Distribution Agreement

Signature:

______________________________	______________
Kristin Bratton Nelson	Date

Transmission patterns of extensively drug-resistant tuberculosis in South Africa: a network approach

By

Kristin Bratton Nelson

M.P.H, Emory University, 2014

B.S. Biochemistry and Molecular Biophysics, University of Arizona, 2012

____

Neel R. Gandhi, MD
Advisor

____

Samuel M. Jenness, PhD
Committee Member

____

Benjamin A. Lopman, PhD
Committee Member

____

Barun Mathema, PhD
Committee Member

Accepted:

____

Lisa A. Tedesco, Ph.D.
Dean of the James T. Laney School of Graduate Studies

____

Date

Transmission patterns of extensively drug-resistant tuberculosis in South Africa: a network approach

By

Kristin Bratton Nelson, M.P.H.

M.P.H, Emory University, 2014

B.S. Biochemistry and Molecular Biophysics, University of Arizona, 2012

Advisor: Neel R. Gandhi, MD

Committee:

Samuel M. Jenness, PhD

Benjamin A. Lopman, PhD

Barun Mathema, PhD

An abstract of

A dissertation submitted to the Faculty of the

James T. Laney School of Graduate Studies of Emory University

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in Epidemiology

2018

Abstract

Transmission patterns of extensively drug-resistant tuberculosis in South Africa: a network approach

By Kristin Bratton Nelson

Tuberculosis (TB) is the leading infectious cause of disease worldwide and there were over half a million cases of drug-resistant TB in 2017. Transmission plays a critical role in the spread of extensively drug-resistant (XDR) TB in South Africa and globally. However, an incomplete understanding of the risk factors for XDR TB transmission prevents design of effective interventions to curtail transmission. Further, there has been little focus on the critical role of missing cases in transmission networks, though understanding missingness is essential to ensure accurate estimation of underlying transmission patterns.

We combined bacterial whole genome sequencing to identify transmission events with network analysis to investigate the clinical and behavioral factors driving XDR TB transmission. In **Aim 1**, we used exponential random graph models (ERGMs) to measure associations between clinical markers of infectiousness and transmission of XDR TB. Cases reporting 2 to 3 months of cough were more highly connected in the network than those reporting no cough and smear-positive cases were more poorly connected than smear-negative cases. In **Aim 2,** we examined associations between social mixing patterns and transmission. Cases who spent time in urban settings were more highly connected in the network than those who did not, and cases with extended hospital stays were less connected that those who reported shorter hospital stays.

In **Aim 3**, we assessed the impact of missing XDR TB cases in the transmission network. We found that no single scenario we tested could account for the missingness in the empirical transmission network. However, missingness was unlikely to be random based on our models; the most likely scenarios involved oversampling of low-transmitting cases or omission of a factor strongly related to transmission from our models. Our results were strongly influenced by several key assumptions. This highlights the uncertainties in our transmission model, and about TB transmission broadly, that preclude more exact inference regarding underlying XDR TB transmission patterns.

Through gaining a clearer understanding of XDR TB transmission patterns in settings of high TB incidence, we can directly inform interventions that will halt the spread of drug-resistant TB in countries with the highest burdens of disease.

Transmission patterns of extensively drug-resistant tuberculosis in South Africa: a network approach

By

Kristin Bratton Nelson, M.P.H.

M.P.H, Emory University, 2014

B.S. Biochemistry and Molecular Biophysics, University of Arizona, 2012

Advisor: Neel R. Gandhi, MD

Committee:

Samuel M. Jenness, PhD

Benjamin A. Lopman, PhD

Barun Mathema, PhD

A dissertation submitted to the Faculty of the

James T. Laney School of Graduate Studies of Emory University

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in Epidemiology

2018

Acknowledgements

First, I want to thank my dissertation committee for their guidance and encouragement over the past four years. To Neel, who has been an invaluable mentor and my biggest advocate; to Sam, who has shown extraordinary patience and generosity as I've stumbled my way through learning network models; to Ben, who has provided balance and fresh perspective even as he insisted he didn't know enough about TB to be on this committee; and to Barun, who has always inspired me to think bigger and more deeply.

To all the doctoral students in the Epidemiology Department and Jena Black, who have celebrated my successes and commiserated when I needed a sympathetic ear.

Lastly, I want to thank my family and friends. This dissertation is dedicated to my parents, who have been my biggest supporters for the past 28 years. I am indebted to them for not only imprinting upon me the value of an education, but more importantly, for showing me what it is to have honor and integrity in all I do. I could not have done this without my husband, Michael, who has always believed that I am smart and capable. When I have forgotten this, you have reminded me. To my sisters, Danielle and Shauna, and to my closest friends, for the laughs and strength that knowing and loving you bring to my life. I am thankful to my second parents, Bob and Kim Nelson, who will attend more PhD defenses for their children than any parent should ever have to, but who are absolute and unwavering in support of their family.

Table of Contents

Chapter 1. Background 1

Chapter 2. Rationale and Specific Aims 33

Chapter 3. Data source 38

Chapter 4. Preliminary Work 38

Chapter 5. Aims 1 and 2 69

Chapter 6. Aim 3 107

Chapter 7. Public health implications and overall significance 160

Chapter 8. References 167

List of Tables and Figures

Tables

Table 4-1. Characteristics of participants in TRAX cohort and comparison to subsets with whole genome sequencing (WGS) results and with genomic links – KwaZulu-Natal Province, South Africa.

Table 4-2. Geographic distribution of XDR TB cases by district.

Table 4-3. Proportions of within- and between-district genomic links (≤ 5 SNPs) in KwaZulu-Natal.

Supplemental Table 4-1. Genomic links (≤ 5 SNPs) by sex, HIV status, and strain type.

Supplemental Table 4-2. Within and between-district links at ≤ 3 and ≤ 1 SNPs.

Supplemental Table 4-3. Within and between-district genomic links (≤ 5 SNPs).

Supplemental Table 4-4. Enrollment fraction bias analysis of between- and within-district genomic links (≤5 SNPs).

Table 5-1. Demographic, clinical and social mixing characteristics of the TRAX cohort.

Table 5-2. Associations between infectiousness and network position.

Table 5-3. Associations between social mixing measures and network position.

Supplemental Table 5-1. Characteristics of networks with alternate SNP thresholds and genomic link definitions.

Supplemental Table 5-2. Associations between infectiousness and network position in models with collapsed variable for smear status.

Supplemental Table 5-3. Associations between infectiousness and network position in networks with alternate genomic link definitions.

Supplemental Table 5-4. Associations between infectiousness and network position in full and parsimonious models.

Supplemental Table 5-5. Associations between infectiousness and network position using conventional regression methods.

Supplemental Table 5-6.

Supplemental Table 5-7. Associations between social mixing measures and network position in alternate models.

Supplemental Table 5-8. Associations between social mixing measures and network position in networks with alternate genomic link definitions.

Supplemental Table 5-9. Associations between social mixing measures and node outdegree using conventional regression methods.

Table 6-1. Missing case scenarios.

Table 6-2. Descriptive characteristics of the empirical sequencing-based network of XDR TB cases from the TRAX study.

Table 6-3. Target statistics of sampled, simulated networks under scenarios (1) and (2) compared to the empirical network.

Table 6-4. Target statistics of sampled, simulated networks under scenarios (3) and (4) compared to the empirical network.

Technical Appendix Table 6-5. Mean degree by cough duration.

Technical Appendix Table 6-6. Mean degree by smear status.

Technical Appendix Table 6-7. Mean degree by smear status.

