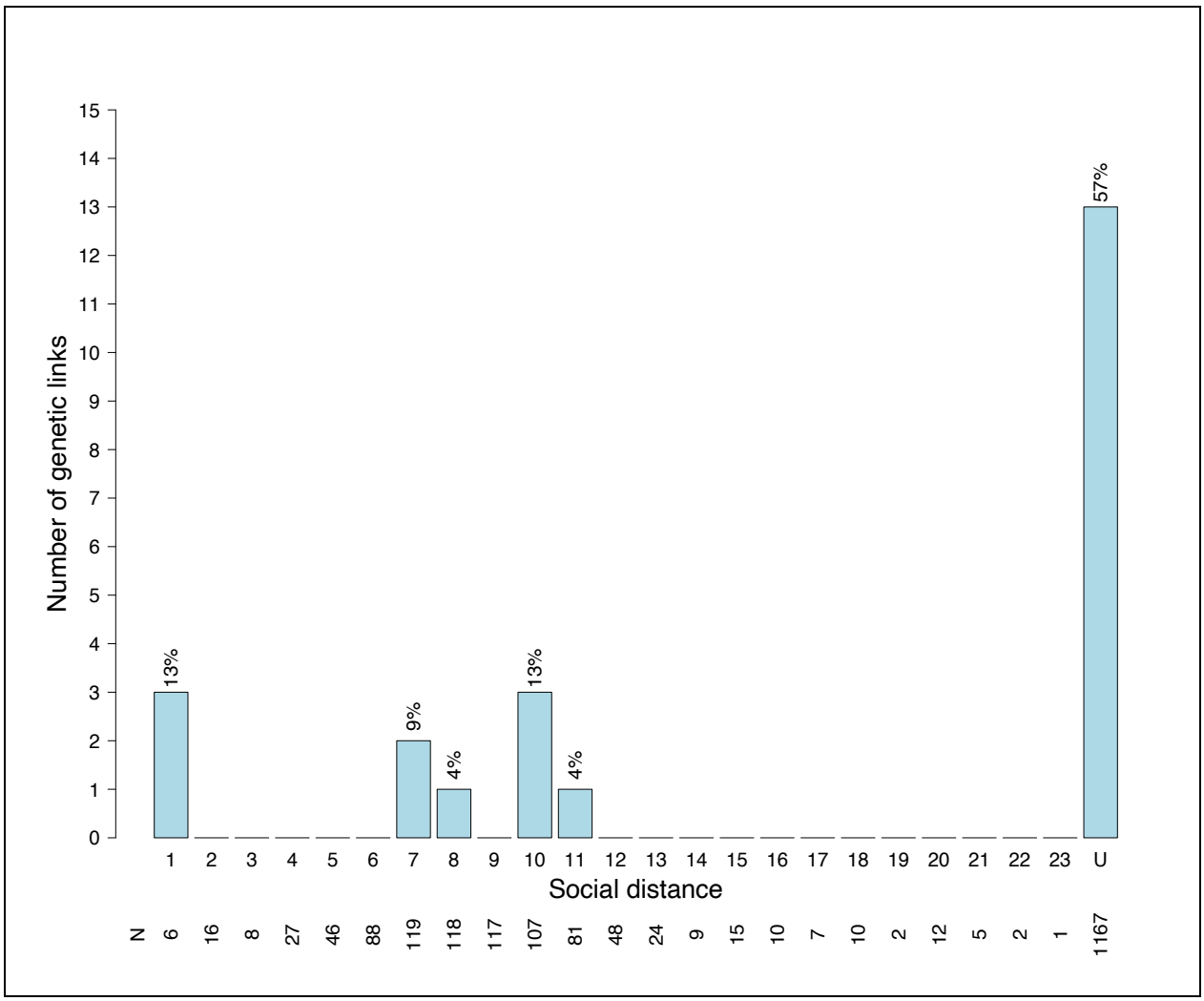


Figure 3.7: Number of genetic links identified per social network distance

**The lower row shows the number of identified social links* (N) *between the tuberculosis patients at a given social network distance.*

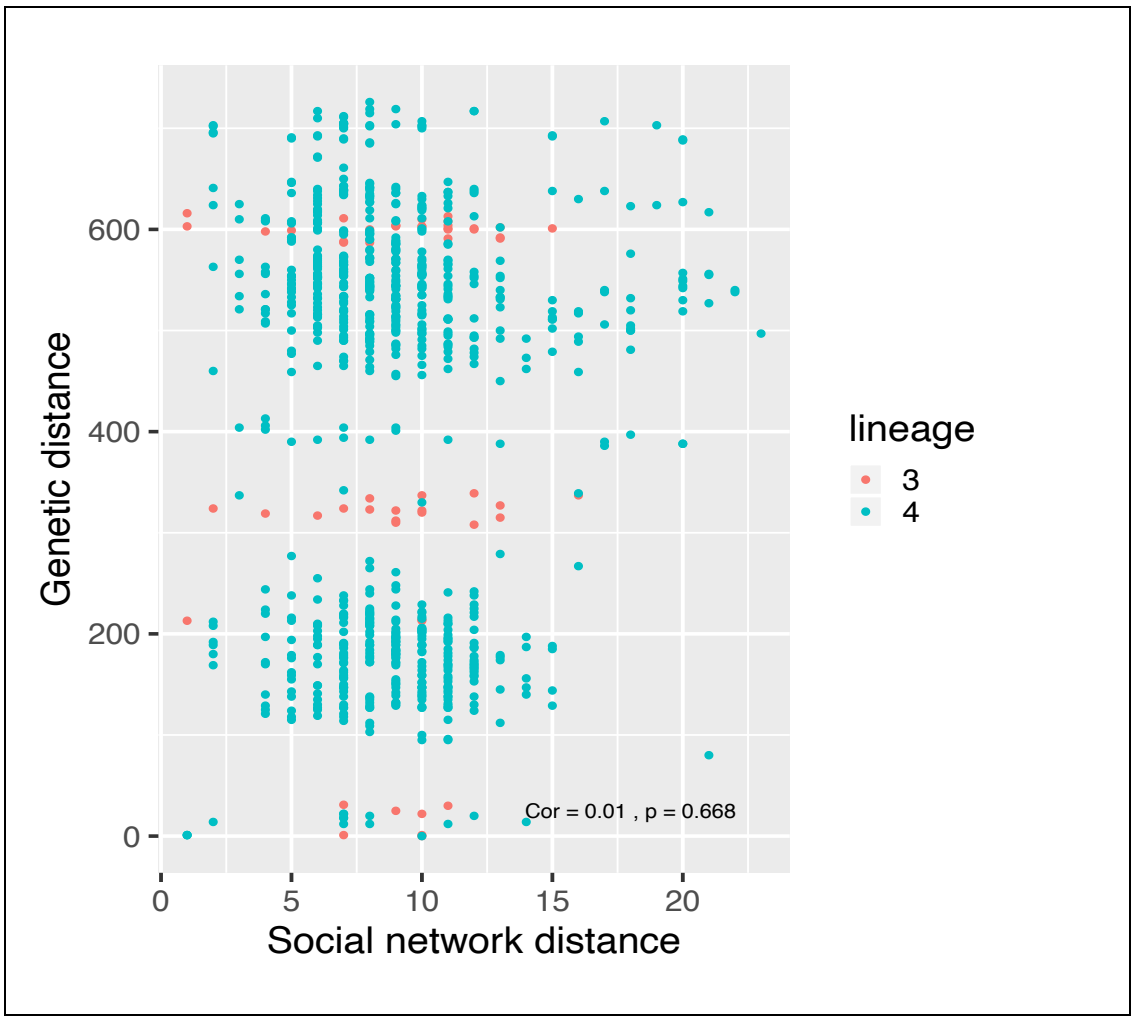
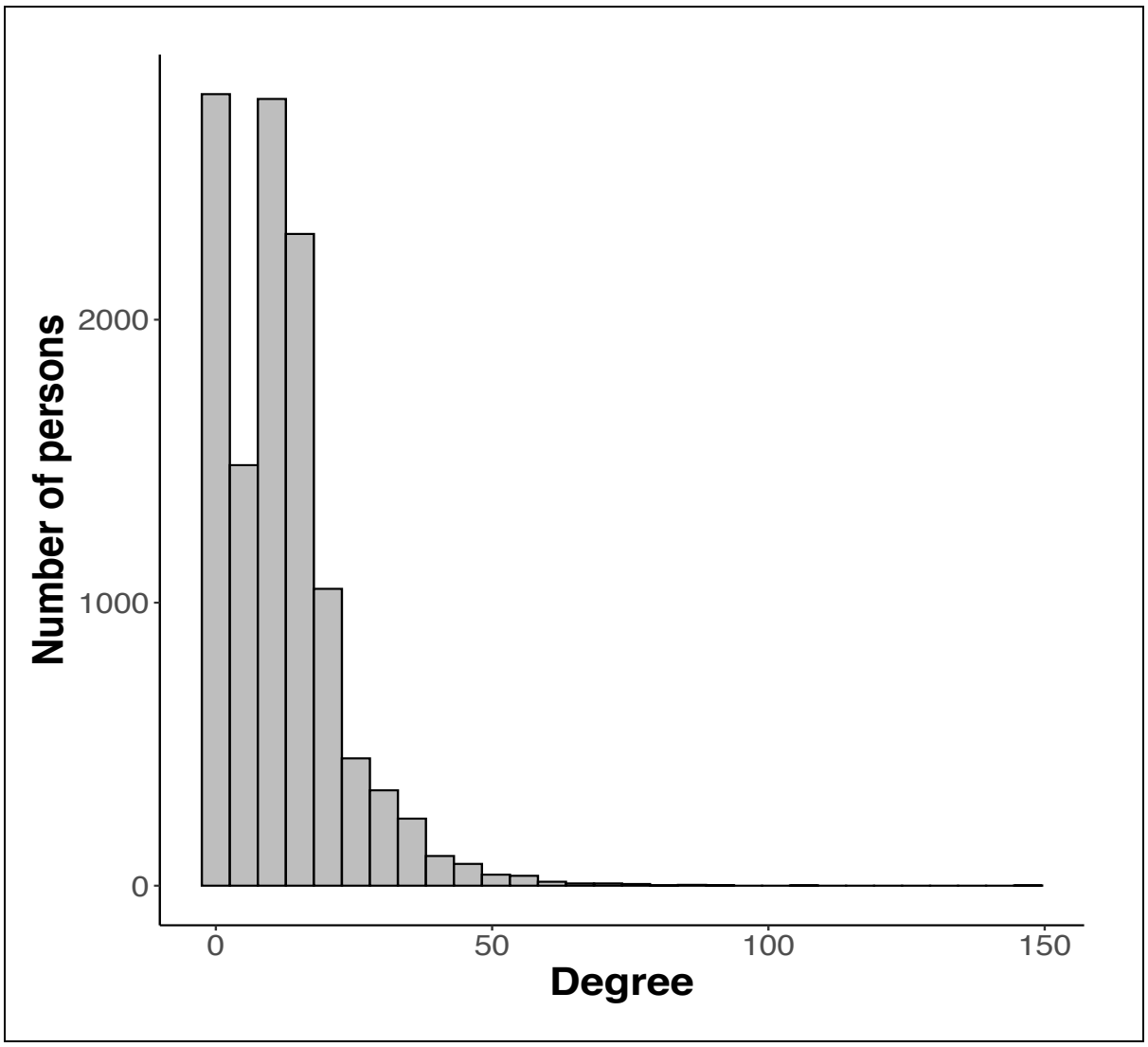
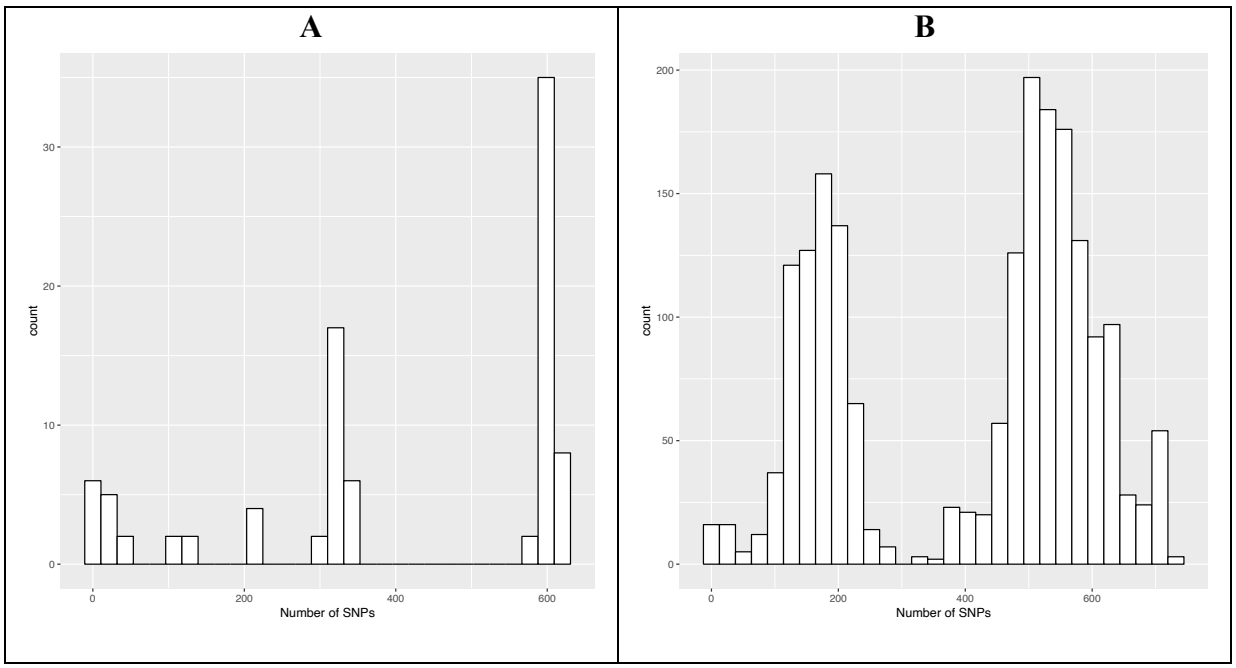


Figure 3.8: Correlation of genetic distance with social network distance

Supplementary tables and figures



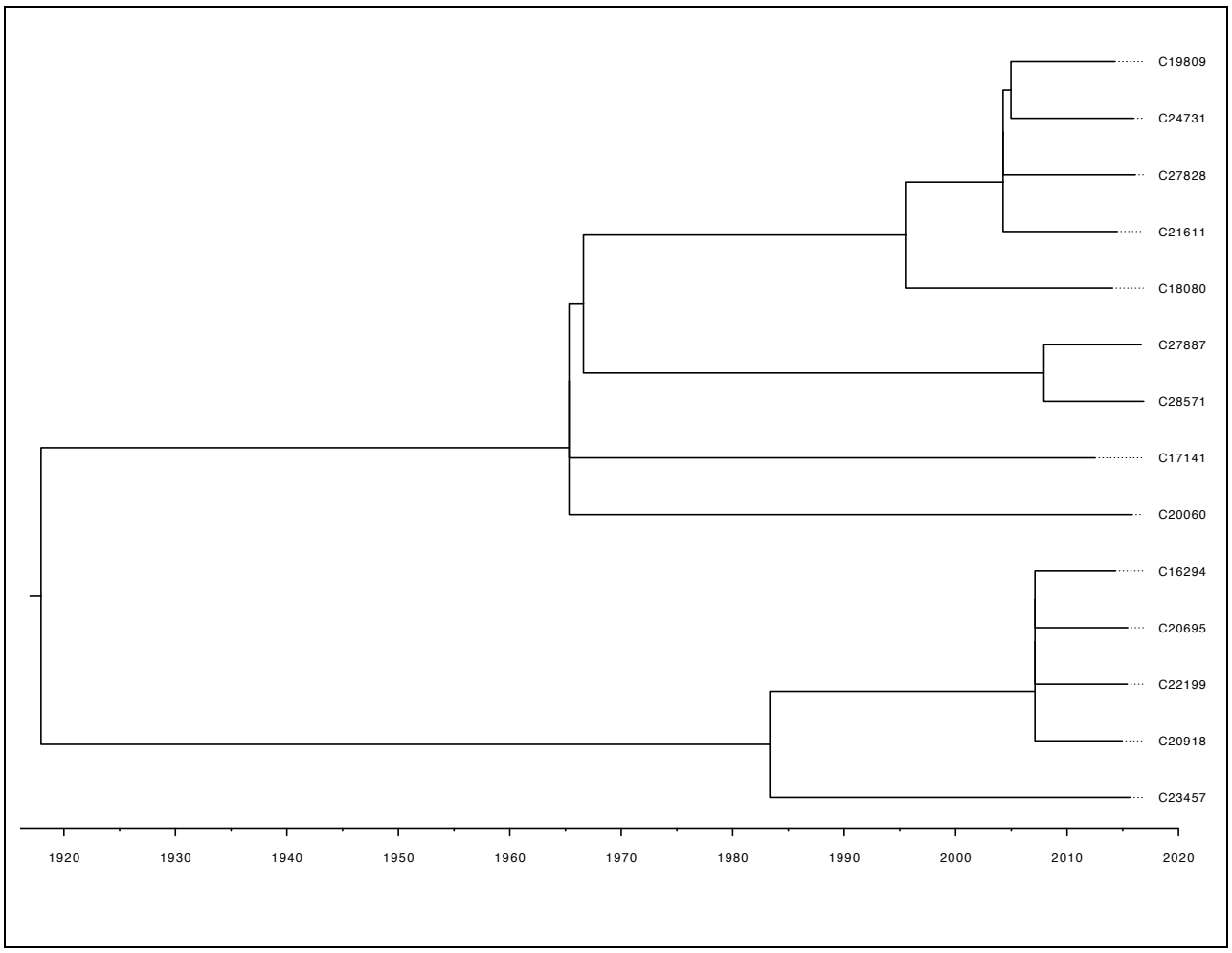
Supplementary figure 3.1: Degree distribution for the aggregated social network



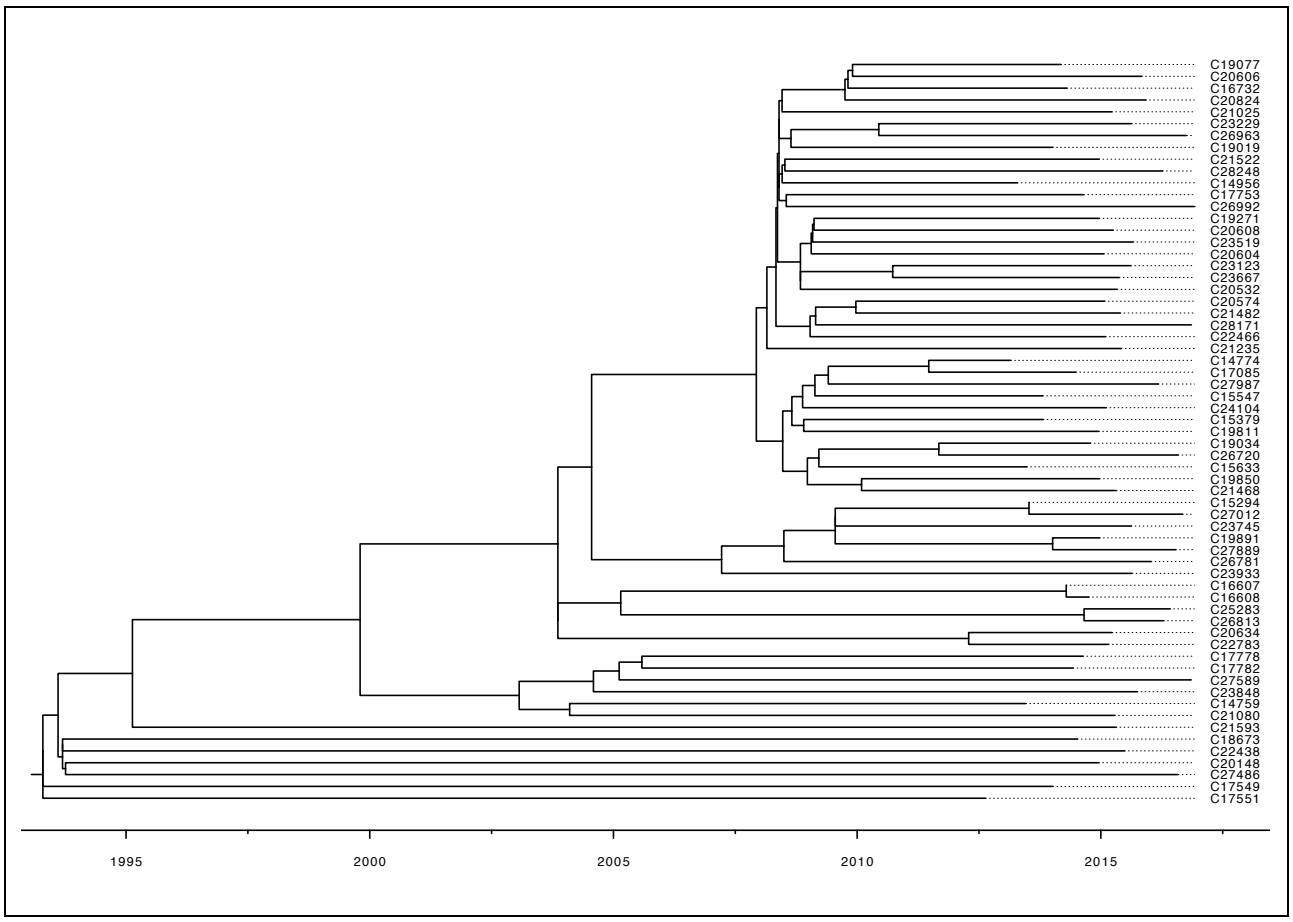
Supplementary figure 3.2: Distribution of SNP differences: lineage 3 (A), lineage 4 (B)

Source	Sink	Probability	**Lineage**	**SNPs**	**Social Network distance**
C23229	C26963	1	4	0	Not connected
C14774	C17085	0.9760048	4	1	1
C17551	C17549	0.88222356	4	0	10
C17549	C17551	0.11777644	4	0	10
C17782	C17778	0.41971606	4	1	1
C17778	C17782	0.26874625	4	1	1
C20695	C22199	1	3	1	Not connected
C20695	C20918	0.7	3	1	7
C20918	C20695	0.29	3	1	7

Supplementary figure 3.3: TransPhylo-inferred genetic links



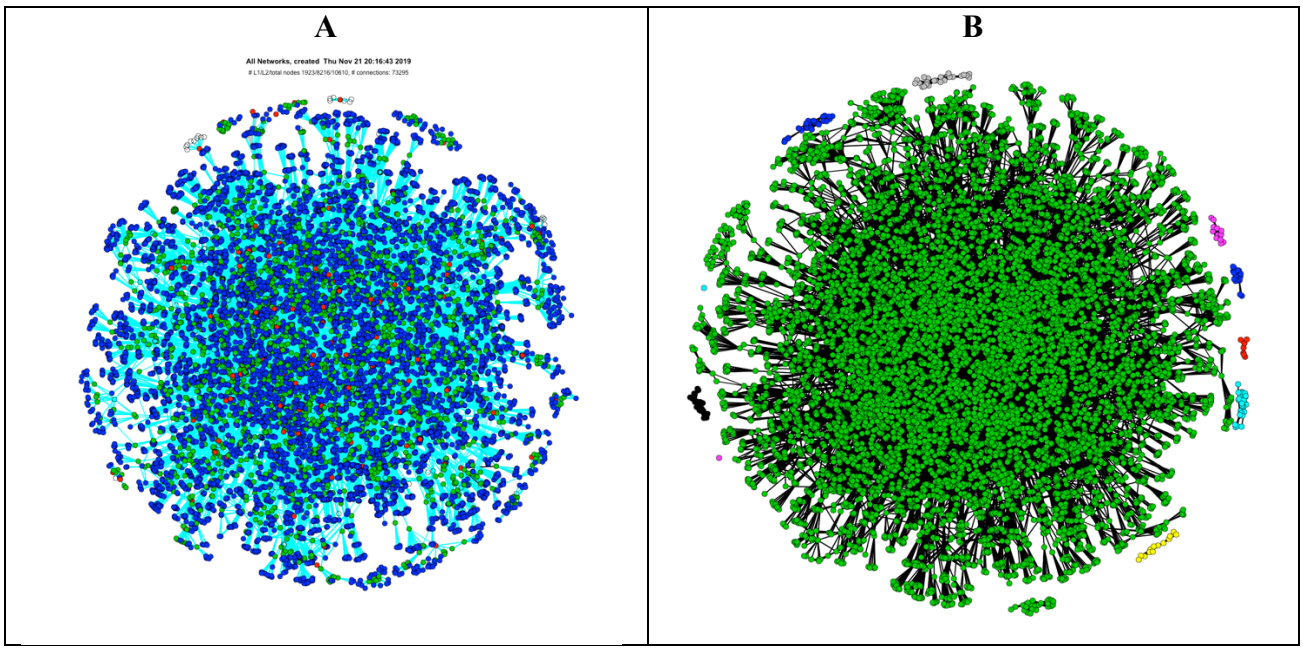
Supplementary figure 3.4: Lineage 3 maximum clade credibility tree



Supplementary figure 3.5: Lineage 4 maximum clade credibility tree

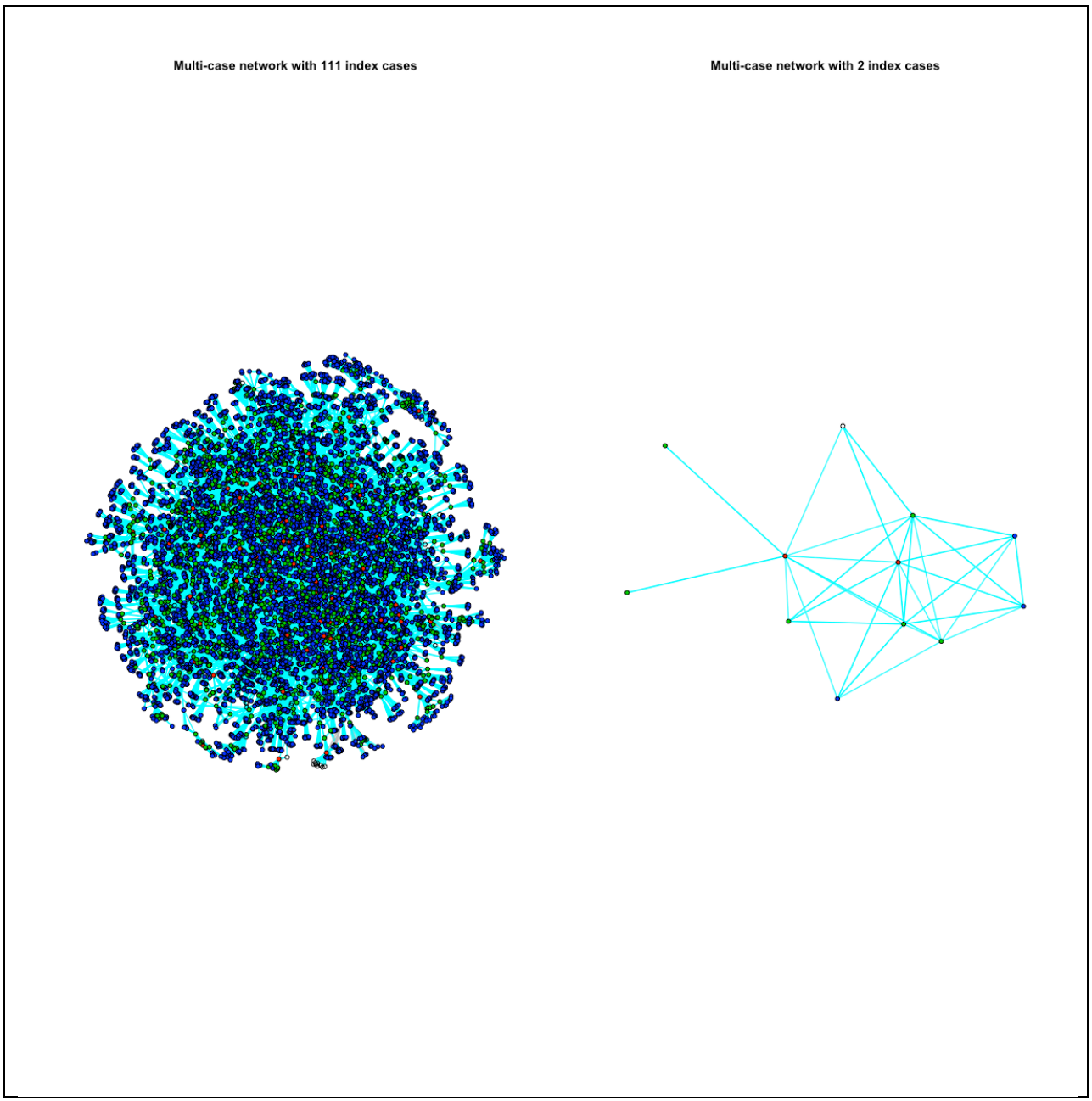
Supplementary table 3.1: Genetic links per SNP threshold among pairs with an identifiable path in the social network

SNP threshold	Median number of SNPs	Number of genetically linked pairs	#Links Lineage 1	#Links Lineage 3	#Links Lineage 4	#Links among pairs at SND=1 (%)	#Links among pairs at SND≤2 (%)	#Links among pairs with a path (%)
5	1	14	1	6	7	3(21)	3(21)	7 (50)
12	1	23	1	6	16	3(13)	3(13)	10 (43)
50	12.5	46	1	13	32	3(7)	4(9)	22 (48)
100	25	75	1	13	61	3(4)	4(5)	27 (36)
SND: Social Network Distance. #: Number of								

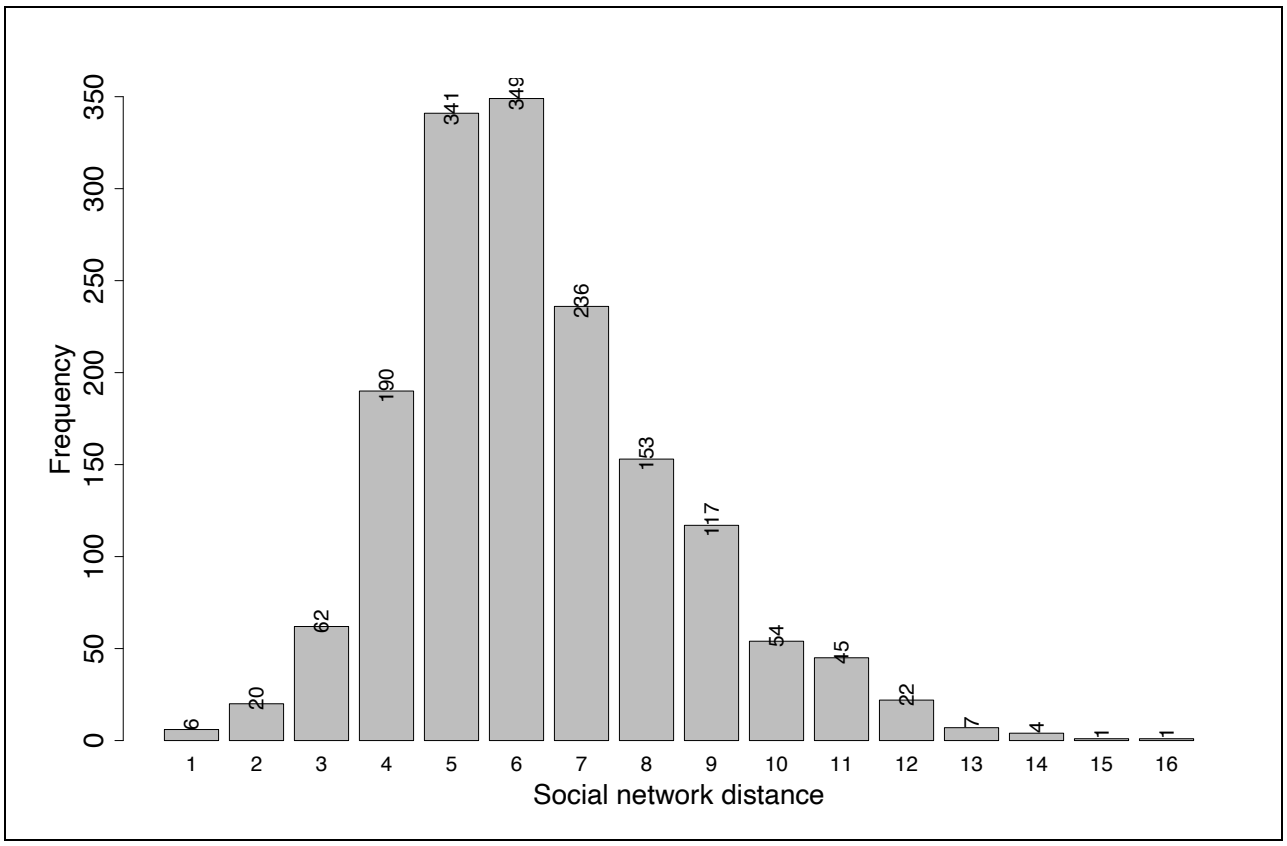


Supplementary figure 3.6: Aggregated social network created and the largest component (Social network built with Fuzzy string matching)