Technical Appendix Table 6-8. Mean degree by *Mtb* strain type.

Technical Appendix Table 6-9. Mean degree by age.

Technical Appendix Table 6-10. Mean degree by age and smear status.

Technical Appendix Table 6-11. Mean degree by age and HIV status.

Supplemental Table 6-1. Descriptive characteristics of the sequencing-based network of XDR TB cases in the TRAX study (≤ 3 SNP threshold).

Supplemental Table 6-2. Effect of modifying network size on network models.

Supplemental Table 6-3. Effect of reducing SNP threshold (≤ 3 SNPs) on the empirical network.

Supplemental Table 6-4. Effect of reducing the SNP threshold (≤ 3 SNPs) on the empirical network, considering an unmeasured factor.

Figures

Figure 4-1. Selection of study participants and identification of genomic links using whole genome sequencing.

Figure 4-2. Geographic distribution of XDR TB cases with genomic links in KwaZulu-Natal province, South Africa.

Figure 4-3. Map and distribution of geographic distances between home residences of genomically linked case-pairs in KwaZulu-Natal.

Figure 4-4. Genomic links (≤ 5 SNPs) within and between districts in KwaZulu-Natal

Supplemental Figure 4-1. Number of genomic links (≤ 5 SNPs) per case.

Supplemental Figure 4-2. Shortest geographic link among genomic links (≤ 5 SNPs) for each case.

Figure 5-1. Variation in the number of links among TRAX cases in the sequencing-based network.

Figure 5-2. Associations between infectiousness and network position.

Figure 5-3. Associations between social mixing measures and network position.

Figure 6-1. Schematic representation of simulation and sampling methods.

Figure 6-2. Degree distributions of sampled, simulated networks under scenarios (1) and (2) compared to the empirical network.

Figure 6-3. Degree distributions of sampled, simulated networks under scenarios (3) and (4) compared to the empirical network.

Supplemental Figure 6-1. Mean degree required to reproduce the maximum degree in the empirical network.

Supplemental Figure 6-2. Effect of modifying complete network size on network models under random sampling.

Supplemental Figure 6-3. Effect of reducing SNP threshold (≤ 3 SNPs) on the empirical network.

Supplemental Figure 6-4. Mean degree required to reproduce the maximum degree in the ≤ 3 SNP empirical network.

1. Background

1.1 Global epidemiology of tuberculosis

1.1.1 Overview

Tuberculosis (TB) is the leading infectious cause of death worldwide: in 2016, an estimated 10.4 million people fell ill with TB and over 1.6 million died from the disease. [1] The etiologic agent of TB is *Mycobacterium tuberculosis (Mtb),* which is one of several mycobacterial species that make up the mycobacterium tuberculosis complex (MTBC). *Mtb* is an obligate pathogen for which humans are the only known reservoir, making human-to-human transmission the most consequential mode of TB spread.

Historically, *Mtb* has been a remarkably successful pathogen. Evidence suggests that some form of TB has caused disease in human populations since the pre-Neolithic period, and between the 17th and 19th centuries, it is estimated that TB killed one in five people in North America and Europe. [2] As living conditions have improved worldwide over the past century, TB incidence has dropped precipitously. This is especially true in high-income countries: annual disease rates in the United States and western European countries today range from 3 to 10 per 100,000 persons, with outbreaks largely confined to high-risk groups (incarcerated, drug-using, and homeless populations) and immigrant communities. [1, 3, 4] Progress has been slower in low and middle-income countries, where the majority of global TB cases are concentrated today. In 2016, only six countries accounted for over half of all notified TB cases: India, Indonesia, China, Nigeria, Pakistan, and South Africa. Collectively, these countries also account for over half of global drug-resistant TB cases (54%) and deaths (53%) from TB. [1] Achieving progress towards ending the global TB epidemic hinges on advancements in TB treatment, care, and prevention in these high-burden areas.

Recent trends in tuberculosis incidence in these high-burden settings have been shaped by the moderate success of modern TB control efforts. The number of diagnosed TB cases in most high-burden countries have seen slow though steady reductions since the mid-2000s. This is due in large part to recent improvements in TB diagnosis and treatment, including the implementation of rapid diagnostic tools and the introduction of shorter, more effective drug therapy regimens. The current global strategy for TB control is laid out by the End TB strategy, the goals of which include reducing worldwide incidence by 50% by 2025. [5] These goals are deliberately ambitious; however, they are a reminder that the global community can and should strive to do better to reduce the massive global burden of TB. Recent trends in TB epidemiology now characterize the most pressing challenges facing TB control: the circulation of increasingly drug resistant strains of TB, the severity of the HIV/TB co-epidemic, and poor living conditions in rapidly urbanizing societies all contribute to the persistence of TB in high-burden settings. These trends threaten to reverse recent, precious gains made in TB control and addressing these challenges will be central to combating the TB epidemic in coming years.

1.1.2 Multidrug and extensively drug-resistant (MDR and XDR) tuberculosis

In the 1940s, the discovery of the first anti-tuberculosis drugs and the development of multidrug therapy made tuberculosis cure possible for the first time. Though this development would revolutionize TB control, it would also lay the foundation for future challenges. Although multidrug therapy is effective in the majority of TB cases, it does not always eradicate *Mtb* infection. Failure to eradicate infection can lead to the development of drug-resistant disease. Indeed, the emergence of drug-resistant TB as a major public health threat was heralded by a severe epidemic in New York City in the early 1990s, prompting concerns about the potential for generalized epidemics of drug-resistant TB. [6] This concern would turn out to be well-founded: the true magnitude of the threat of drug-resistant TB would soon be fully realized in sub-Saharan Africa, Eastern Europe, and Southeast Asia, where epidemics would flourish in the absence of robust public health infrastructure and effective disease control measures.

Today, the World Health Organization (WHO) estimates that 580,000, or approximately one-fifth of all TB cases worldwide, are caused by TB that is resistant to first-line tuberculosis drugs isoniazid and rifampicin. These cases are termed multidrug-resistant, or MDR, TB. [1] As MDR strains became increasingly prevalent, additional, ‘second-line’ drugs were employed to treat MDR TB patients, causing further development of resistance to aminoglycosides (kanamycin, amikacin, capreomycin, and streptomycin) and fluoroquinolones (levofloxacin, moxifloxacin, gatifloxacin, and ofloxacin). [7, 8] TB strains resistant to at least two first-line and two classes of second-line drugs are termed extensively drug-resistant, or XDR, TB. Today, the WHO estimates that XDR TB cases make up approximately 10% of all MDR TB cases worldwide. [5]

1.1.3 Extensively drug-resistant (XDR) tuberculosis in South Africa

In the wake of the global HIV epidemic, drug-resistant TB successfully gained a foothold and spread rapidly in sub-Saharan Africa. Perhaps the most severely affected country was South Africa, where a dismal response by the South African government to the HIV/AIDS crisis left HIV largely uncontrolled for much of the 1990s. [9] The brewing HIV epidemic, combined with increasing circulation of drug-resistant TB strains, laid the foundation for a public health crisis. In 2005, the first cluster of XDR TB cases was identified and described in a hospital in the rural town of Tugela Ferry, South Africa. [10] By the of end 2015, only 11 years after it was first identified, XDR TB had been reported in 117 countries. [1, 11] Today, South Africa remains among the countries with the highest burdens of XDR TB. HIV co-infection among all drug-resistant TB cases in South Africa is 63%; among XDR TB cases, it exceeds 70%. [1, 12, 13] The province of KwaZulu-Natal, which was home to the 2005 Tugela Ferry outbreak, continues to have the highest incidence of XDR TB of any province in South Africa. [14]

Though XDR TB cases account for a relatively small proportion of the TB burden in South Africa and globally, the disease is significantly more difficult and costly to treat than drug-susceptible TB. In South Africa, the survival rate from XDR TB is 28%. [15] The drugs that are available to treat XDR TB are toxic and cause the majority of patients to experience severe side effects. [16] In addition to the physical costs of contracting XDR TB, disease and treatment can exact a devastating financial burden on patients. An estimated 20,000 drug-resistant TB patients experience catastrophic costs (defined as costs that exceed 20% of individual annual income) in South Africa each year, which can have profound effects on their lives and families even if patients are eventually cured of disease. [17, 18] XDR TB also represents a dire economic threat to already-overburdened health systems. Drug therapy for each XDR case in South Africa is estimated to cost over $26,000 to the healthcare system, and treatment for XDR TB cases consumes over two-thirds of the national budget for TB control despite accounting for fewer than 5% of all TB cases. [19, 20] In light of the overwhelmingly poor prognosis for XDR TB patients and the healthcare systems charged with managing them, prevention is at present the most powerful tool for reducing morbidity and mortality from XDR TB.

1.2 Natural history of tuberculosis

1.2.1 Overview

The clinical course of pulmonary tuberculosis is often divided in two stages: latent infection and active disease. The period of latent infection is one of the hallmarks of the natural history of TB, although its biology and epidemiology remain poorly understood. During this period, *Mtb* is sequestered by the host immune system and the infected person is typically asymptomatic. The duration of the latent period can be highly variable, ranging from weeks to decades, and the vast majority of infected persons never progress to active TB disease. Latently-infected individuals make up a vast reservoir of *Mtb*: it is estimated that as many as one-quarter of the global population is infected. [21, 22]

Five to fifteen percent (5-15%) of those infected will progress to active pulmonary TB disease within several years, and among those cases, the median incubation period is between several months and two years. [23, 24] TB patients with active disease typically present with clinical symptoms of night sweats, weight loss, fever, and a productive cough. [24] Although TB primarily infects the lungs, TB infection can occur in virtually any organ; TB infection in a location outside the lungs is referred to as extrapulmonary TB disease. Generally, persons with active, pulmonary TB are considered infectious and those with latent infection or with only extrapulmonary disease are not. However, it is worth noting that the classification of TB infection into two discrete stages is useful for many clinical and public health purposes but increasingly at odds with our understanding of the biology of TB infection. This is supported by a range of epidemiologic and biologic evidence that support the existence of a continuum of disease states: for example, the observation that some individuals can be culture- or smear-positive, and thus presumably infectious, prior to the onset of clinical disease. [25, 26]

Transmission of tuberculosis occurs via the airborne route: *Mtb* bacteria are released into the air in aerosolized droplet nuclei by infectious cases and inhaled by susceptible persons. The duration of infectiousness of TB cases has been estimated to be several months, with HIV+ cases typically exhibiting a shorter pre-diagnosis infectious period than HIV-negative cases. [27] Among undiagnosed cases in high burden settings, some estimates suggest that TB cases may be infectious for over a year prior to diagnosis and initiation of treatment. [28] Classically, smear-positive TB cases are thought to be more infectious than cases who have undetectable levels *Mtb* in sputum (‘smear-negative’). Bacterial load, and thereby infectiousness, is expected to decline after initiation of TB therapy.

Beyond this, our understanding of TB transmission is surprisingly limited. This can be partially attributed to the fact that transmission risk is modulated by a complex interaction of host, pathogen, and environmental influences, making it difficult to isolate and study the effect of a single factor. Indeed, this complexity is true of many infectious diseases and is not specific to TB. However, several features unique to the natural history of TB present challenges to studying transmission. First, the variable and often lengthy period of latent infection complicates efforts to infer timing of transmission and definitively identify sources of infection. Since tests for latent infection are generally poor and cannot distinguish between remote and recent, or resolved and current infection, the results of these tests are difficult to interpret. Second, mode of TB transmission, via droplet nuclei, makes it difficult to establish risk of TB exposure. Ideally, identifying exposure risk would involve a complete accounting of an individual’s history of sharing air with others. Of course, this is nearly impossible to measure accurately and completely. This is further complicated by the fact that droplet nuclei can remain suspended in air for days after they are expelled from the lungs, making it difficult to definitively identify interactions leading to transmission. Finally, the variable and often lengthy duration of infectiousness provides the opportunity for TB cases to expose countless others to disease, resulting in an intractably large pool of potential secondary cases. The period between latent disease and clinical illness is often called the 'subacute' or 'subclinical' period, during which an individual may be shedding bacteria (smear-positive) but feeling sufficiently well to maintain their daily routine. Increasingly, this period is recognized as potentially important for driving transmission. [25, 28]These features present serious challenges for the design of studies aiming to understand transmission dynamics, and existing studies should be interpreted cautiously in light of these important limitations.

The above limitations notwithstanding, our current understanding of TB transmission has benefited from a spectrum of scientific approaches. Experimental studies of TB transmission began in the 1950s, with Wells' and Riley's seminal experiments showing airborne transmission of *Mtb* from humans to guinea pigs in hospital TB wards. [29] (More recently, similar studies have updated these findings to include the impact of drug resistance and HIV infection on infection and transmission. [30, 31]) These experimental studies revealed an important principle of TB transmission that has been borne out in epidemiologic studies and will be a central focus of this dissertation: there appears to be significant heterogeneity in infectiousness, and therefore transmission, among cases of TB.

1.2.2 Factors influencing transmission

1.2.2.1 Transmission heterogeneity

Relative to other respiratory infectious diseases, tuberculosis is considered to be only moderately infectious. The reproduction number, often denoted R_0, is a fundamental property of an infectious disease and defined as the number of individuals a single case of disease is expected to infect in fully susceptible population. The reproduction number of TB is thought to range between 1 and 5, depending on the epidemiologic context. [32, 33] (For comparison, the reproduction number of measles is estimated to be 12-18, and influenza is estimated to be between 1 and 2. [34, 35]) However, these estimates represent the average of a distribution, reflecting the fact that some individuals will cause fewer secondary cases and others will cause more. Variation in the number of secondary cases per index case is often referred to as ‘transmission heterogeneity’. [36] Such heterogeneity is generated by differences across individuals with respect to several factors: an individual's duration and extent of infectiousness, the rate at which they contact susceptible individuals, and the intensity of such contacts. [37] Broadly speaking, individual reproduction numbers for TB tend to follow a right-skewed distribution with a long right tail: this is to say that while most cases of disease will cause approximately the expected number of secondary TB cases, a minority of cases will cause many more than expected. Individuals that cause an unexpectedly high number of secondary cases are often colloquially called ‘superspreaders’ and have been implicated as critically important in understanding the transmission dynamics of many diseases. [38-41]

Such large variation in individual reproduction numbers can be attributed to the unequal distribution of host, pathogen, and environmental characteristics that influence transmission. Defining the combination of clinical, demographic, and social factors that promote differences in transmission across individuals and groups is the first step towards understanding the complex interactions that give rise to the patterns by which disease spreads through populations. From a public health standpoint, explaining the roots of this heterogeneity can also provide the basis for targeting interventions towards groups and individuals responsible for a disproportionate amount of transmission, which offers opportunities to optimize use of scarce public health resources.