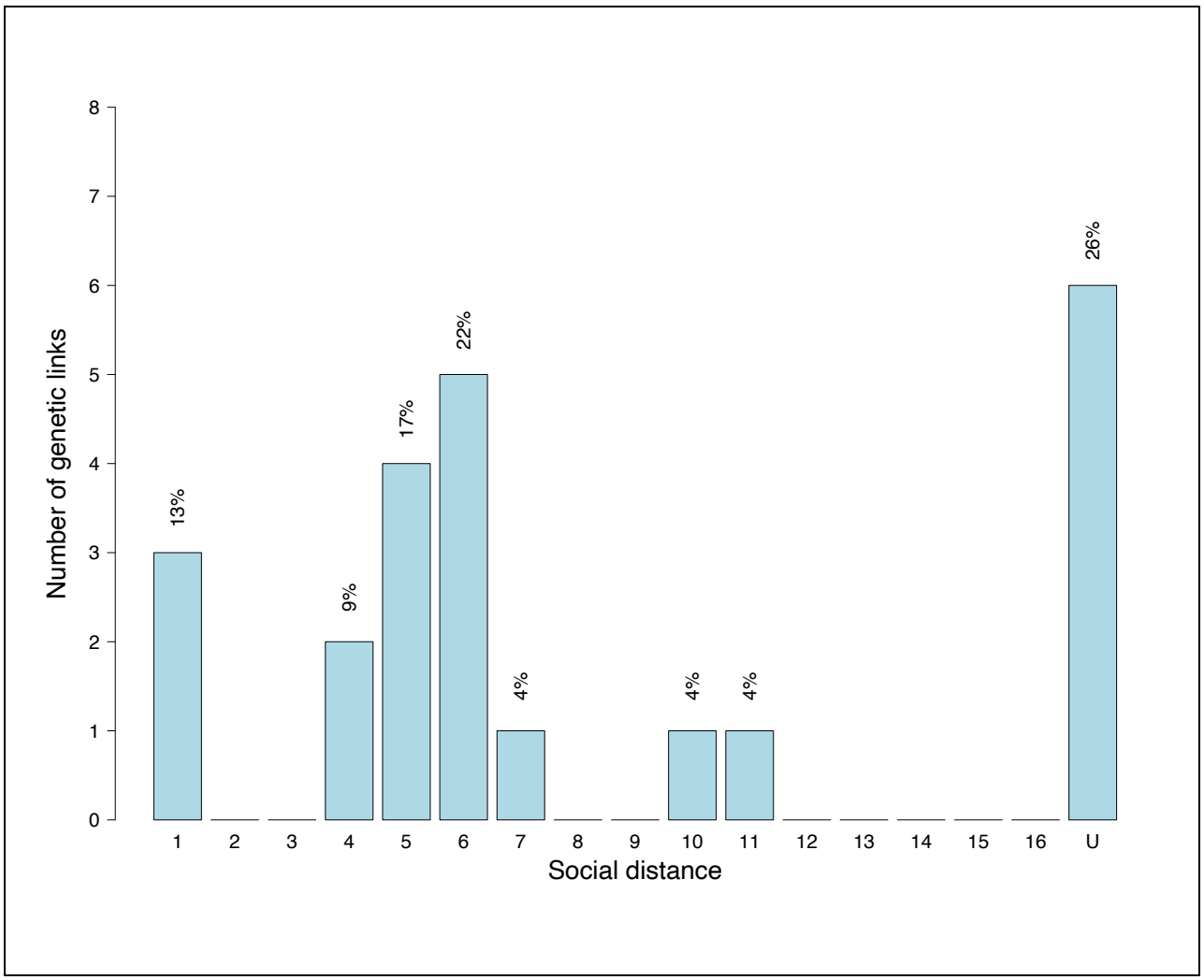
A: Aggregated social network. B: Aggregated social network with colors showing the different component networks that make up the full network. In blue is the biggest component.



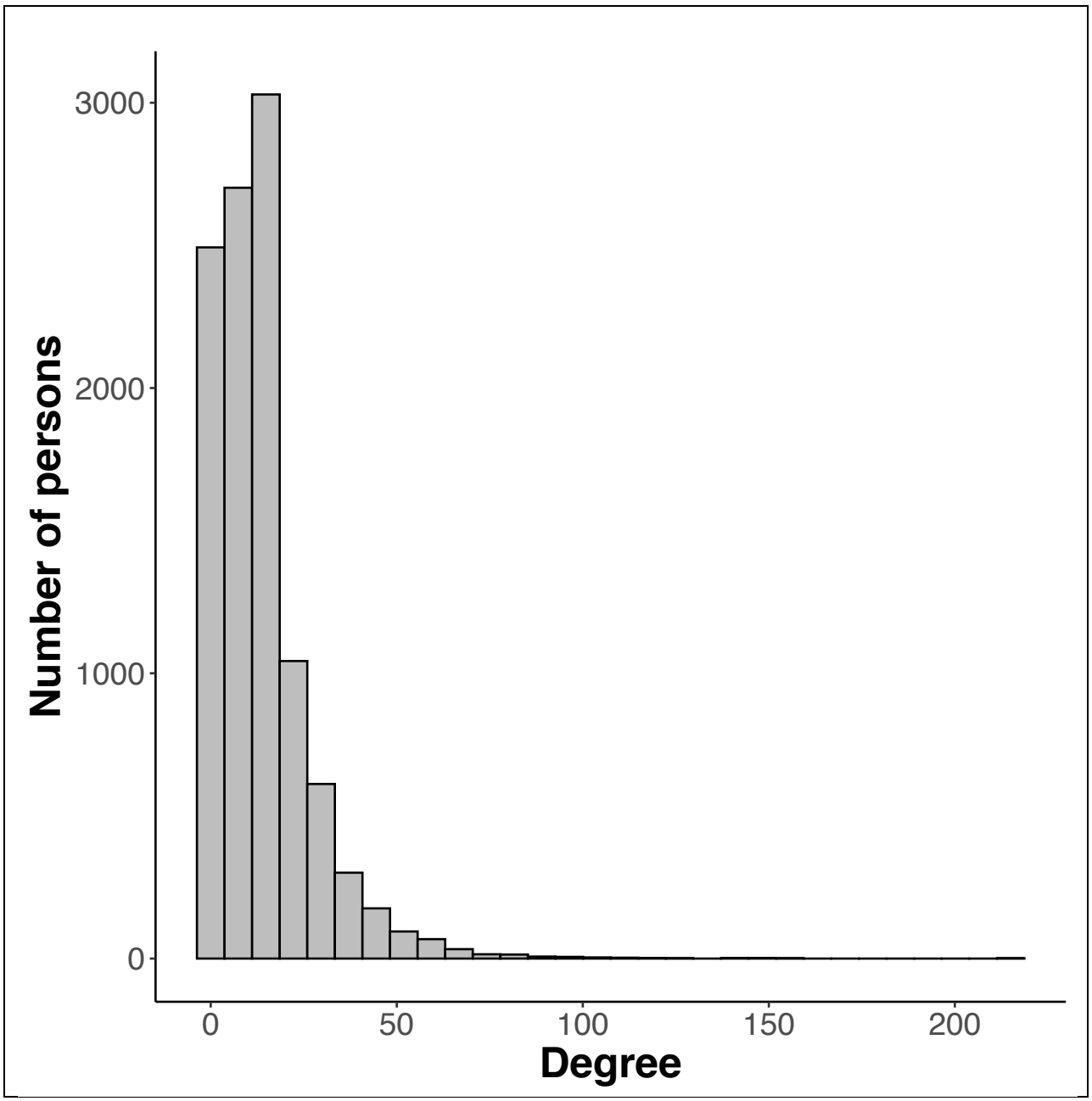
Supplementary figure 3.7: Multi-case networks (Social network built with Fuzzy string matching)



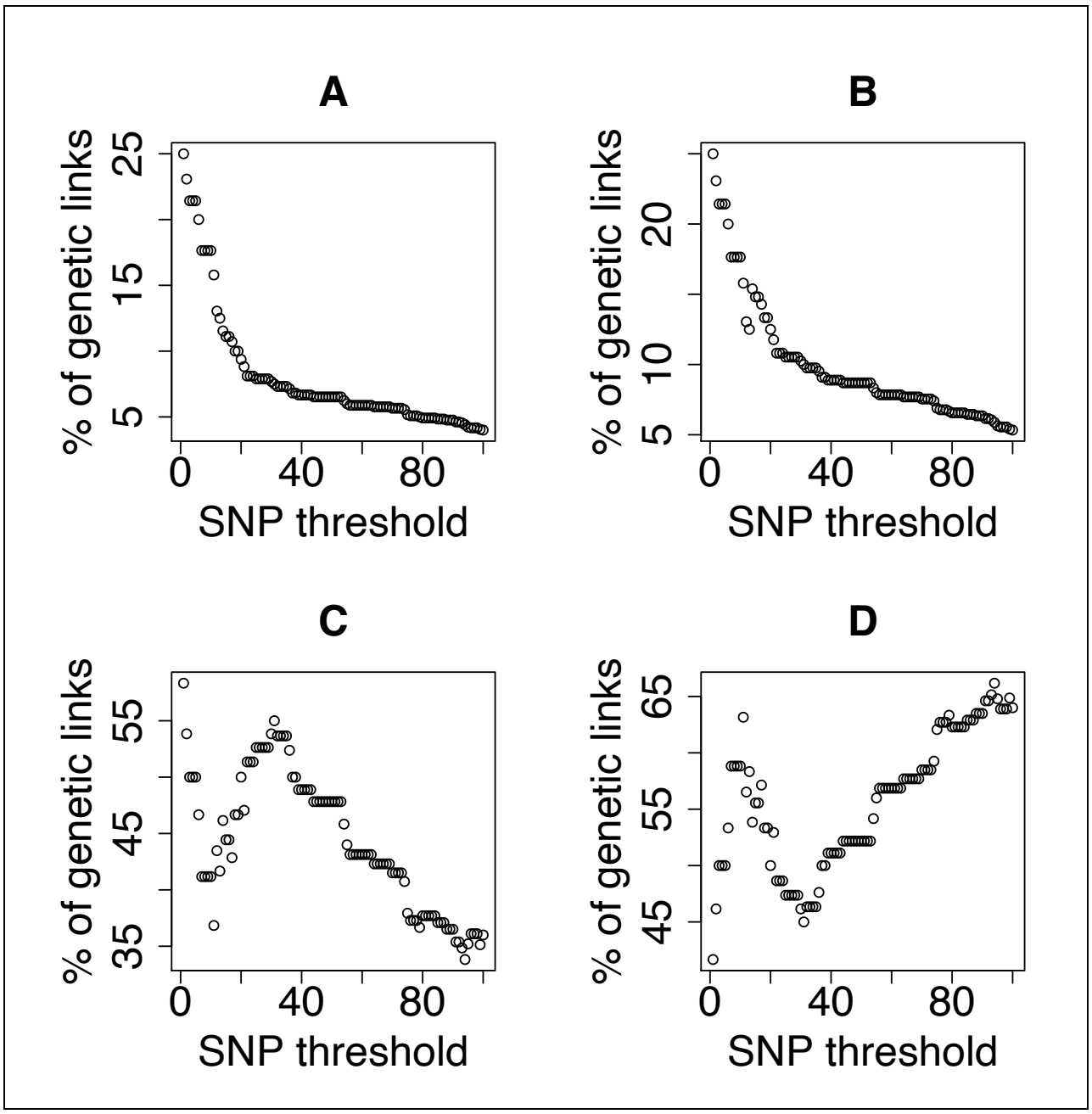
Supplementary figure 3.8: Distribution of pairwise social network distance between tuberculosis patients with an identifiable path between them in the aggregated social network (Social network built with Fuzzy string matching)



Supplementary figure 3.9: Number of genetic links identified per social network distance (Social network built with Fuzzy string matching)

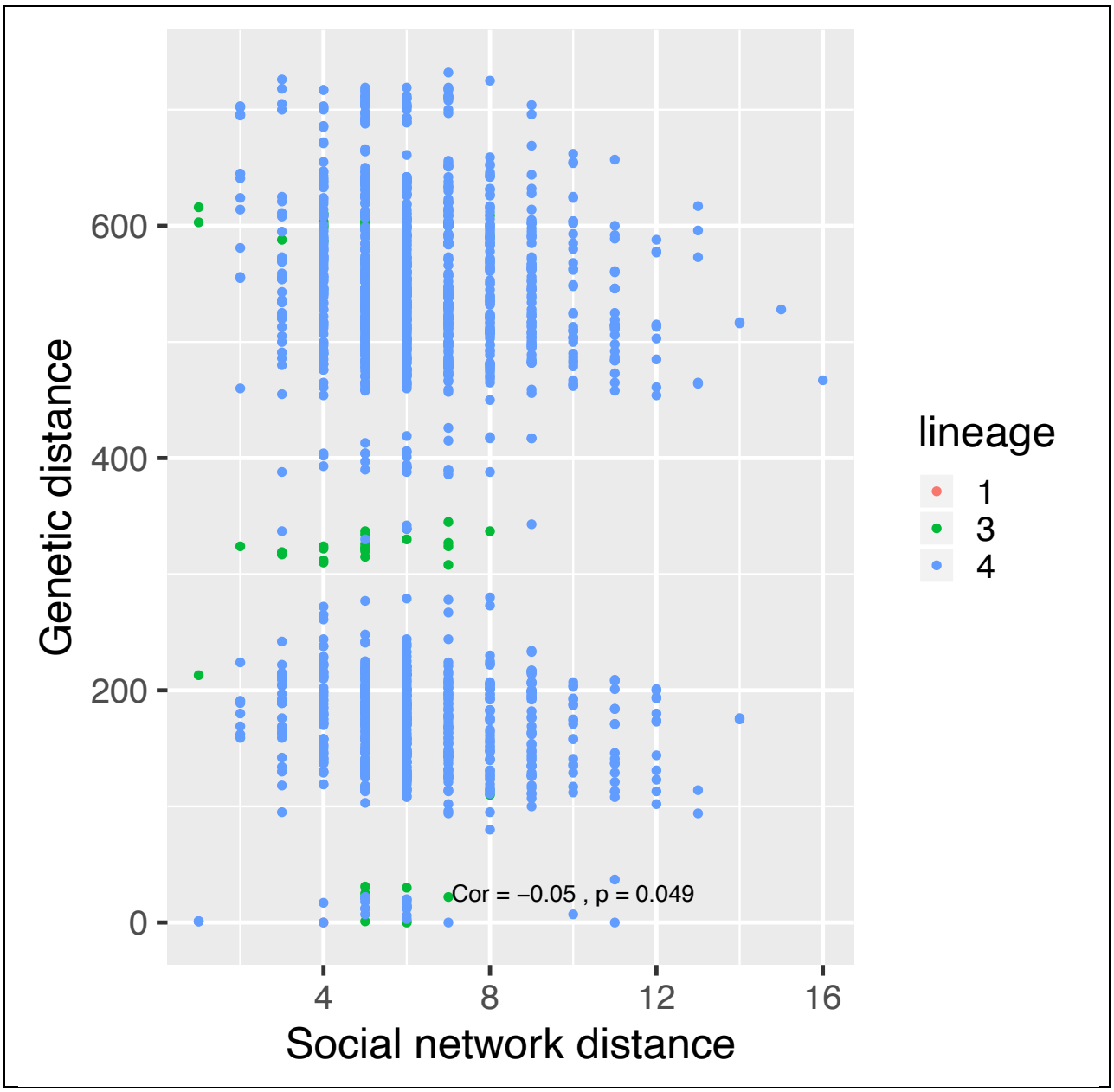


Supplementary figure 3.10: Degree distribution for the aggregated social network (Social network built with Fuzzy string matching)



Supplementary figure 3.11: Percentage of genetic links at different social network distances

Percentage of genetic links between index TB patients (A: social network distance = 1, B: social network distance = 2, C: where an identifiable path exists between pairs in the social network, D: where NO identifiable path exists between pairs in the social network.