1.2.2.2 Infectiousness

‘Infectiousness’ refers to the likelihood that a TB case will spread tuberculosis to others. Biologically, this is defined by the quantity and viability of *Mtb* that an infectious person releases into the surrounding environment. Measuring infectiousness is complex: although sophisticated methods analyzing cough aerosols have shown that it is possible to measure the rate at which cases expel *Mtb*-containing particles, it is technically and methodologically challenging to measure the ability of such particles to cause infection in a susceptible host. [42-45] However, clinical characteristics of TB cases may partially reflect infectiousness and be useful for identifying cases who are more likely to transmit TB. Classically, untreated, smear-positive patients with cough and cavitary disease are thought to be most contagious, and this is generally supported by epidemiologic studies. [46] Given that these clinical features are easily measurable, combining them to create an overall measure of extent and duration of the infectious period represents a practical approach to estimating the infectiousness of a TB case.

The quantity of bacteria found in the sputum or in the lungs of a TB case, often referred to as bacillary burden, is considered one marker of infectiousness. The effect of high bacillary burden on infectiousness is twofold. First, cases with a high bacillary burden may aerosolize higher quantities of *Mtb*. Second, although infectiousness is typically thought to decline after treatment initiation, cases with a high bacillary burden may require a longer period on treatment until they are non-infectious. This may effectively increase their infectious period relative to cases with a lower bacillary burden. Several clinical tests can estimate bacillary burden in a TB patient. Smear microscopy is a microbiological technique that is performed by staining a sample of sputum with a dye that adheres to the mycobacterial cell wall. Among cases who test positive for mycobacteria, smear grade can be determined by counting the number of mycobacteria visible on the stained slide. Larger quantities of bacteria in sputum assigned a higher smear grade. These markers have previously been associated with transmission: positive smear status has previously been associated with being ‘genotypically-linked’, or harboring a similar TB strain, to at least one other sampled TB case. Similarly, smear-positive TB cases have been shown to be genotypically-linked to *higher* numbers of secondary cases than smear-negative cases. Among smear-positive cases, higher smear grade has been linked with incremental increases in the number of secondary cases caused by an index case. [47-50]

Bacillary burden can also be measured in the lungs using findings from chest radiograph. Chest radiograph can be used to detect areas of acutely damaged lung tissue, or ‘cavities’, indicative of high concentrations of bacteria. The presence of cavities is thought to increase transmission risk through improving the ability of *Mtb* in the lungs to access the respiratory tract, and indeed, presence of cavitary disease has been linked with a higher likelihood of transmission. [51] Because both tests measure the bacillary burden in different organs, combined measures of smear grade and cavitary disease provide a composite marker of infectiousness that captures both the quantity of bacteria and the time required until anti-TB therapy controls *Mtb* infection and the case is rendered non-infectious. Indeed, a recent study showed that a combined measure of smear grade and chest radiograph findings is linked with time to sputum conversion. [52]

The duration and nature of TB symptoms may also be related to infectiousness. Studies show that coughing produces more particles of the appropriate size and velocity for establishing infection than either talking or breathing. Thus, cough is considered the primary method of aerosolization and release of infectious *Mtb* particles. Recent work has shown that measures of cough aerosol production may identify individuals more likely to contribute to community transmission. [44, 45] While characteristics of the aerosol reflect the probability that a particular contact results in transmission, the duration of cough symptoms can represent the number of potential contacts affected. Some evidence suggests that length of the period during which a TB patient has cough may represent the time that they are most likely to be infectious. [53-56] Increased cough frequency, though not well-studied, has also been linked to higher numbers of secondary TB cases. [57] Collectively, these characteristics of cough may provide further insight into the infectiousness of a TB case.

This dissertation will examine smear status, chest x-ray, and cough duration as markers of transmission potential, though it is important to note several important limitations to this approach. Smear status and lung pathology may vary significantly over the course of TB disease, and therefore single chest x-rays or sputum samples cannot provide a complete picture of transmission over the full clinical course of TB. Moreover, the relationships between these clinical markers and infectiousness may be complex. For example, transmission from consistently smear-negative TB cases is well-documented, and there is evidence that cough aerosols may vary significantly in their capacity to cause infection. [25, 42, 43, 58] Moreover, it remains to be understood to what extent the quantity of *Mtb* in sputum reflects the amount of *Mtb* that is aerosolized, and of those *Mtb* particles, the proportion that is fit to cause infection in a susceptible host. Despite these limitations, information on patient symptoms combined with clinical tests may provide a reasonable picture of the infectiousness of a TB case.

1.2.2.3 HIV infection

The single most influential biologic factor shaping TB epidemiology, particularly in South Africa, is HIV infection. The primary effect of HIV infection on TB disease epidemiology is that it shortens the latent period: the immunosuppressive effects of HIV cause co-infected cases to progress rapidly to active TB disease. [59, 60] However, the link between HIV status and TB transmission is more complex. Patients with TB/HIV co-infection often have atypical clinical presentations of TB, such as extrapulmonary and smear-negative disease. [61-63] These clinical features are associated with low infectiousness and may reduce the likelihood of transmission. On the other hand, smear-negative patients are less likely to receive a prompt TB diagnosis. Delays in diagnosis could instead *increase* the likelihood of transmission, by extending the period during which a case is infectious. [61-64] Indeed, studies of the effect of HIV on transmission have yielded inconclusive results. Human-to-guinea pig transmission studies have shown that HIV-positive TB cases cause more secondary TB cases than HIV-negative cases. [31] (Notably, the validity of these results has been questioned. [65]) Observational studies have also examined the relationship between HIV status and transmission, where transmission is measured by determining whether each case has a TB genetic ‘fingerprint’ similar to that of another sampled case. These studies have produced mixed results, perhaps because they cannot distinguish between risk of TB infection, progression and transmission. [51, 66-69] Studies that aim to establish the *direction* of transmission can theoretically parse the effect of risk factors on infection and transmission. A study from Malawi that established directionality of transmission events found that HIV co-infection reduces the likelihood of transmission, and a recent modeling study from South America suggested that transmission from HIV-negative to HIV-positive individuals may be driving TB spread in areas of with high HIV burden. [69, 70] However, this same South American study suggested little, if any, association between HIV status and transmission. Taken together, the evidence suggests that HIV-positive TB cases may be less likely to transmit TB than HIV-negative cases, but that counterbalancing forces may lead to similar transmission potential between HIV-negative and HIV-positive TB cases.

1.2.2.4 Demographic factors

Demographic factors, such as age and sex, modulate transmission dynamics through influencing the biology of TB infection as well as the social mixing patterns that drive transmission. Although the biology of TB infection may not differ significantly by setting, age and sex-specific contact patterns are dependent upon sociocultural norms that drive patterns of person-to-person interactions, and therefore their effects on transmission may depend largely on the population of study. Recent work has aimed to characterize age- and sex-specific social contact patterns in areas with high incidence of TB in order to provide a better understanding of the types of contact that lead to transmission in these settings. [71-74]

Although their root causes are poorly understood, sex differences in TB incidence and prevalence are well-described. Crude TB prevalence is typically higher among men than women, though it is uncertain whether this is the result of increased rates of diagnosis or increased biologic susceptibility to infection or disease progression. Among diagnosed TB cases, men tend to have poorer treatment outcomes than women, which result in longer infectious periods and may increase the likelihood of transmission to secondary cases. [75] In low- and middle-income countries, men are more likely than women to exhibit social mixing patterns that promote transmission, including frequent congregation at bars and in other social settings conducive to transmission. Conversely, women tend to spend a larger share of their time at home, and thus may be at higher risk for contracting TB from a household member. [72, 75]

Age also influences both the biology of TB disease and the social contact patterns that drive transmission. Adults are more likely to transmit TB than children, since the typical clinical features of pediatric TB disease, including smear-negativity and lack of cough symptoms, makes forward transmission from pediatric TB cases less likely. [76] Instead, adults are thought to responsible for most TB transmission events: in fact, a recent study using social contact data in Zambia and South Africa estimated that over 50% of TB infections could be attributed to contact with adult men. [72] Age and sex-related risks for infection and transmission are closely linked with other behaviors, including smoking, drug, and alcohol use, which are also known to increase risk for TB. These factors may increase risk for disease progression, severity, and poor treatment outcomes, thus increasing the extent and duration of infectiousness. [77-79]

1.2.2.5 Pathogen genetic factors

The genetic makeup of *Mtb* can contribute to biologic differences in the processes of infection and transmission. Although not all genetic variation is important, genetic variation that gives rise to differences in the ability of *Mtb* to cause and propagate disease may affect transmission dynamics. On a population level, certain strains of TB have been shown to be more 'successful' than others: that is, they are associated with higher numbers of secondary cases, are more geographically widespread, or are more likely than other strains to cause disease. [80, 81] The hypothesis that these strains are more 'fit' is supported by laboratory evidence: links between specific genetic variants and pathogenicity or virulence have been described extensively in murine and rabbit models. [82-86] Ongoing research focused on the genetic underpinnings of *Mtb* airborne survival represent new and exciting frontiers in this area of research. [42, 43, 87]

However, testing whether associations between genetic makeup and transmissibility are causal on a population level is not straightforward. Linking observations of increased virulence *in vitro* with epidemiologic events of interest, such as particularly large TB outbreaks, can only generate hypotheses regarding whether certain strains are more transmissible than others. [86] Epidemiologic studies that have examined the effect of *Mtb* lineage and sublineage on whether an individual is involved in a transmission cluster suffer from methodological limitations; namely, it is difficult to disentangle the effects of host factors from those of pathogen lineage. [66, 67, 88] Studies linking certain TB genotypes with poorer response to treatment are subject to similar limitations. [89-91] Certainly, some TB strains appear to be well-adapted to a wide range of human populations, with evidence to suggest that this flexibility has genetic origins. For example, an extensive amount of laboratory and epidemiologic evidence suggests that

Lineage 2, or Beijing, strains, may have particular genetic characteristics that explain their modern epidemiologic success worldwide. [92]

The drug resistance profile of a TB strain has also been hypothesized to impact transmission potential. Until very recently, it was thought that drug-resistant TB strains were significantly less transmissible than drug-susceptible strains, due to fitness losses incurred through the process of developing drug resistance. [93, 94] Several lines of evidence support this theory: *in vitro* and epidemiologic studies have shown that specific mutations associated with drug resistance can reduce the growth rate, virulence, and transmissibility of *Mtb*. [95-100] However, that MDR and XDR TB strains have persisted over the past several decades is crude but clear evidence of their evolutionary success. (Of note, some partially attribute this phenomenon to the large reservoir of immunocompromised, and highly susceptible, individuals as a result of the HIV epidemic. [70, 101]) Indeed, recent studies have demonstrated that MDR and XDR TB strains are transmissible in both institutional and community settings, and some have suggested that they may be even more transmissible than drug-susceptible strains. [66, 102, 103] Recent studies of high-burden countries examining the proportion of XDR TB cases attributed to primary transmission, rather than inadequate treatment, suggest that as many as three-quarters of drug-resistant TB cases are the result of transmission. [12, 104] While it is difficult to establish the *relative* transmissibility of drug-resistant and drug-susceptible strains, it is expressly clear that transmission of drug-resistant TB plays major role in its epidemiology.

1.2.2.6 Social and environmental factors

Social and environmental factors, defined broadly, are those that influence the conditions in which exposure to *Mtb* occurs. The scale on which these factors act on transmission can range from the individual to the societal, and factors acting on different scales may influence one another. For example, a poorly ventilated and crowded household may be the result of community-level social structures that create and reinforce social inequalities. Both these macro- and micro-level environmental factors can give rise to favorable conditions for transmission.

Although the notion of TB as a 'social disease' was formalized by Rene Dubos in the 1950s, TB has long been recognized as a disease of poverty. [105] Today, TB remains concentrated among those with low social and economic status. Broader social conditions, including poverty and homelessness, are consistently associated with tuberculosis transmission in epidemiologic studies, acting through downstream factors like crowding, poor living conditions, and malnutrition. [66, 68] Social and economic factors driving transmission may take different forms depending upon the setting, but ultimately lead to similar impacts on risk of transmission and disease progression. A recent study found that among cities in high-income European countries, tuberculosis cases concentrated in areas where homelessness, poverty, migrants, overcrowding, and substance abuse were more common. [106, 107] In South Africa, India and in other low and middle-income countries experiencing accelerated economic development, rapid urbanization has drawn rural populations into the cities in massive numbers. The crowded, informal housing that supports these new occupants is fertile grounds for disease transmission. [108] In high- and low-income settings alike, social conditions are the driving force behind TB epidemics.

Larger social and economic forces can influence more proximal, physical environmental factors affecting risk of transmission. The local environmental profile of a particular setting– temperature, humidity, ventilation, and the level of UV light– plays an important role in mediating transmission between individuals, through modifying conditions that enable aerosolized *Mtb* to travel from infected to susceptible persons. [53] Perhaps the most compelling link between environmental conditions and transmission exists for ventilation. Crowded and poorly-ventilated institutional settings (in hospitals, homeless shelters, and prisons) are well-recognized as being conducive to transmission, and lack of ventilation may also explain higher TB incidence in colder areas, where individuals tend to spend more time indoors. [109-111] Interestingly, it may also account for the paradoxically low incidence of TB among the poorest segment of the population in South Africa, since they may be less likely to have glass windows and to meet indoors. [112, 113] Importantly, ventilation is often a modifiable factor in households, hospitals, or other locales with high risk for transmission, and improving ventilation can be a relatively inexpensive and effective method of preventing transmission. Similarly, increasing UV light has been shown to kill TB bacteria and is often useful, particularly in healthcare settings, to reduce risk of transmission. [114] Given that settings of transmission can be clearly identified, interventions targeting modifiable, proximal environmental factors (e.g., UV light, ventilation) in specific locations can be practical methods for location-specific TB control.

1.2.2.7 Nature of contact required for transmission

Identifying the types of contact that lead to TB transmission and the locations where it occurs is a critical step in the effort to design targeted interventions. In high- and low-TB incidence areas, households and institutional settings, such as hospitals and prisons, are often implicated in transmission. While these locations and the type of close contact that occurs in them are often prioritized in TB control efforts, increasing evidence points towards a large role for non-close contacts for transmission, especially in high-incidence areas.

'Close' contact, defined as contact that occurs in households or other locations where an individual spends substantial time with others whom they know, is an important source of TB transmission. Studies have shown that within households, the probability of TB transmission is high: it is estimated that a case of drug-resistant TB will infect approximately half of their household contacts, though this proportion is highly variable across settings. [115, 116] Though only about 10% of household contacts progress to TB disease, they and other close contacts are practical targets for prevention measures, since they are more easily enumerated than casual contacts. Moreover, household transmission disproportionately affects children, whose TB disease is typically more difficult to diagnose and treat. [76, 117] As such, preventing household transmission has long been a key element of TB control and remains the primary focus of the public health response to incident TB cases in low-incidence settings.