Supplementary figure 3.12: Correlation of genetic distance with social network distance (Social network built with Fuzzy string matching)

CHAPTER 4 : WHOLE GENOME SEQUENCING IDENTIFIES CLUSTERS OF RECENT TRANSMISSION AND FACTORS ASSOCIATED WITH RECENT TRANSMISSION IN AN ENDEMIC SETTING IN KAMPALA-UGANDA[3]

[3] R Galiwango, E Yassine, A Handel, J Sekandi, L Liu, R Kakaire, S Zalwango, N Kiwanuka and C Whalen. To be submitted to Nature Scientific Reports

ABSTRACT

Introduction: Uganda is one of the top 30 countries with the highest burden of Tuberculosis (TB) and HIV coinfection. In 2018, the country had an incident rate of 200/100,000 for TB disease which represents a 33.3% decline in incidence from 2000. There is a need to interrupt ongoing transmission if the country is to achieve the targets of elimination spelt out in the End TB strategy. Whole Genome Sequencing (WGS) aids in interrupting transmission by identifying chains of recent transmission as it assumes that cases separated by a few SNP differences are more likely to be part of the same transmission chain.

Methods: We investigated genetic linkage among TB patients in the Community Health and Social Networks of TB (COHSONET) study using a threshold of 12 SNPs to identify clusters of recent transmission, and covariates associated with clustering.

Results: Twenty-nine (36.7%) patients of the 79 sequenced isolates formed 12 clusters. Most (nine) of the clusters were of size 2, one cluster was of size 3 while two were of size 4. In the univariate analysis, clustered patients were more likely to be male and current/past smokers. The multivariate analysis showed that clustered cases were more likely to be current/past smokers.

Conclusion: There is a need for targeted interventions among identified risk groups in order to interrupt transmission.

INTRODUCTION

In 2014, WHO set an ambitious target to end TB by 2035 [7] which has at its core the early detection and treatment of existing cases. While diagnosis and treatment of index cases are essential for the proper management of the individual case, they may not be sufficient to control the epidemic. Like most infectious diseases, tuberculosis creates the next generation of new

cases through transmission before the diagnosis is made and treatment begun in the index case. This transmission may sustain the epidemic in the community by replacing one case with another over time [8]. Therefore, efforts to end TB will depend on our ability to halt ongoing transmission [119].

The drivers of TB transmission differ by setting. This is because, countries (or regions) differ in the burden of prevalent tuberculosis, HIV burden, capacity of healthcare and public health systems to identify and effectively treat individuals with infectious forms of tuberculosis, and the ways in which individuals live, work, and interact i.e., social mixing patterns [1,44].

Uganda is part of the list of top 30 countries with the highest burden of TB and HIV coinfection (WHO Global TB Report, 2019). The incident rate of TB disease was 200/100,000 (95% Confidence Interval= 118/100,000–304/100,000) in 2018 (WHO Global TB Report, 2019). This represents a 33.3% decline in incidence from the 300/100,000 new cases in 2000. There is a need to turn off the tap of new cases of disease by interrupting ongoing transmission if you Uganda is to achieve the targets of elimination that are spelt out in the End TB strategy. A better understanding of risk groups involved in recent transmission chains is required to effectively target interventions.

Previously, traditional genotyping has been used to identify factors associated with recent transmission, using clustering of isolates based on their genotypic profiles as a measure of recent transmission [45]. In this approach, individuals with identical or similar fingerprint patterns over a given time frame usually 2 years are considered to be clustered. Patients whose isolates cluster together are considered to be part of the same recent transmission chains while those with unique (un-clustered) isolates are more likely to be cases of reactivated TB disease that was acquired in

the past. The covariates associated with clustering are determined by comparing the characteristics of clustered and non-clustered TB patients.

Whole Genome Sequencing (WGS) has been shown to separate isolates that had previously been identified as part of the same transmission chain using traditional genotyping techniques leading to smaller distinct clusters and less clustering [46–51]. This is why more recently, WGS has replaced traditional molecular typing as routine in *Mycobacterium tuberculosis* transmission studies. WGS aids in interrupting transmission by identifying chains of recent transmission as it assumes that cases separated by a few SNP differences are more likely to be part of the same transmission chain.

In this study, we used WGS data of pathogen isolates for patients in the Community Health and Social Networks of TB (COHSONET) study to identify tuberculosis patients involved in chains of recent transmission (clusters). We identified factors associated with clustering of tuberculosis patients.

MATERIALS AND METHODS

Study population

The study population consisted 123 index tuberculosis patients from the Community Health and Social Networks of TB (COHSONET) study enrolled between 2012 and 2016. The COHSONET study was a study of index cases and their contact networks. For comparison purposes, the study included a random sample of controls. However, for purposes of this study, we only analyzed the index patients since we were interested in their pathogen isolates, except were we extracted their social network information. The study population (including the eligibility criteria), enrolment procedure, data collection and study measurements, whole genome sequencing of the isolates as

well as the social network study have been described elsewhere (chapter 3). Isolates of 79 of the 123 index participants had successful sequencing.

Data collection

Data used in this study from the COHSONET study were: demographics (age, patient's identified sex, education level, income), HIV coinfection, social risk factors (alcohol use, smoking), clinical factors (sputum smear status, dates of cough onset, BCG vaccination status, previous TB diagnosis), lineage and degree of each index participant in the aggregated COHSONET social network.

Ethical considerations and data availability

The COHSONET study was approved by both the Institutional Review Board at Makerere University and one at the University of Georgia. Whole genome sequence data is available upon request to the corresponding author.

Definition of a clustered case

A clustered case was defined as any case whose isolate was within 12 SNPs of at least one other case's isolate during the study period [21,94,103]. A non-clustered case was defined as any TB case from the study population whose isolate was >12 SNPs from any other case's isolate.

We performed a sensitivity analysis with a more stringent threshold of 5 SNPs to define a clustered tuberculosis case.

Data analysis

We calculated the proportion of clustered cases from the number of cases with at least one genetic link with another case divided by the total number of cases. We compared characteristics of patients whose isolates were sequenced with those whose isolates were not sequenced using chi-square tests for categorical variables (or Fisher exact test where necessary),

and a t-test for continuous variables to make sure that there was no bias in sampling of isolates for sequencing. We performed an item analysis on each of the collected variables and excluded variables with a lot of missing data, variables that were highly correlated with other variables and those whose distributions were not appreciable. This provided a subset of variables that was used in subsequent analysis.

We performed univariate logistic regression to identify individual covariates associated with clustering and multivariate logistic regression including age, as a potential confounder along with covariates associated with clustering in univariate analysis. Considering that logistic regression tends to overestimate the measure of effect [120], we performed a sensitivity analysis using Modified Poisson Regression.

The outcome was clustering (clustered vs un-clustered). All explanatory variables relating to the characteristics of each index patient were assessed for their relationship with clustering at univariate level. These were: social network characteristics (degree, betweenness and closeness of each index participant in the social network), sex, education level, income, age in years, HIV status, previous history of TB, Body Mass Index (BMI), cough duration in months, BCG, reported contact with a person known to have TB, smoking and alcohol use. Covariates were included in the multivariate model if the p-value for the univariate odds ratio (OR) was ≤0.2. We assessed possible interactions between the covariates i.e., alcohol with sex, alcohol with smoking, sex with smoking, alcohol with HIV, alcohol with education and smoking with education.

RESULTS

Description of the study population

The study enrolled 123 index tuberculosis patients. Eleven of the patients were identified during field evaluation of index contacts. There was no statistically significant difference between patients whose isolates were sequenced and those whose isolates were not sequenced isolates (Table 2.1). Two of the sequenced isolates were lineage 2, fourteen were of lineage 3 while sixty-three belonged to lineage 4. 68.35% of them were male while the rest were female (Table 4.1). 15.19% were HIV positive. Their median age was 29 years.

Identified clusters

Twenty-nine tuberculosis patients (36.7%) were clustered (Figure 4.1A). The 29 patients formed 12 clusters (Supplementary table 4.7). Most (nine) of the clusters were of size 2 (Figure 4.1B). One cluster was of size 3 while two were of size 4.

Characteristics of clustered patients

82.75% of the clustered patients were male, 55.17% reported alcohol use (table 1). They had a median age of 27 years (range=20, 49) and a median BMI of 19.13kg/m^2 (range=13.55, 23.67), a median cough duration of 2 months (range=0.46, 24) and a median social network degree of 10 contacts (range=23, 56).

Factors associated with clustering

In the univariate analysis, clustered patients were more likely to be male (Odds Ratio=3.20, 95% Confidence Interval=1.11, 10.75; $p<0.05$) and were more likely to be current or past smokers (Odds Ratio =9.14, 95% Confidence Interval =2.08, 64.12; $p<0.01$) (Table 4.1). The odds of clustering increased with decrease in BMI (Odds Ratio =0.86, 95% Confidence Interval=0.72, 0.99; $p=0.05$). There was no association between clustering and social network

centrality characteristics (degree, betweenness and closeness). No interaction between the covariates was statistically significant (i.e., each had $p>0.2$) thus no interaction terms were added to the multivariate model.

Sex, HIV status, smoking, alcohol use, known contact with a TB patient and BMI had p-value ≤ 0.2 in the univariate analysis and these variables together with age (a potential confounder) were included in the multivariate analysis. The multivariate analysis showed that clustered cases were more likely to be current or past smokers (Adjusted Odds Ratio=9.14, 95% Confidence Interval=2.08, 64.12; $p<0.01$) (Table 4.2).

The results of the Modified Poisson regression analysis were similar to those of the logistic regression analysis (Supplementary Table 4.1, Supplementary Table 4.2)

The results were generally robust to a change in the definition of a clustered TB patient using a more stringent threshold of 5 SNPs (Supplementary Table 4.3, Supplementary Table 4.4, Supplementary Table 4.5, Supplementary Table 4.6) even though the number of clustered TB patients and size of clusters reduced (Supplementary Figure 4.1, Supplementary Table 4.8).

DISCUSSION

In this study, use of whole genome sequencing enabled us to identify clusters of recent tuberculosis transmission and covariates associated with clustering with a high degree of accuracy. We found that clustered patients were more likely to be past or current smokers. This study adds to the growing literature on the increased risk of acquiring tuberculosis by current smokers or persons who have ever smoked [121–123]. Our study has illustrated the association between tuberculosis and smoking using whole genome sequence data. The results were generally robust to a change in the definition of a clustered TB patient from 12 SNPs to a more stringent threshold of 5 SNPs.

Smoke particles have been shown to impair macrophages, which are critical immune cells in fighting mycobacterium tuberculosis [124]. This may explain why current and past smokers were associated with being involved in a recent transmission event.

To test if cigarette smoking is a marker of some cultural behavior, we tested for associations between smoking and other patient characteristics using pairwise logistic regression models. We found associations with education level, age, cough duration and past contact with a person known to have TB. Smoking was neither associated with alcohol use nor patient reported sex.

The Community Health and Social Networks of TB (COHSONET) study is the largest social network study of tuberculosis in Africa to be reported. Unlike most social network studies that are egocentric in nature, meaning the index tuberculosis patient is asked to list their contacts who are normally not enrolled into the study, the COHSONET study used an extended form of egocentric sampling were in addition to what is done in classic egocentric network sampling, an additional layer of contacts (the contacts of contacts) was added. The study also included a sample of index controls, their first level contacts and second level contacts. Indexes (cases and controls) and first level contacts were asked to describe social relations between the first level contacts and second level contacts, respectively. This comprehensive social network approach provides a better representation of the social network.

A limitation of the study is that we did not enroll all consecutive tuberculosis patients during the study period and not all isolates were sequenced. It is therefore possible that we underestimated the proportion of clustered patients. However, there was no statistically significant difference between characteristics of patients whose isolates were sequenced and those whose isolates were not sequenced.

In conclusion, targeting high risk groups such as smokers for interventions could help interrupt ongoing transmission.

Data availability

Whole Genome Sequence data is available upon request to the corresponding author.

REFERENCES

1. Alavi-Naini, R., Sharifi-Mood, B., & Metanat, M. (2012). Association Between Tuberculosis and Smoking. *International Journal of High-Risk Behaviors and Addiction*, *1*(2), 71–74. https://doi.org/10.5812/ijhrba.5215
2. Auld, S. C., Kasmar, A. G., Dowdy, D. W., Mathema, B., Gandhi, N. R., Churchyard, G. J., … Shah, N. S. (2017). Research Roadmap for Tuberculosis Transmission Science: Where Do We Go from Here and How Will We Know When We're There? *Journal of Infectious Diseases*, *216*(January), S662–S668. https://doi.org/10.1093/infdis/jix353
3. Bates, M. N. (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
4. Berg, R. D., Levitte, S., O'Sullivan, M. P., O'Leary, S. M., Cambier, C. J., Cameron, J., … Ramakrishnan, L. (2016). Lysosomal Disorders Drive Susceptibility to Tuberculosis by Compromising Macrophage Migration. *Cell*, *165*(1), 139–152. https://doi.org/10.1016/j.cell.2016.02.034
5. Churchyard, G., Kim, P., Shah, N. S., Rustomjee, R., Gandhi, N., Mathema, B., … Cardenas, V. (2017). What We Know about Tuberculosis Transmission: An Overview. *Journal of Infectious Diseases*, *216*(August), S629–S635. https://doi.org/10.1093/infdis/jix362
6. Cohen, K. A., Abeel, T., McGuire, A. M., Desjardins, C. A., Munsamy, V., Shea, T. P.,

… Earl, A. M. (2015). Evolution of Extensively Drug-Resistant Tuberculosis over Four Decades: Whole Genome Sequencing and Dating Analysis of Mycobacterium tuberculosis Isolates from KwaZulu-Natal. *PLOS MEDICINE*, *12*(9). https://doi.org/10.1371/journal.pmed.1001880

7. Den Boon, S., Van Lill, S. W. P., Borgdorff, M. W., Verver, S., Bateman, E. D., Lombard, C. J., … Beyers, N. (2005). Association between smoking and tuberculosis infection: A population survey in a high tuberculosis incidence area. *Thorax*, *60*(7), 555–557. https://doi.org/10.1136/thx.2004.030924
8. Fok, A., Numata, Y., Schulzer, M., & Fitzgerald, M. J. (2008). Risk factors for clustering of tuberculosis cases: a systematic review of population-based molecular epidemiology studies. *International Journal of Tuberculosis and Lung Disease*, *12*(March 2007), 480–492.
9. Gurjav, U., Outhred, A. C., Jelfs, P., Mccallum, N., Wang, Q., Hill-Cawthorne, G. A., … Sintchenko, V. (2016). Whole Genome Sequencing Demonstrates Limited Transmission within Identified Mycobacterium tuberculosis Clusters in New South Wales, Australia. *Australia. PLoS ONE*, *11*(10). https://doi.org/10.1371/journal.pone.0163612
10. Jajou, R., de Neeling, A., van Hunen, R., de Vries, G., Schimmel, H., Mulder, A., … van Soolingen, D. (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). https://doi.org/10.1371/journal.pone.0195413
11. Jamieson, F. B., Teatero, S., Guthrie, J. L., Neemuchwala, A., Fittipaldi, N., & Mehaffy, C. (2014). Whole-genome sequencing of the Mycobacterium tuberculosis Manila sublineage results in less clustering and better resolution than mycobacterial interspersed

repetitive-unit-variable-number tandem-repeat (MIRU-VNTR) typing and spoligotyping. *Journal of Clinical Microbiology*, *52*(10), 3795–3798. https://doi.org/10.1128/JCM.01726-14

12. Mathema, B., Andrews, J. R., Cohen, T., Borgdorff, M. W., Behr, M., Glynn, J. R., … Wood, R. (2017). Drivers of Tuberculosis Transmission. *Journal of Infectious Diseases*, *216*(April), S644–S653. https://doi.org/10.1093/infdis/jix354
13. Nelson, K. N., Shah, N. S., Mathema, B., Ismail, N., Brust, J. C. M., Brown, T. S., … Gandhi, N. R. (2018). Spatial Patterns of Extensively Drug-Resistant Tuberculosis Transmission in KwaZulu-Natal, South Africa. *JOURNAL OF CLINICAL MICROBIOLOGY*, *30322*. https://doi.org/10.1093/infdis/jiy394
14. Nikolayevskyy, V., Kranzer, K., Niemann, S., & Drobniewski, F. (2016). Whole genome sequencing of Mycobacterium tuberculosis for detection of recent transmission and tracing outbreaks: A systematic review. *Tuberculosis*, *98*, 77–85. https://doi.org/10.1016/j.tube.2016.02.009
15. Roetzer, A., Diel, R., Kohl, T. A., Rueckert, C., Nuebel, 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
16. Stucki, D., Ballif, M., Egger, M., Furrer, H., Altpeter, E., Battegay, M., … Fenner, L. (2016). Standard Genotyping Overestimates Transmission of Mycobacterium tuberculosis among Immigrants in a Low-Incidence Country. *Journal of Clinical Microbiology*, *54*(7), 1862–1870. https://doi.org/10.1128/JCM.00126-16

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

18. Whalen, C. C. (2016). The replacement principle of tuberculosis. *American Journal of Respiratory and Critical Care Medicine*, *194*(4), 400–401. https://doi.org/10.1164/rccm.201603-0439ED

19. WHO. (2014). *The End TB Strategy*.

20. Wyllie, D. H., Davidson, J. A., Smith, E. G., Rathod, P., Crook, D. W., Peto, T. E. A., … Campbell, C. (2018). A Quantitative Evaluation of MIRU-VNTR Typing Against Whole-Genome Sequencing for Identifying Mycobacterium tuberculosis Transmission: A Prospective Observational Cohort Study. *EBioMedicine*. https://doi.org/10.1016/j.ebiom.2018.07.019

21. Yang, C., Lu, L., Warren, J. L., Wu, J., Jiang, Q., Zuo, T., … Cohen, T. (2018). Internal migration and transmission dynamics of tuberculosis in Shanghai, China: an epidemiological, spatial, genomic analysis. *LANCET INFECTIOUS DISEASES*, *18*(7), 788–795. https://doi.org/10.1016/S1473-3099(18)30218-4

22. Yang, C., Luo, T., Shen, X., Wu, J., Gan, M., Xu, P., … Gao, Q. (2017). Transmission of multidrug-resistant Mycobacterium tuberculosis in Shanghai, China: a retrospective observational study using whole-genome sequencing and epidemiological investigation. *www.thelancet.com/infection*, *17*. https://doi.org/10.1016/S1473-3099(16)30418-2

Acknowledgements

My doctoral training was supported by the Fogarty International Center of the National Institutes of Health (Award Number D43TW010045). The COHSONET study was supported by the National Institute of Allergy and Infectious Diseases of the National Institutes of Health under Award Number R01AI093856.

TABLES AND FIGURES

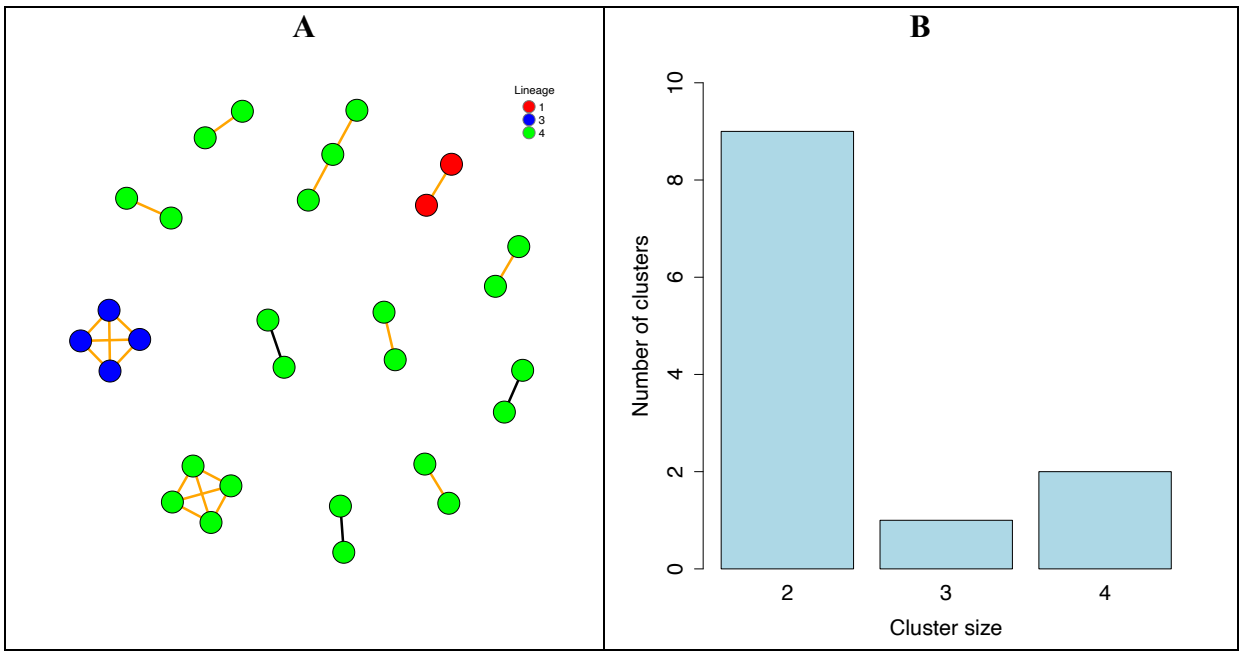


Figure 4.1: A: Identified clusters. B: Number of clusters identified for each cluster size

Black edges represent links for patients who had a social link while orange edges represent links for patients who had no social link.

Table 4.1: Factors associated with clustering in the univariate logistic regression analysis (SNP threshold=12)

Variable	Total number of cases (N=79) (%)	No (%) clustered (n=29); 36.7%	OR (95% CI)
Sex			
Male	54 (68.35)	24 (44.44)	3.20 (1.11, 10.75)
Female	25 (31.65)	5 (20.00)	1
HIV status			
Positive	12 (15.19)	2 (16.67)	1
Negative	66 (83.54)	27 (40.90)	3.46 (0.83, 23.70)
Missing	1 (1.27)		
Monthly income			
<200,000 UGSHS	23 (29.11)	6 (26.08)	0.51 (0.16, 1.43)
≥200,000 UGSHS	56 (70.89)	23 (0.41)	
Alcohol use			
Yes	34 (43.04)	16 (47.06)	2.19 (0.87, 5.65)
No	45 (56.96)	13 (28.89)	1
Missing			
Smoking			
Past/current smoker	10 (12.66)	8 (80.00)	9.14 (2.08, 64.12)

Never smoked	69 (87.34)	21 (30.43)	1
Education Below high school At least high school	 47 (59.50) 32 (40.50)	 18 (38.30) 11 (34.38)	 1.18 (0.47, 3.08) 1
Previous TB Yes No	 15 (18.99) 64 (81.01)	 6 (40.00) 23 (35.94)	 1 0.84 (0.27, 2.79)
BCG scar Present Absent/uncertain	 21 (26.58) 58 (73.42)	 20 (95.24) 9 (15.52)	 0.70 (0.25, 1.98) 1
Ever had contact with a person known to have TB Yes No Missing	 53 (67.09) 22 (27.85) 4 (5.06)	 11 (20.75) 18 (81.82)	 1.94 (0.71, 5.41)
Median age in years (range)	29 (17-59)	27(20, 47)	1.00 (0.95, 1.05)
Median BMI in kg/m²(range)	32.57 (23.41, 52.08)	19.13 (13.55, 23.67)	0.86 (0.72, 0.99)
Median cough duration in months (range)	2 (0.46,24)	2 (0.46, 24)	1.00 (0.88, 1.12)
Median social network degree	23 (10, 87)	10 (23, 56)	1.00 (0.96, 1.03)
Closeness centrality			1.00 (0.999, 1.00)
Betweenness centrality			1.00 (1.00, 1.00)

Table 4.2: Factors associated with clustering in the multivariate logistic regression analysis (SNP threshold=12)

Variable	Adjusted OR (95% CI)
Sex Male Female	 1.44 (0.36, 6.03) 1
HIV status Positive Negative	 1.41 (0.25, 11.37) 1
Smoking Past/current smoker Never smoked	 18.11(1.90, 459.62) 1
Alcohol use Yes No	 1.87 (0.49, 7.37) 1

Ever had contact with a person known to have TB Yes No	 1.02 (0.26, 3.66) 1
Median age in years (range)	0.94 (0.86, 1.01)
Median BMI in kg/m^2(range)	0.89 (0.73, 1.04)

Supplementary figures

Supplementary table 4.1: Factors associated with clustering in the univariate Modified Poisson analysis (SNP threshold=12)

Variable	Total number of cases (N=79) (%)	No (%) clustered (n=29); 36.7%	PR (95% CI)
Sex Male Female	 54 (68.35) 25 (31.65)	 24 (44.44) 5 (20.00)	 2.22 (0.96, 5.14) 1
HIV status Positive Negative Missing	 12 (15.19) 66 (83.54) 1 (1.27)	 2 (16.67) 27 (40.90)	 1 2.46 (0.67, 8.99)
Monthly income <200,000 UGSHS ≥200,000 UGSHS	 23 (29.11) 56 (70.89)	 6 (26.10) 23 (41.10)	 0.64 (0.30, 1.35)
Alcohol use Yes No Missing	 34 (43.04) 45 (56.96)	 16 (47.06) 13 (28.89)	 1.63 (0.91, 2.91) 1
Smoking Past/current smoker Never smoked	 10 (12.66) 69 (87.34)	 8 (80.00) 21 (30.43)	 2.63 (1.64, 4.22) 1
Education Below high school At least high school	 47 (59.50) 32 (40.50)	 18 (38.30) 11 (34.38)	 1.11 (0.61, 2.03) 1
Previous TB Yes No	 15 (18.99) 64 (81.01)	 6 (40.00) 23 (35.94)	 1 1.14 (0.60, 2.17)
BCG scar Present Absent/uncertain	 21 (26.58) 58 (73.42)	 20 (95.24) 9 (15.52)	 0.81 (0.44, 1.48) 1

Ever had contact with a person known to have TB Yes No Missing	 53 (67.09) 22 (27.85) 4 (5.06)	 11 (20.75) 18 (81.82)	 1.47 (0.84, 2.58)
Median age in years (range)	29 (17-59)	27(20, 47)	1.00 (0.97, 1.03)
Median BMI in kg/m^2(range)	32.57 (23.41, 52.08)	19.13 (13.55, 23.67)	0.91 (0.84, 0.99)
Median cough duration in months (range)	2 (0.46,24)	2 (0.46, 24)	1.00 (0.93, 1.08)
Median social network degree	23 (10, 87)	23 (10, 56)	1.00 (0.98, 1.02)
Closeness centrality			1.00 (1.00, 1.00)
Betweenness centrality			1.00 (1.00, 1.00)

Supplementary table 4.2: Factors associated with clustering in the multivariate Modified Poisson analysis (SNP threshold=12)

Variable	Adjusted PR (95% CI)
Sex Male Female	 1.54 (0.65, 3.65) 1
Smoking Past/current smoker Never smoked	 **2.83 (1.26, 6.34)** 1
Alcohol use Yes No	 1.45 (0.73, 2.90) 1
Median age in years (range)	0.96 (0.92, 1.00)
Median BMI in kg/m^2(range)	0.93 (0.86, 1.01)

Supplementary table 4.3: Factors associated with clustering in the univariate logistic regression analysis (SNP threshold=5)

Variable	Total number of cases (N=79) (%)	No (%) clustered (n=20); 25.3%	OR (95% CI)
Sex Male Female	 54 (68.35) 25 (31.65)	 17 (31.50) 3 (12.00)	 3.37 (0.99, 15.60) 1
HIV status Positive Negative Missing	 12 (15.19) 66 (83.54) 1 (1.27)	 2 (16.70) 18 (27.30)	 1 1.88 (0.44, 12.97)

Monthly income <200,000 UGSHS ≥200,000 UGSHS	 23 (29.11) 56 (70.89)	 5 (21.70) 15 (26.80)	 0.76 (0.22, 2.30)
Alcohol use Yes No Missing	 34 (43.04) 45 (56.96)	 11 (32.40) 9 (20.00)	 1.91 (0.69, 5.45) 1
Smoking Past/current smoker Never smoked	 10 (12.66) 69 (87.34)	 6 (60.00) 14 (20.30)	 **5.89 (1.49, 25.88)** 1
Education Below high school At least high school	 47 (59.50) 32 (40.50)	 13 (27.70) 7 (21.90)	 1.37 (0.49, 4.09) 1
Previous TB Yes No	 15 (18.99) 64 (81.01)	 5 (33.30) 15 (23.40)	 1 0.61 (0.19, 2.22)
BCG scar Present Absent/uncertain	 21 (26.58) 58 (73.42)	 14 (66.70) 6 (10.30)	 0.80 (0.27, 2.58) 1
Ever had contact with a person known to have TB Yes No Missing	 53 (67.09) 22 (27.85) 4 (5.06)	 10 (18.90) 10 (45.50)	 **3.58 (1.21, 10.85)**
Median age in years (range)	29 (17-59)	27.5(21, 46)	1.01 (0.96, 1.07)
Median BMI in kg/m^2(range)	32.57 (23.41, 52.08)	18.9 (13.6, 23.5)	**0.83 (0.68, 0.98)**
Median cough duration in months (range)	2 (0.46,24)	2 (0.46, 24)	1.03 (0.90, 1.16)
Median social network degree	23 (10, 87)	23 (10, 56)	0.99 (0.95, 1.03)
Closeness centrality			1.00 (0.999, 1.00)
Betweenness centrality			1.00 (1.00, 1.00)

Supplementary table 4.4: Factors associated with clustering in the multivariate logistic regression analysis (SNP threshold=5)

Variable	Adjusted OR (95% CI)
Sex Male Female	 1.85 (0.45, 9.59) 1
Smoking Past/current smoker Never smoked	 4.97 (0.82, 37.69) 1
Ever had contact with a person known to have TB Yes No	 2.54 (0.70, 9.11) 1
Median age in years (range)	0.97 (0.90, 1.04)
Median BMI in kg/m^2(range)	0.84 (0.66, 1.02)

Supplementary table 4.5: Factors associated with clustering in the univariate Modified Poisson analysis (SNP threshold=5)