Institutional settings have also been implicated as key locations of TB transmission. In areas with low TB incidence, TB outbreaks are commonly linked with homeless shelters and prisons. [118, 119] In high-incidence settings, institutions can also play a critical role by creating reservoirs of disease that can drive incidence in the broader community. Indeed, the role of 'institutional amplifiers' in driving transmission in high-incidence settings is well-documented. [10, 120-123] In South Africa, studies of the initial XDR TB outbreak in Tugela Ferry suggested that cases who were infected while in the hospital may have contributed to community transmission. [103] Identifying and stopping transmission in these locations may be an efficient TB control strategy, as research has suggested that it can reduce incidence in both the setting itself and in the wider community. [124]

While the close contact that occurs in households and institutional settings may provide the most convenient targets for TB control measures, molecular epidemiology studies suggest that 'casual' (non-close) contact may account for a majority of transmission in high-incidence settings. In Malawi and South Africa, molecular epidemiology studies have estimated that household contacts account for less than 20% of transmission. [69, 125-127] In Malawi, these studies have shown that half of TB cases with another smear-positive case in their household had a different TB strain from that case, indicating their infection was acquired elsewhere. [126, 127] Importantly, interpretation of molecular evidence of transmission has limitations that must be carefully considered, and are subject to sampling biases that may partially explain 'missing' transmission links. [128] Even accounting for this source of error, the evidence suggests that a high proportion of transmission results from non-close contacts.

Other types of evidence further support an important role for casual, non-household contact in TB transmission, particularly in the high-incidence setting of South Africa. Social mixing studies in South African townships have suggested that public transport, schools and workplaces are likely settings for transmission. [71, 129] Studies employing environmental sampling approaches further support these hypotheses: recent modeling work using empirical data on rebreathed air fractions have estimated that over 80% of TB transmission in South Africa takes place outside of the home. [73, 129] Although this seems to contradict the conventional wisdom that close, intense contact is required for TB transmission, recent work has shown that this may be explained by the combined effect of a long infectious period and a high frequency of non-repeated contacts. [130] In other words, the cumulative number of casual interactions over a long infectious period represent a significantly larger pool of possible contacts, thereby accounting for the observation that a majority of transmission appears to occur through casual contact.

1.3 Measuring tuberculosis transmission

1.3.1 Epidemiologic (contact) investigation

Retrospective investigations of TB transmission employ several complimentary methods to reconstruct transmission events, one of which is epidemiologic, or contact investigation. These investigations typically involve an interview with the patient to elicit information on time of illness onset, symptoms, names of possible contacts, and locations at which contact leading to infection may have occurred. [131] Contacts and locations named by the case are then investigated to identify additional cases with latent infection or with active TB disease who can then be referred for treatment. Contact investigation is recommended by the WHO as one of several methods of active case finding, primarily because these investigations are generally high-yield: latent TB is detected in about 30% of household contacts in low incidence countries and 50% of household contacts in high incidence countries. [125, 132] These activities are valuable in establishing person-to-person interactions that may be responsible for transmission and can provide actionable information for public health authorities that helps identify additional cases and reduce further transmission.

However, there are several major concerns with relying solely on epidemiologic investigations to completely enumerate possible transmission links. First, many countries with a high TB burden do not have the resources to investigate contacts of every TB case, and routine contact investigation activities are therefore limited in most countries with high incidence of TB. Second, epidemiologic investigations have an important limitation in that they are subject to recall error by TB cases. A recent study that examined agreement in contact reporting found that fewer than 30% of identifiable contacts were reported by both persons involved in the event, indicating that some individuals report contacts in a more complete manner than others. [133] Unsurprisingly, casual contacts seem to be more prone to recall error than close contacts: short person-to-person encounters are less likely to be recalled than more extended encounters. [134] This may be of minor concern to public health authorities charged with TB control, since transmission may be more likely to occur to individuals with whom cases had more intense contact. In this sense, epidemiologic investigations maximize efficiency by capturing contacts with the highest likelihood of progressing to TB disease. However, in light of increasing evidence that casual contact may be an important driver of transmission, this bias is problematic when studying population-level transmission patterns.

In short, epidemiologic contact investigations, and the contacts enumerated through them, are more likely to reflect close contact rather than casual contact. For examining patterns of transmission, therefore, reliance upon these investigative methods alone to identify transmission links may reveal only a subset of transmission events.

1.3.2 Molecular genotyping

Molecular epidemiology offers a complementary method to contact investigation for uncovering transmission links. The premise of molecular epidemiology studies of transmission is that TB strains sampled from cases between whom transmission occurred should be more similar than strains sampled from cases between whom transmission did not occur. In other words, *Mtb* collected from epidemiologically-related cases should have identical, or at least similar, genetic ‘fingerprints’. Changes in the *Mtb* genome occur slowly, but at epidemiologically relevant timescales (months to years), allowing these genetic changes to provide information about the likelihood that cases may be linked through transmission. Several classes of repetitive elements in the *Mtb* genome have been exploited to develop assays to detect these molecular ‘fingerprints’ in *Mtb* sampled from TB cases. These methods, reviewed elsewhere, include restriction fragment length polymorphism (RFLP) typing, mycobacterial interspersed repetitive unit-variable number tandem repeat (MIRU-VNTR) typing, and spacer oligonucleotide, or spoligotyping, and they have all been applied extensively in epidemiologic studies. [135] Typically, cases with the same or very similar genetic fingerprints are classified as ‘clustered’, or linked through recent transmission. Cases with genotypes that differ from other cases’ are termed ‘unique’and are assumed to have developed disease due to reactivation of a TB infection acquired in the past. An important caveat to this construct is that unique cases may in fact be the result of transmission if the source case was not sampled.

While these genotyping techniques are widely used to characterize TB epidemiology and detect ongoing TB transmission, they have several important limitations. First, these techniques only examine a small proportion of the genome (<10%) and thus do not represent potentially important genetic variation that exist in other parts of the *Mtb* genome. Second, these methods are limited by the discriminatory power of the finest classification in their respective typing schemes. When the determination of whether a case is 'genotypically linked' relies upon a typing scheme that provides only a coarse, discrete measurement of genetic similarity, inferences regarding transmission links are heavily reliant upon the proportion of TB cases sampled. [128, 136] Given the same true number of genotypically linked cases in a TB outbreak, the fewer cases that are sampled, the lower the proportion of genotypically linked cases appears. Continuous measures of genetic similarity that examine a larger fraction of the *Mtb* genome are less sensitive to these methodologic limitations.

1.3.3 Whole genome sequencing

With the advent of next generation sequencing platforms, bacterial whole genome sequencing (WGS) has recently emerged as a tool with unprecedented ability to identify fine-scale genetic differences between TB strains. In comparison to conventional genotyping techniques, WGS examines a larger portion of the *Mtb* genome at a finer resolution, creating a more complete picture of genomic variation between bacterial organisms sampled from TB cases. WGS has successfully been used to resolve recent transmission events in the context of TB outbreaks [137, 138], as well as characterize broader transmission patterns over extended time periods [67, 127, 139, 140]. However, it has been employed primarily to confirm or refute previously identified epidemiologic links between cases or to further differentiate members of a cluster detected through conventional genotyping, rather than as a stand-alone method to identify transmission links. [141] Using sequencing approaches to generate hypotheses about transmission events may be an efficient way to fully capture transmission links due to both close and casual contact, since molecular methods are unprejudiced by patient report of contacts. Given the limitations of epidemiologic investigations, particularly in high-incidence settings, and improved capabilities for resolving transmission links afforded by WGS, new techniques for investigating transmission that rely on sequencing data represent a promising way forward for studying TB transmission.