Variable	Total number of cases (N=79) (%)	No (%) clustered (n=29); 36.7%	PR (95% CI)
Sex			
Male	54 (68.35)	24 (44.44)	2.62 (0.85, 8.15)
Female	25 (31.65)	5 (20.00)	1
HIV status			
Positive	12 (15.19)	2 (16.67)	1
Negative	66 (83.54)	27 (40.90)	1.64 (0.44, 6.16)
Missing	1 (1.27)		
Monthly income			
<200,000 UGSHS	23 (29.11)	6 (26.10)	0.81 (0.33, 1.97)
≥200,000 UGSHS	56 (70.89)	23 (41.10)	
Alcohol use			
Yes	34 (43.04)	16 (47.06)	1.62 (0.76, 3.46)
No	45 (56.96)	13 (28.89)	1
Missing			
Smoking			
Past/current smoker	10 (12.66)	8 (80.00)	**2.96 (1.49, 5.89)**
Never smoked	69 (87.34)	21 (30.43)	1
Education			
Below high school	47 (59.50)	18 (38.30)	1.26 (0.57, 2.82)
At least high school	32 (40.50)	11 (34.38)	1
Previous TB			
Yes	15 (18.99)	6 (40.00)	1
No	64 (81.01)	23 (35.94)	1.25 (0.55, 2.83)
BCG scar			
Present	21 (26.58)	20 (95.24)	0.85 (0.37, 1.91)
Absent/uncertain	58 (73.42)	9 (15.52)	1
Ever had contact with a person known to have TB			
Yes	53 (67.09)	11 (20.75)	**2.41 (1.17, 4.96)**
No	22 (27.85)	18 (81.82)	
Missing	4 (5.06)		
Median age in years (range)	29 (17-59)	27(20, 47)	1.01 (0.97, 1.05)
Median BMI in kg/m²(range)	32.57 (23.41, 52.08)	19.13 (13.55, 23.67)	0.87 (0.78, 0.98)
Median cough duration in months (range)	2 (0.46,24)	2 (0.46, 24)	1.02 (0.94, 1.11)
Median social network degree	23 (10, 87)	23 (10, 56)	0.995 (0.968, 1.022)

Closeness centrality			1.00 (1.00, 1.00)
Betweenness centrality			1.00 (1.00, 1.00)

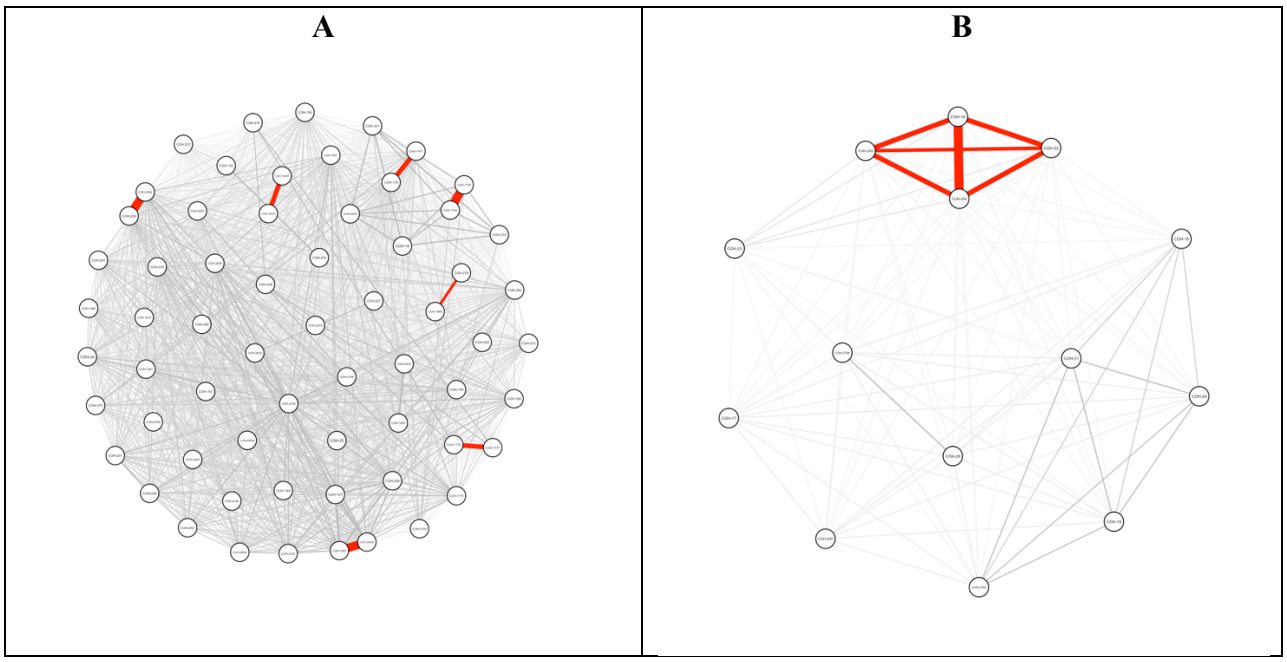
Supplementary table 4.6: Factors associated with clustering in the multivariate Modified Poisson analysis (SNP threshold=5)

Variable	Adjusted PR (95% CI)
Sex Male Female	 1.62 (0.51, 5.18) 1
Smoking Past/current smoker Never smoked	 2.36 (0.89, 6.27) 1
Ever had contact with a person known to have TB Yes No	 1.79 (0.85, 3.80) 1
Median age in years (range)	0.98 (0.94, 1.02)
Median BMI in kg/m^2(range)	0.90 (0.81, 1.00)

Supplementary table 4.7: Genomic clusters

Cluster ID	Cluster size	Lineage	Number of genetic links	Number of SNPs	Number of social links	Social network distance
1	2	1	1	0	0	
2	4	3	6	0,1,1,1,1,2	3	7,10,10
3	4	4	6	0,1,7,7,11,11	0	
4	3	4	2	12,12	2	7,8
5	2	4	1	0	1	10
6	2	4	1	1	1	1
7	2	4	1	6	0	
8	2	4	1	1	1	1
9	2	4	1	0	0	
10	2	4	1	12	1	1
11	2	4	1	1	1	11
12	2	4	1	3	0	
Sum	29		23		10	

**Genetic clusters identified with a threshold of 12 SNPs*



Supplementary figure 4.1: Identified clusters (SNP threshold=5). A: Lineage 4. B: Lineage 3.

Supplementary table 4.8: Description of clusters (SNP threshold=5)

Cluster ID	Cluster size	Lineage	Number of genetic links
1	2	1	1
2	4	3	6
3	2	4	1
4	2	4	1
5	2	4	1
6	2	4	1
7	2	4	1
8	2	4	1
9	2	4	1
Sum	20		14

CHAPTER 5 : DEVELOPMENT OF A STOCHASTIC NETWORK MODEL TO STUDY THE TRANSMISSION DYNAMICS OF MYCOBACTERIUM TUBERCULOSIS[4]

[4] R Galiwango, A Handel, J Sekandi, L Liu and C Whalen. To be submitted to PLOS Computational Biology.

ABSTRACT

Background: Unlike deterministic compartmental models, individual-based models such as network models allow us to account for heterogeneity in mixing of individuals in the population. Network models have been used for the study of transmission dynamics of other infectious diseases such as HIV but not so much for tuberculosis yet like for HIV, network structure plays a critical role in the transmission of *Mycobacterium tuberculosis*.

Methods: I developed a stochastic network model to be used to study the transmission dynamics of *Mycobacterium tuberculosis*. I implemented an individual-based version (particularly a network model) of a deterministic model with two latency compartments on a dynamic network simulated from a static, cross-sectional network of indexes (cases and controls) and their contacts from the Community Health and Social Networks of TB (COHSONET) study. I used the Statnet suite of packages to build the COHSONET social network, to simulate a dynamic network and to run the epidemic model on the dynamic network. I assessed the viability of the model by running simulations at different values of the input parameters and observed the effect on the overall dynamics. I compared the results with those of a deterministic version of the model.

Results: The model worked as expected with the number of susceptible individuals decaying exponentially with time (since there was no replenishment of susceptible individuals) and the number of latently infected individuals and TB diseased individuals increasing exponentially with time until all susceptible individuals were depleted. Increasing the infection probability or the contact rate quickened the epidemic as expected. A deterministic version of the model less to fewer infections

Future direction: The model will be extended to make it more realistic by accounting for drug resistance. I will test network-based interventions such as giving the intervention to only first level contacts of index TB cases and compare this with giving the intervention to both their first level and second level contacts. I will then develop an optimal combination of interventions that is necessary to achieve the targets of elimination spelt out in the end-TB strategy, in an endemic setting in Sub-Saharan Africa and in similar settings. The model could be used to answer the question on whether infections in the household are sufficient to maintain the epidemic in the community, and if not so, simulate different scenarios that can explain the observed infections in the community.

INTRODUCTION

Deterministic compartmental models are useful for studying the transmission dynamics of an infectious disease and consequently for informing public health interventions. However, these models are so simplistic in that they assume random (homogeneous) mixing of individuals in the population meaning that all susceptible persons have equal probabilities of getting infected which is not always true. In practice, each infectious individual has a finite set of contacts to whom they can pass infection. Individual-based models on the other hand such as network models allow us to account for this variability in mixing of individuals in the population. We can thus explore the effect of the underlying structure of the network on dynamics occurring on the network.

It has always been known that compartmental models are too simplistic. What has been lacking are the necessary tools to implement more accurate connection structures. With the emergency of tools such as the Statnet suite of packages [56], we can explore transmission dynamics of infectious diseases using more realistic stochastic network transmission models.

Network models are compelling for studying transmission dynamics of respiratory pathogens that are transmitted via close contact (or airborne) *Mycobacterium tuberculosis* inclusive since infectious individuals generally pass on infection to their contacts.

Network models have been used for the study of transmission dynamics of other infectious diseases such as HIV [13–16] but not so much for tuberculosis yet like for HIV, network structure plays a critical role in the transmission of *Mycobacterium tuberculosis*. For example, household contacts of index TB cases particularly children are at an elevated risk of acquiring TB though the proportion of transmission that attributed to household contact has been shown to be low in household contact studies [32].

I developed a stochastic network model to be used to study the transmission dynamics of *Mycobacterium tuberculosis*. I implemented an individual-based (particularly, a network model) version of a deterministic model with two latency compartments [11,12] on a dynamic network simulated from a static, cross-sectional network of indexes (cases and controls) and their contacts from the Community Health and Social Networks of TB (COHSONET) study. I used the Statnet suite of packages [56] to develop the COHSONET social network, to simulate a dynamic network and to run the epidemic model on the dynamic network. Social/contact networks are not static but rather they are dynamic structures. New relations form between individuals in the social network with time (relational formation) while existing ones are dissolved over time (relational dissolution). Therefore, studying the spread of infectious diseases in general and tuberculosis in particular is more realistic if done on dynamic networks.

METHODS

I used the Statnet suite of packages [56], particularly the 'tergm' package and the 'networkDynamic' package implemented in R's statistical software (www.r-project.org) to estimate a dynamic network from a static, cross-sectional social network of indexes (TB cases and matched controls) and their first and second level contacts in the Community Health and Social Networks of TB (COHSONET) study (supplementary materials). Statnet [56] uses Separable Temporal Exponential-family Random Graph Models (STERGMs) to estimate a dynamic network from a static network based on observed statistical properties of the static network such as density, degree and clustering. In this approach, two Exponential-family Random Graph Models (ERGMs) are used to model the dynamic network: one ERGM is used to model relational formation while the other is used to model relational dissolution.

I used the Bernoulli (Erdős–Rényi) model [125] for the formation formula but added a term for the number of completed triangles. The Bernoulli model has only one term (the number of edges) which captures the density of the network as a function of a homogenous edge probability. Two individuals in a network are said to form a triangle if they share a contact. The triangle is said to be closed if the two individuals who share a contact are also connected in the network i.e., are contacts of each other. The number of triangles in the network are often used as a measure of clustering for the network. The higher the number of triangles the, higher the clustering coefficient (degree of compactness of the network). On the other hand, I specified a simple dissolution model with only the edges term. This model implies that the probability of edge dissolution at each discrete time point is a homogeneous, constant hazard across ties i.e., it doesn't depend on the specific configuration of individuals forming a tie.

I simulated a network version of a deterministic model with two latency compartments [11,12] on the resultant dynamic network using the EpiModel package [126] which is also part of the Statnet suite of packages [56]. A model with two latency compartments (one for low-risk latently infected persons and the other for high-risk latently infected persons) was shown to reproduce actual transmission dynamics (Ragonnet et al., 2017). On the other hand, models with one latency compartment were shown to produce unreasonably poor fits to empirical data.

This deterministic model consists of four compartments: S(t) for the number of susceptible individuals at time t, LA(t) for the number of high-risk latently infected individuals (the fast progressors) at time t, LB(t) for the number of low-risk latently infected individuals (the slow progressors) at time t and I(t) for the number of active TB diseased individuals at time t. Only individuals in compartment I(t) are infectious. On infection, a proportion, g, of infected individuals moves into compartment LB(t) while the rest move into compartment LA(t). High-risk latently infected individuals progress to active disease at some rate ϵ while low-risk latently infected individuals progress to active disease at a lower rate v.

To implement a network version of this deterministic model, I modified the in-built infection module in the EpiModel package [126] to include two latency compartments. I also developed a new disease progression module for progression from latency to active TB disease. The per tie (relation) transmission rate is calculated given by $1 - (1 - p)^{\mathbf{c}}$ where p is the probability of infection per transmissible contact between a susceptible individual and a person with TB disease and **c** is the average number of transmissible contacts per contact per unit time. Transmission is a Bernoulli trial (binomial with n=1 trials) with probability of infection = the per tie transmission rate. Progression to active disease was also modeled by a Bernoulli distribution with a lower probability of progression for slow progressors compared to fast progressors.

The model was parameterized with local data from the COHSONET study (supplementary materials). Additional parameters were obtained from a study of tuberculosis transmission in S. Africa [127]. The model was run for 5 years (an equivalent of 60 months) and the dynamics were observed for different levels of the transmission probability and the contact rate. All rate parameters were converted to units of months. I assessed the viability of the model. I run the model with an initial number of 123 diseased individuals, a figure equivalent to the number of index TB cases in the COHSONET study.

I run a deterministic version of the model and compared the results with those of the network model. The parameters used in the model were: proportion of infected individuals transitioning to a low-risk compartment (LB) immediately after infection, g=0.86, rate of progression to active TB from the high-risk compartment (LA), ϵ=0.88/12, rate of progression to active TB from the low-risk compartment (LB), v=0.00011/12 (Table 5.1).

I observed the dynamics for probability of infection per transmissible contact=0.01 and a lower rate of 0.001. I also run the model with an average number of transmissible acts per tie (average number of contacts between diseased individuals and susceptible individuals) per month, c=1 and compared that with c=5.

RESULTS

Viability of the model

As expected, the number of susceptible individuals decreased exponentially with time (figure 5.2) since there was no replenishment of susceptible individuals. The number of latently infected individuals increased exponentially with time until steady state (when all susceptible individuals were depleted) with the number of low-risk latently infected individuals being lower than the number of high-risk latently infected individuals throughout the epidemic curve as

expected. The number of TB diseased individuals increased exponentially with time until steady state.

Increasing the infection probability from 0.001 to 0.01 quickened the epidemic with susceptible individuals being depleted at an earlier time of 10 months compared to 20 months initially. This is expected as susceptible population is being infected at a higher rate. The number of incident TB cases for p=0.01 was higher at the peak of the epidemic compared to p=0.001 (figure 5.3). Similarly, increasing the contact rate from c=1 to c=5 depleted the susceptible population at a faster rate (figure 5.4). The number of incident TB cases at c=1 was also higher at the peak of the epidemic compared to c=5 (figure 5.5).

In the deterministic model, few new infections resulted compared with the network model, at the same parameter values (figure 5.5 and 5.6).

DISCUSSION

I implemented a network version of the deterministic model for *Mycobacterium tuberculosis* transmission with two latency compartments [11,12] on a dynamic network simulated from a static, cross-sectional network of indexes (cases and controls) and their contacts from the COHSONET study. I assessed the viability of the model by running simulations at different values of the input parameters and observing their effect on the overall dynamics.

The model worked as expected with the number of susceptible individuals decaying exponentially with time (since there was no replenishment of susceptible individuals) and the number of latently infected individuals and TB diseased individuals increasing exponentially with time until all susceptible individuals were depleted. Increasing the infection probability or the contact rate quickened the epidemic as expected.

In a model with two latency compartments, the activation dynamics are driven by two exponential components that are associated with two independent growth rates. Models with two latency compartments have been shown to accurately replicate empirically observed dynamics [11,12]. On the other hand, models with one latency compartment were shown to produce unreasonably poor fits to empirical data. Such models only involve a single exponential function, which is not sufficient to replicate the two distinct patterns observed in the dynamics of activation—a high risk of disease activation over the first few months, followed by a dramatically lower risk in a second phase [11].

Deterministic compartmental models assume uniform mixing of individuals in the population which is not always true. Individual-based models on the other hand such as network models explored here allow us to account for mixing patterns of individuals in the population. We can thus explore the effect of the underlying structure of the network on dynamics occurring on the network. Network models are compelling for studying transmission dynamics of respiratory pathogens, *Mycobacterium tuberculosis* inclusive, that are transmitted via close contact (or airborne) since infectious individuals generally transmit to their contacts. It has always been known that compartmental models are too simplistic. What has been lacking are the necessary tools to implement more accurate connection structures. With the emergency of tools such as the Statnet suite of packages [56], we can explore transmission dynamics of infectious diseases using more realistic stochastic network transmission models.

The biggest limitation of network models in particular and individual based models in general is that they are computationally involved. Relational data is usually big in itself and therefore running a transmission model on it creates a complex system that requires fast computers with bigger memory (RAM) and software tools that are capable of handling big data

and making complex simulations. This is why deterministic compartmental models have mainly been the go-to methodology for studying transmission dynamics of infectious diseases and testing interventions. However, these models are less realistic compared to individual based models. It is thus a trade-off between simplicity and realism.

The model I have developed is a basic model that only includes four compartments (susceptible, low-risk latently infected, high-risk latently infected and diseased). The model can be made more realistic with regard to *Mycobacterium tuberculosis* transmission by including two parallel compartmental structures with one representing transmission of a drug sensitive strain and another for transmission of a drug resistant strain. Interaction between the two structures occurs when a proportion of diseased individuals on defaulting treatment and acquire drug resistance thus crossing from the drug sensitive structure to the drug resistance structure. Persons may also develop drug resistance when they are infected by a drug resistance strain (transmitted drug resistance). Susceptible individuals can be replenished at a constant rate or one that includes birth and in-migration while all individuals in the population can die due to natural causes or due to TB disease. Individuals with TB disease can recover spontaneously or after completing treatment and they can be re-infected at a given rate on recovery. Latently infected individuals can also be re-infected.

On top of treating diseased individuals, other interventions such as BCG vaccination and chemotherapy of latently infected persons can be applied to the system. Since the vaccine (BCG) is not 100% efficacious, a proportion of vaccinated individuals would become latently infected with either a drug sensitive strain or a drug resistant strain and would thus move to the respective latent compartments. Vaccination reduces the number of susceptible individuals in the population that diseased individuals would come into contact with in the network. On the other

hand, a proportion of latently infected individuals would be given Isoniazid Preventive Therapy (IPT) at a given rate and they return to the susceptible compartment.

We can then explore which individuals in the network can be targeted for interventions for example the effect of giving interventions to only first level contacts of index TB cases compared with the difference made when the intervention is given to both first level contacts and second level contacts (the contacts of first level contacts). We can also determine an optimal combination of interventions that is necessary to achieve the targets of elimination spelt out in the End-TB strategy.

Another potential application of the model is to answer the question on whether infections in the household are sufficient to maintain the epidemic in the community, and if not so, simulate different scenarios that can explain the observed infections in the community.

REFERENCES

1. Basu, S., Andrews, J. R., Poolman, E. M., Gandhi, N. R., Shah, N. S., Moll, A., … Friedland, G. H. (2007). Prevention of nosocomial transmission of extensively drug-resistant tuberculosis in rural South African district hospitals: an epidemiological modelling study. *Lancet*, *370*(9597), 1500–1507. https://doi.org/10.1016/S0140-6736(07)61636-5
2. Goodreau, S. M., Hamilton, D. T., Jenness, S. M., Sullivan, P. S., Valencia, R. K., Wang, L. Y., … Rosenberg, E. S. (2018). Targeting Human Immunodeficiency Virus Pre-Exposure Prophylaxis to Adolescent Sexual Minority Males in Higher Prevalence Areas of the United States: A Modeling Study. *Journal of Adolescent Health*, *62*(3), 311–319. https://doi.org/10.1016/j.jadohealth.2017.09.023
3. Hamilton, D. T., Goodreau, S. M., Jenness, S. M., Sullivan, P. S., Wang, L. Y., Dunville, R. L., … Rosenberg, E. S. (2018). Potential impact of HIV preexposure prophylaxis among

black and white adolescent sexual minority males. *American Journal of Public Health*, *108*, S284–S291. https://doi.org/10.2105/AJPH.2018.304471

4. Hamilton, D. T., Rosenberg, E. S., Jenness, S. M., Sullivan, P. S., Wang, L. Y., Dunville, R. L., … Goodreau, S. M. (2019). Modeling the joint effects of adolescent and adult PrEP for sexual minority males in the United States. *PLoS ONE*, *14*(5), 1–12. https://doi.org/10.1371/journal.pone.0217315
5. Handcock, M. S., Hunter, D. R., Butts, C. T., Goodreau, S. M., & Morris, M. (2003). statnet: Software Tools for the Statistical Modeling of Network Data . *Statnet Project Http: //Statnetproject.Org/*, *24*(1), Seattle, WA. R package version 2.0, URL http://CRA.
6. Jenness, S. M., Goodreau, S. M., & Morris, M. (2018). Epimodel: An R package for mathematical modeling of infectious disease over networks. *Journal of Statistical Software*, *84*(8). https://doi.org/10.18637/jss.v084.i08
7. Jenness, S. M., Maloney, K. M., Smith, D. K., Hoover, K. W., Goodreau, S. M., Rosenberg, E. S., … Sullivan, P. S. (2019). Addressing Gaps in HIV Preexposure Prophylaxis Care to Reduce Racial Disparities in HIV Incidence in the United States. *American Journal of Epidemiology*, *188*(4), 743–752. https://doi.org/10.1093/aje/kwy230
8. Martinez, L., Shen, Y., Mupere, E., Kizza, A., Hill, P. C., & Whalen, C. C. (2017). 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
9. 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

10. Newman, M. E. J., Watts, D. J., & Strogatz, S. H. (2002). Random graph models of social networks, *99*. https://doi.org/10.1073/pnas.012582999
11. 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

TABLES AND FIGURES

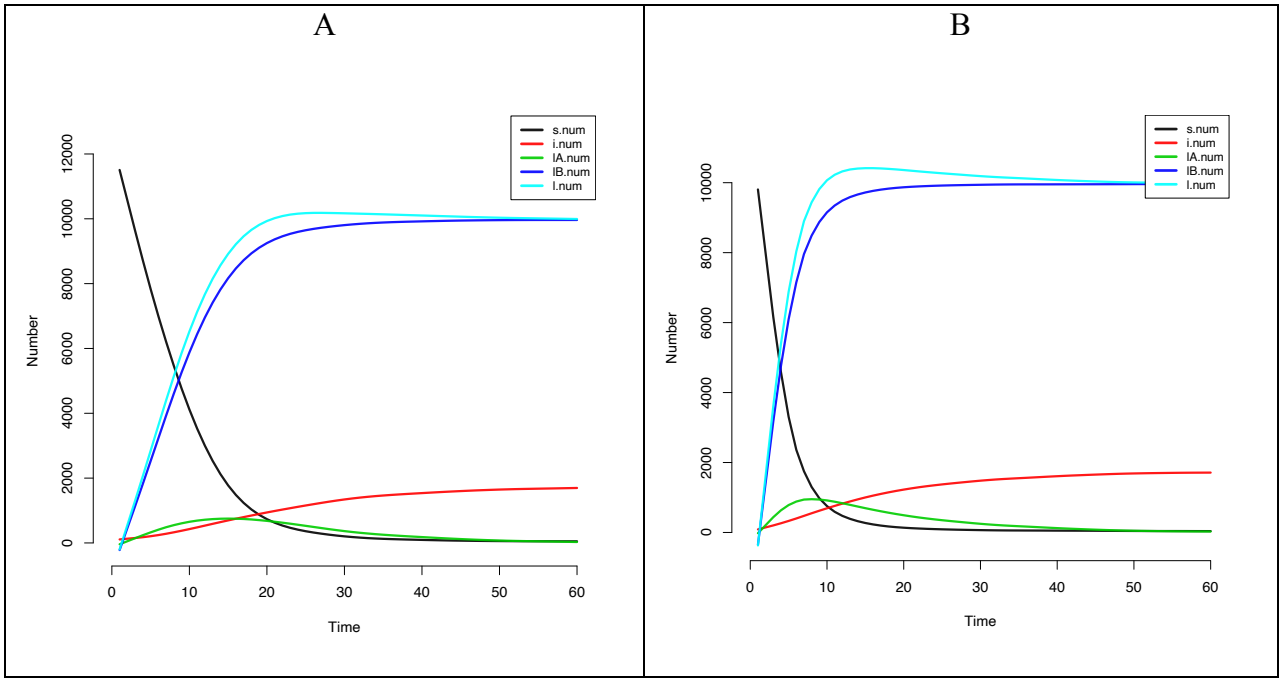


Figure 5.1: Transmission dynamics for different levels of the transmission probability

Transmission dynamics: s.num (number of susceptibles), i.num (number of diseased), fast lA.num (fast progressors), lB.num (slow progressors) and total number of latently infected individuals (l.num). A: transmission probability=0.001. B: transmission probability=0.01.

Table 5.1: Parameters used in the model

Parameter	Description	Value
g	proportion of infected individuals transitioning to a low-risk compartment (LB) immediately after infection	0.86
ϵ	rate of progression to active TB from the high-risk compartment (LA)	0.88/12
v	rate of progression to active TB from the low-risk compartment (LB	0.00011/12
c	effective contact rate	4.9/day
p	transmission probability	0.011

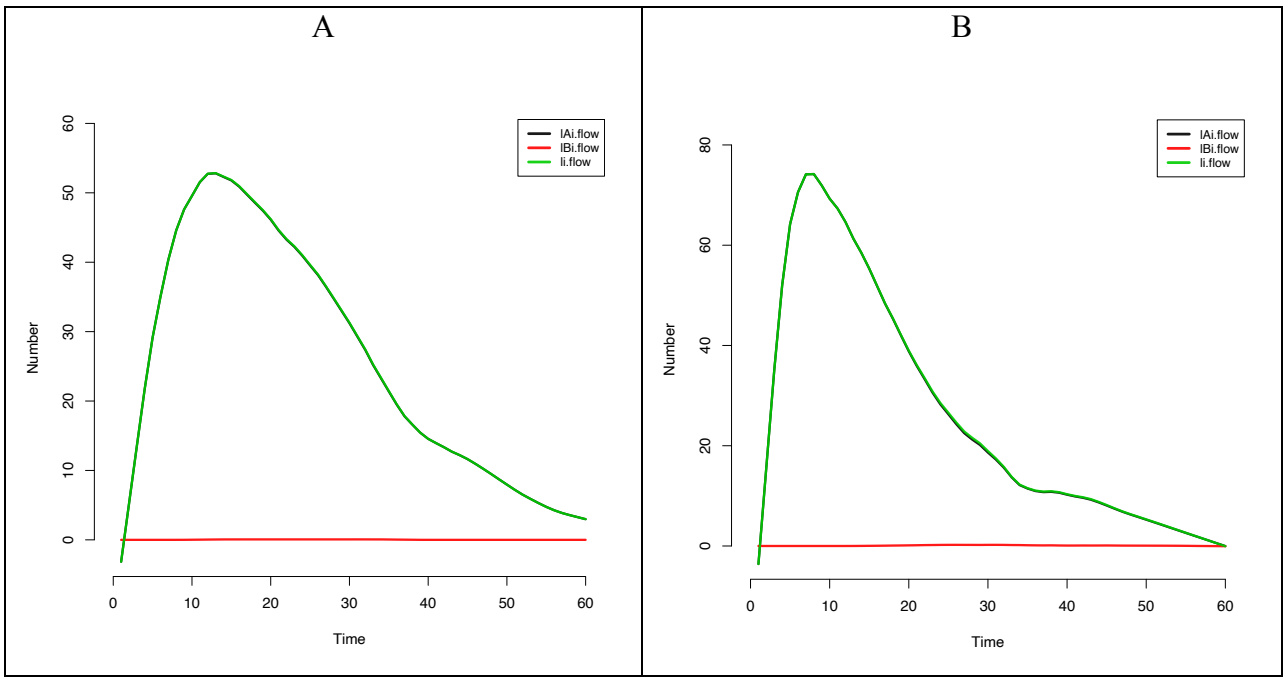


Figure 5.2: Incidence at different values of the transmission probability

Incidence. A: transmission probability=0.001; B: transmission probability=0.01

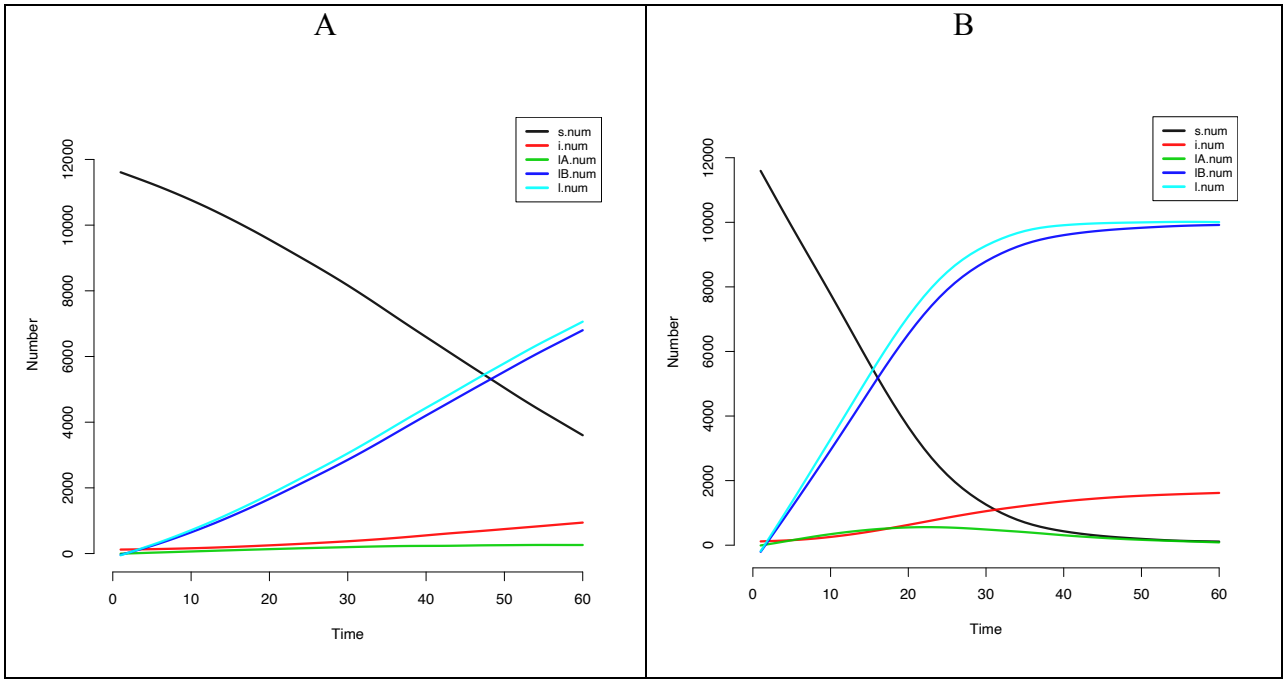


Figure 5.3: Transmission dynamics at different levels of the contact rate

Transmission dynamics: s.num (number of susceptibles), i.num (number of diseased), fast lA.num (fast progressors), lB.num (slow progressors) and total number of latently infected individuals (l.num). A: Contact rate=1. B: Contact rate=5