It is instructive when interpreting *Mtb* sequencing data to define the genomic epidemiology of *Mtb* more broadly. Generally, *Mtb* is a slow-growing organism that is relatively resistant to the horizontal gene exchange that accounts for much of the genetic variation in other bacterial species. [142] As a result, almost all genetic variance in *Mtb* populations is the result of genetic changes arising from errors during genome replication, or *de novo* mutations, in the DNA sequence. These mutations arise relatively slowly in *Mtb* as compared to other bacteria: the mutation rate of *Mtb* is estimated to be approximately 0.2-0.5 SNPs/year. [143] Collectively, these characteristics give rise to largely monomorphic, or genetically uniform, *Mtb* populations. [144, 145] However, the long history of *Mtb* as a human pathogen has given modern *Mtb* a strong phylogeographic structure. Phylogenetic classification schemes split *Mtb* into seven lineages, six modern and one ancient, each named for its world region of origin. [146] Though this geographic specificity still persists, increasing human movement and the recent widespread epidemiologic success of certain lineages have diversified the modern genomic epidemiology of *Mtb*. [147] Today, all seven lineages of *Mtb* can be found almost anywhere in the world; however, local epidemics tend to maintain their own, unique genomic epidemiology. In New York City, for example, Lineage 4, or Euro-American, strains predominate. [148] In China, Lineage 2 strains, including the globally prevalent Beijing strain, are most common. Among XDR strains in KwaZulu-Natal, South Africa, a specific Lineage 4 strain named KZN/LAM4 accounts for over 80% of cases. [149, 150] With only several key exceptions, the genomic epidemiology of TB in high-incidence settings tends to be highly clonal. [149]

Despite its slow mutation rate and exceptional clonality, *Mtb* evolves at a rate that is sufficiently high to detect epidemiologically-relevant genetic variation, or genetic variation between *Mtb* sampled from cases linked through transmission. However, the exact genetic distance indicative of transmission between two TB cases is not clearly defined. It has been suggested that the threshold may vary depending on the local epidemiologic context, since the dynamics of *Mtb* infection may be influenced by setting-specific host and pathogen factors. [151] In general, our understanding of the genetic changes that occur over the course of TB infection and disease is limited. Estimations of *Mtb* mutation rates generally converge to an estimate of 0.2- 0.5 SNPs/yr, although there remains little certainty regarding the extent to which these estimates are appropriate for the purposes of making inference about individual transmission events. [67, 152-154] This particular issue will be discussed further in later chapters of this dissertation.

1.3.3.1 Other applications of whole genome sequencing in tuberculosis

In addition to providing insights about transmission, whole genome sequencing has the potential to revolutionize other aspects of the clinical and public health response to TB. Current methods for diagnosis and detection of drug resistance in many low-income countries continue to rely heavily on bacterial culture. Although sensitive, a culture diagnosis can take up to two months and delay initiation of effective treatment. Next generation sequencing technologies hold promise to dramatically improve the speed of TB diagnosis through the development of new diagnostic tools for sequencing bacteria directly from sputum samples, effectively eliminating the lag time between presentation at the clinic and a TB diagnosis. Moreover, the availability of sequencing results at the time of diagnosis would also allow clinicians to determine the drug resistance profile of the infecting strain and tailor a drug regimen accordingly. [155-160] Although sequencing-based TB diagnosis and drug susceptibility testing directly from sputum is not yet feasible, recent progress in these areas has prompted predictions that TB may be the first pathogen for which a ‘complete genomic approach’ may be implemented. [155, 156]

Whether the lofty promises of NGS for diagnosis, treatment, and control of TB will bear out remains to be seen. However, it is clear that the quantity and quality of sequencing data collected through routine TB surveillance activities will continue to increase. Several high-income countries, including the US and the UK, are in various stages of implementing universal whole genome sequencing for all TB cases, and in South Africa, the national laboratory service has begun to sequence isolates from all drug-resistant TB cases. As more countries follow suit, sequencing data will increasingly be available as a possible tool for understanding the local epidemiology and drivers of TB transmission, and TB control programs will be charged with using this data to inform TB prevention and control activities. The evaluation of systematic approaches to use WGS data to provide actionable information on TB transmission will be necessary to facilitate integration of WGS data into routine TB control activities.

1.4 Network approaches for studying infectious disease transmission

1.4.1 Overview

Networks and network models are a useful tool for representing and analyzing relationships between actors in a system of related components. Network science draws from concepts in mathematical graph theory and has been applied in scientific disciplines ranging from neuroscience to ecology, to represent biological processes from the cellular level up to that of entire human communities. The application of these models in a wide range of contexts reflects the fundamental concept that biological systems, even very complex ones, are not arranged randomly. Instead, biological processes occur according to proscribed rules, and therefore it is possible to theorize, model, and predict relationships between the entities, or 'actors' that make up a network. [161] A key conceptual and statistical advantage of analytic methods designed expressly for networks is that they acknowledge dependencies between network actors, in contrast to conventional modeling methods that typically assume independence of all subjects under study. [162] The field of epidemiology has been a particularly enthusiastic adopter of network analytic methods, because their relational nature easily lends itself to making inferences about infectious disease transmission. [161-165]

Networks and their associated analytic methods can provide insight into a range of epidemiological parameters of interest in infectious disease dynamics. Network analyses can characterize drivers of transmission by identifying disease 'superspreaders' and revealing the types of relationships and social mixing patterns important for disease spread. [166-168] They have been used to inform and evaluate disease control activities, by guiding the targeting of interventions towards specific individuals or groups [169-171] and assessing intervention effectiveness. [172] Comparative analysis of networks has been used to track the progression of epidemics over time [173, 174] and to characterize the structural features unique to the transmission networks of different diseases. [175, 176] Network models can be used to forecast future disease trends: either to predict trajectories of outbreaks [177] or to compare putative patterns of disease spread under different intervention scenarios. [178] In short, different types of network models and methods have been applied to a wide range of epidemiologic questions and successfully adapted to different pathogens, settings, and research questions.

The motivation for constructing a network dictates who, or what, is represented by network nodes and the manner in which the data used to create the network should be collected. For example, understanding determinants of the spread of a disease is well-suited to a network composed of social relationships between diseased and susceptible persons in a population. A network such as this could identify factors influencing susceptibility to infection or the likelihood of transmission. Alternatively, a network aiming to identify cases responsible for the most number of transmission events may only include cases of disease and the other cases whom they infected. [179] Studies that make use of compartmental models to study disease dynamics may require only data on general population structure and characteristics to model networks that are representative of artificial populations.

In addition to choices about *who* should be sampled, the method of sampling also has important implications for network analysis and inference. [180] Egocentric sampling, in which a random sample of the population at-risk for disease is surveyed about their contacts, is an attractive option for data collection because, given a clearly-defined sampling frame, it can be relatively simple and cost-effective. The analytic methods appropriate to analyze networks constructed through egocentric sampling are well-characterized and straightforward. [181] Adaptive sampling is another method of recruiting individuals that make up a network, and is used primarily to capture harder-to-reach populations that may be difficult to randomly sample. A common type of adaptive sampling is respondent-driven sampling, in which the first sampled individual is used to ‘seed’ a network, and the network is constructed outwards from that individuals’ contacts and the contacts of their contacts. [182] This type of sampling requires more careful statistical analysis, since there are often dependencies in the manner that cases are sampled which must be accounted for in order to make inferences about a target population. [183]

Once an appropriate set of network members, or nodes, is identified, information on interactions between them determines the construction of links, or edges, in the network. The transmission route of a pathogen determines the type of interactions that constitute links between individuals in the network. For respiratory pathogens, which are passed from person to person through airborne transmission, links may represent any situation that results in two individuals sharing air, which could include shared households, workplaces, classrooms, or transit routes. For HIV and other sexually-transmitted diseases, links may indicate sexual encounters. For pathogens with environmental drivers, such as mosquito-borne diseases, a network might be embedded in space, with links that represent geographic distances between locations. [184] Still other types of networks may not represent directly measured contacts between individuals; rather, information that *implies* links may be used to construct the network. For example, if two individuals both harbor a pathogen with a similar genetic makeup, this might strongly suggest a transmission event occurred between them, even if there is no direct evidence of a specific physical interaction. [185] Edges in a network may be considered static or to vary with time: recent extensions of network models have addressed the dynamic nature of actor relationships by explicitly modeling the formation and dissolution of relationships between individuals in the network. [178]

The unique arrangement of these two fundamental elements - nodes and edges- define the structure of a network. The network models that I will discuss in the following two sections and will use in this dissertation aim to represent the generative processes that give rise to the structure of a transmission network comprised of TB cases.