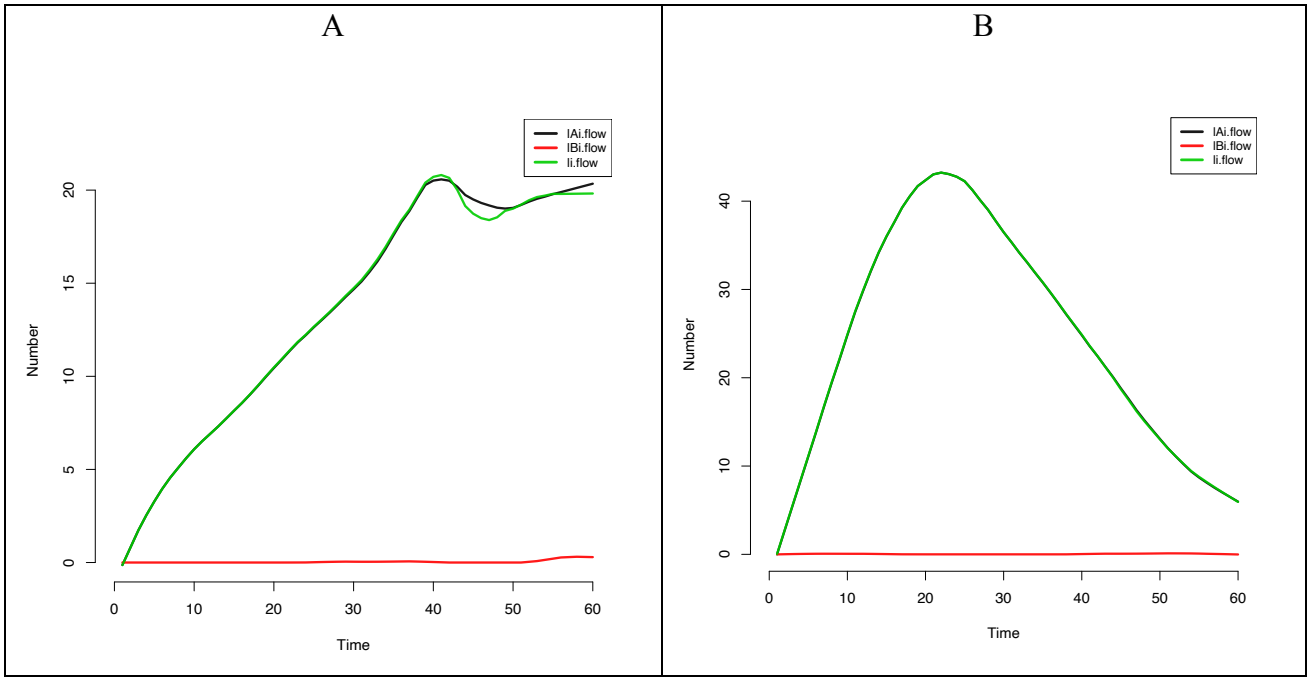


Figure 5.4: Incidence at different values of the contact rate

Incidence. A: Contact rate=1; B: Contact rate=5

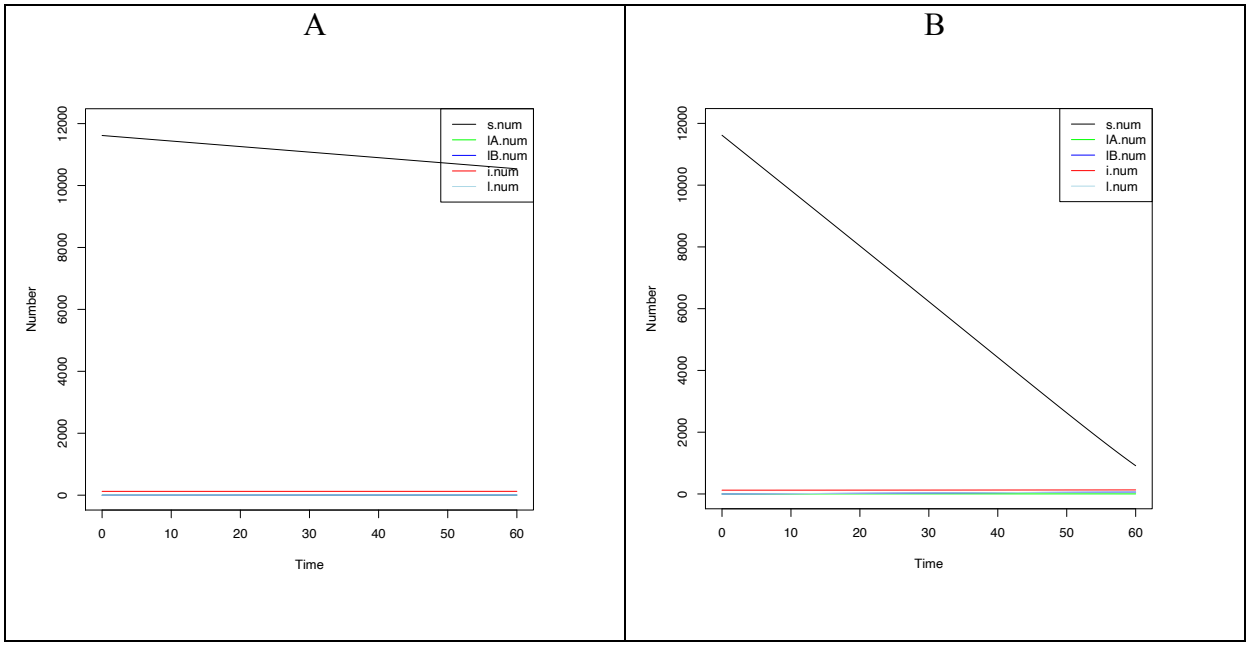


Figure 5.5: Transmission dynamics for different levels of the transmission probability in the deterministic model

A: transmission probability=0.001. B: transmission probability=0.01

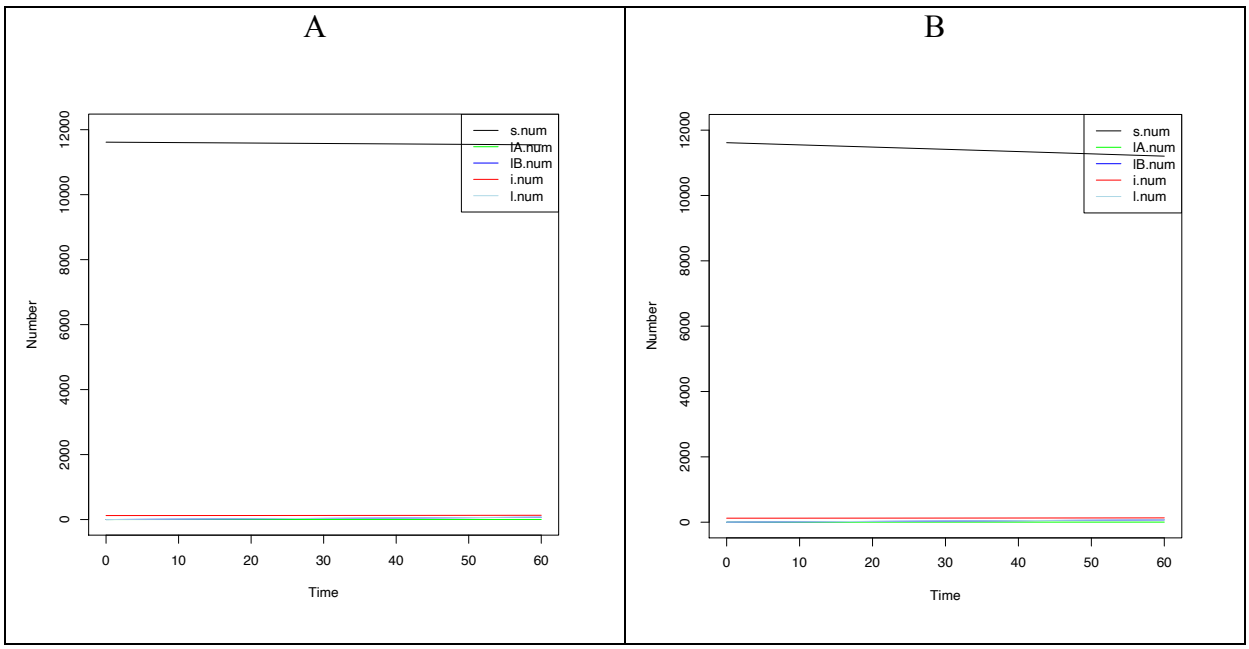


Figure 5.6: Transmission dynamics for different levels of the contact rate in the deterministic model

A: contact rate=0.001. B: contact rate=0.01

CHAPTER 6 : CONCLUSION

MOTIVATION

This research aimed at filling four gaps identified during the review of the literature. First, I aimed at discussing current approaches for processing WGS data from TB pathogen isolates for purposes of making inferences on transmission of *Mycobacterium tuberculosis* and to update existing literature on the methods used to make direct transmission inferences. Second, I aimed to explore the relevance of the social network of an index tuberculosis case in the transmission of *Mycobacterium tuberculosis*. Third, I aimed to identify the critical drivers of *Mycobacterium tuberculosis* in an endemic urban setting in Sub-Saharan Africa. Fourth, I aimed to develop a stochastic network model to be used to study *Mycobacterium tuberculosis* transmission.

SYNTHESIS OF MAIN FINDINGS

Aim 1: In a systematic review, we found heterogeneity in processing of WGS data among studies and some areas of consensus especially in recent literature. Standardization of data processing methodology such as with creation of standardized computational pipelines could improve comparability of transmission inference results. SNP thresholds are the most widely used method for inferring transmission because of their simplicity, with a threshold of 12 SNPs the most widely used. Bayesian transmission modeling attempts to address their limitation and is increasingly being used in transmission studies.

Aim2: In a large social network study of tuberculosis (COHSONET), we found that transmission often happens outside of the defined social network of an individual case. Further

exploration of other mechanisms of extra-household transmission of *Mycobacterium tuberculosis* is required. One way of doing this is by studying mobility of tuberculosis patients several months prior to diagnosis so as to identify community venues and geographical locations in the community where transmission occurs. We can also reconstruct community networks of index TB cases by identifying geographical locations spanned by each TB case using cellphone meta data.

I found no correlation between genetic distance and social network distance. For a disease that requires adequate contact for effective transmission to occur, our hypothesis before the study was that patients at close social network distance are more likely to have genetically similar strains but this wasn't the case. Previous studies have investigated the relationship between genetic distance and geographic distance [51,116–118] and found that patients living at close proximity were more likely to have genetically similar strains. Geographical distance could be a better measure of proximity than social network distance.

Aim 3: We identified clusters of recent transmission in the COHSONET study using high resolution whole genome sequencing. We found that clustered cases were more likely to be current or past smokers. This study adds to the growing literature on the increased risk of acquiring tuberculosis by current smokers or persons who have ever smoked [121–123]. Smoke particles have been shown to impair macrophages, which are critical immune cells in fighting mycobacterium tuberculosis [124]. There is a need for targeted interventions among identified risk groups in order to interrupt transmission.

Aim 4: Unlike deterministic compartmental models, network models account for heterogeneity in mixing patterns. I implemented a network version of a deterministic model with two latency compartments on a dynamic network simulated from a static network. The model

depicted expected dynamics in a viability analysis when compared with a deterministic version. The model will be used to answer research questions such as whether infections in the household are sufficient to maintain the epidemic in the community, and if not so, different scenarios explaining the observed infections in the community will be simulated.

STUDY LIMITATIONS

When conducting the systematic review, information from the studies was extracted as reported. Researchers may have done a particular data processing step during the analysis but may have not reported it. Nevertheless, the major data processing steps should be reported because each step in the pipeline influences the inferences made.

In the COHSONET study, we did not enroll all consecutive TB patients during the study period and not all isolates were sequenced. It is therefore possible that we underestimated the proportion of clustered patients. However, there was no statistically significant difference between characteristics of patients whose isolates were sequenced and those whose isolates were not sequenced.

It is also possible that some nodes and edges were miss-specified during the search for duplicates. However, use of local content experts when matching records who were knowledge in local names and their sex affiliation decreased the likelihood of this occurring. Despite these limitations, this study represents the largest most comprehensive social network study of tuberculosis in Africa.

PUBLIC HEALTH RECOMMEDATIONS

There is a need to standardize data processing methodology such as with creation of standardized computational pipelines so as to improve comparability of transmission inference results.

Since transmission often happens outside of the defined social network of an individual case, studying mobility of tuberculosis patients several months prior to diagnosis could enable us to better understand extra-household transmission of *Mycobacterium tuberculosis*.

There is a need for targeted interventions among identified risk groups, for example current or past smokers found in this study, in order to interrupt transmission.

FUTURE DIRECTION

Other potential transmission routes of tuberculosis could be explored by identifying locations in the community where transmission occurs. Such hotspots of transmission could be identified by studying mobility of index TB patients several months prior to diagnosis. By so doing, we use mobility of tuberculosis patients as an indicator of TB transmission. We can reconstruct community networks of index TB cases using their cellphone meta data and link these cases using these data. When coupled with Whole genome sequencing of pathogen isolates from diseased persons, these data could help improve our understanding of extra-household transmission.

The stochastic network model developed will be extended to make it more realistic by accounting for drug resistance. The model will be used to answer research questions such as whether infections in the household are sufficient to maintain the epidemic in the community, and if not so, different scenarios explaining the observed infections in the community will be simulated. I will also test network-based interventions such as giving the intervention to only

first level contacts of index TB cases and compare this with giving the intervention to both their first level and second level contacts. I will then develop an optimal combination of interventions that is necessary to achieve the targets of elimination spelt out in the end-TB strategy, in an endemic setting in Sub-Saharan Africa and in similar settings.

REFERENCES

1. Deatherage, D. E., & Barrick, J. E. (2014). Identification of mutations in laboratory-evolved microbes from next-generation sequencing data using breseq. *Methods in Molecular Biology (Clifton, N.J.)*, *1151*, 165–188. https://doi.org/10.1007/978-1-4939-0554-6_12
2. Didelot, X., Fraser, C., Gardy, J., Colijn, C., & Malik, H. (2017). Genomic infectious disease epidemiology in partially sampled and ongoing outbreaks. *Molecular Biology and Evolution*, *34*(4), 997–1007. https://doi.org/10.1093/molbev/msw275
3. Kohl, T. A., Utpatel, C., Schleusener, V., De Filippo, M. R., Beckert, P., Cirillo, D. M., & Niemann, S. (2018). MTBseq: a comprehensive pipeline for whole genome sequence analysis of Mycobacterium tuberculosis complex isolates . *PeerJ*, *6*, e5895. https://doi.org/10.7717/peerj.5895
4. 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
5. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352
6. Petkau, A., Mabon, P., Sieffert, C., Knox, N. C., Cabral, J., Iskander, M., … Van Domselaar, G. (2017). SNVPhyl: a single nucleotide variant phylogenomics pipeline for microbial

genomic epidemiology. *Microbial Genomics*, *3*(6), e000116. https://doi.org/10.1099/mgen.0.000116

7. Sahl, J. W., Lemmer, D., Travis, J., Schupp, J. M., Gillece, J. D., Aziz, M., … Keim, P. (2016). NASP: an accurate, rapid method for the identification of SNPs in WGS datasets that supports flexible input and output formats. *Microbial Genomics*, *2*(8), e000074. https://doi.org/10.1099/mgen.0.000074
8. Walker, T. M., Ip, C. L. C., Harrell, R. H., Evans, J. T., Kapatai, G., Dedicoat, M. J., … Peto, T. E. A. (2013). Whole-genome sequencing to delineate Mycobacterium tuberculosis outbreaks: a retrospective observational study. *The Lancet Infectious Diseases*, *13*, 137–146. https://doi.org/10.1016/S1473-3099(12)70277-3