1.4.2 Why model networks?

A modeling approach to network analysis reflects assumptions about the nature of the network and the processes that occurred to create it. Specifically, the benefits of building a model of a network, rather than simply interrogating the features of an empirical network, are twofold. [186] First, as with any mathematical model, it allows us to recognize that the processes giving rise to a network are regular, if stochastic, in nature. Defining the model allows us to simultaneously establish rules that govern the formation of connections in the network and acknowledges that these processes, as well as the parameters estimated from our hypothesized model of the network, are subject to a degree of uncertainty. Second, statistical models of networks, as with other mathematical models, can answer questions that extend beyond the observed network. For example, a well-fitting model can be used to simulate networks with similar, or slightly different, features in order to understand the potential impact of perturbations in the system giving rise to network. The value of specifying these models using empiric data is that they are grounded in real-world, if imperfectly-measured, observations and processes.

As discussed previously, the structure of network data is fundamentally different from that of conventional epidemiologic data. The unit of analysis of interest in a network is often relationships between individuals, in contrast to conventional epidemiologic analysis which most often examines relationships between two or more traits within individuals. As a result, many of the tools for statistical inference that are used in conventional epidemiology are not appropriate for networks, since these methods are predicated on the assumption that each unit of analysis, or subject, is independent from every other. In a network, if Case A is connected with Case B and Case A is also connected with Case C, it may not be true that each connection is independent, since Case A is involved in two of them. (For example, if Case A and B know one another, Case C might be more likely to be connected to both of them rather than to only Case A or Case B. Thus, the connections of Case C are not independent of the connections of Cases A and B). Using standard epidemiologic methods to analyze these relationships may result in incorrect estimation of both point estimates and their associated standard errors, in much the same manner that a naïve analysis of correlated longitudinal data or clustered data may result in incorrect inference. As a result, it is important to either explicitly model, or to at least account for, these dependencies by using network methods that are uniquely suited for relational data.

1.4.3 Exponential random graph models (ERGMs)

Exponential random graph models, or ERGMs, are a class of network models that incorporate node and edge attributes as well as small-scale structural features to answer questions about the generative processes that give rise to relationships between actors in a network.

A brief history of the development of ERGMs illustrates the iterative additions of important features that make them an attractive option for modeling networks. In 1981, Holland and Leinhardt first proposed p* models, which are log-linear models that describe the formation of a tie between two actors as a random variable. [187] P* models (like ERGMs) utilize an exponential, or log, link to describe the probability of a tie between actors in a graph as a random variable. However, these early network models were subject to the unrealistic assumption that dyads, or pairs of nodes, were independent of one another. To address this limitation, Markov random graph models, a special class of p* models, were proposed in 1986 by Frank and Strauss, who were able to eliminate the dyadic dependence assumption. [188] Strauss and Ikeda extended this model to include methods for estimation of model parameters using pseudolikelihoods; these methods were later replaced by simulation approaches for estimating posterior parameter distributions, namely Monte Carlo estimation, which produces parameter estimates with more straightforward statistical properties than those produced by pseudolikelihood methods [189-191] Modern ERGMs, then, have several convenient statistical properties. First, they can be specified in log-linear form; second, they are not subject to the assumption that network actors and dyads of network actors are independent; and third, parameter estimates can be obtained using relatively simple Markov Chain Monte Carlo (MCMC) methods.

An ERGM specifies the probability distribution for a set of random graphs or networks. The probability of observing a specific network is a function of key predictors of network structure. These predictors can describe network structural features, for example, the total number of edges in the network or the number of isolates (unconnected nodes). They can also explicitly express local dependencies among nodes, for example, the tendency of actors to form ‘triangles’, or three-way relationships between 3 nodes. Actor attributes may also be incorporated into the model, if one hypothesizes that certain actors may preferentially interact with other actors based on their characteristics. This could be expressed by specifying terms for particular dyadic (two-node) configurations. After the desired model terms are specified, the resulting ERGM represents the distribution of possible networks given the structural constraints defined in the model. [186] The general form of an ERGM is given in Eq. 1:

Eq. 1. $\Pr(Y = y) \equiv (1/\kappa) \exp\{\Sigma_A \eta_A g_A(y)\}$

$g_A(y)$ = any possible network statistic, where A indexes the multiple statistics included in a model vector $g(y)$

η_A = coefficients for model terms; their value reflects the change in the conditional log odds of a tie for each unit increase in g_A that the tie would create

κ = normalizing constant, the sum of $\exp\{\Sigma_A \eta_A g_A(y)\}$ over all possible networks with n actors

1.4.4 Selection (sampling) bias in network models

While transmission networks may hold great promise for understanding disease dynamics, it is rare to observe the complete set of all actors and connections in a network. To the extent that all epidemiologic bias can be considered as missing data, this simply means that network studies are subject to the same biases as conventional epidemiologic studies. Networks can suffer from misclassification bias, as a result of non-response or inaccurate responses of participants, and selection bias, if the method of sampling participants is ill-suited to the research question or population of interest. A missing data consideration unique to network studies, however, is the concept of the network boundary problem, which is the challenge of defining which individuals and ties should be eligible for inclusion in a network. [192] Given the potential for bias, it is prudent to consider any inference made from a transmission network in light of the potential for missing data.

Consideration of the effects of missing data is particularly important when using networks to study an endemic infectious disease. Epidemics that occur over a clearly defined time period involving a finite number of cases can often be sampled nearly in full, obviating the network boundary issue.

However, endemic diseases, especially those with long latent periods, give rise to transmission networks that can involve connections between individuals that occurred months or years in the past, making missing data a virtual certainty. Moreover, the ongoing nature of an endemic disease means that cases who are linked to individuals prior to the start or after the end of the sampling period are not represented, resulting in a sample that is, temporally speaking, ‘truncated’ on both ends. A wide geographic range of cases, as is often the case with endemic diseases, can also complicate complete sampling. This can result in a failure to capture cases and links crossing arbitrary geographic borders, especially since surveillance data is most commonly collected and compiled on the level of administrative units that may not effectively communicate with one another. Although not specific to endemic disease, deficiencies in diagnosing cases can also contribute to the incompleteness of the network.

Missing data can represent a significant barrier to making inference using network models, and as a result, understanding the implications of missing data is increasingly a focus in the field of network analysis. Part of the challenge in clearly identifying the consequences of missing data in networks results from the interdependencies in a network: missing actors and ties may change the position of other elements in the network. Missing even very few actors could change network structure dramatically, especially if those actors are highly connected in the network. Recent work has shown that these effects may, in certain cases, be less severe than feared. For example, studies have shown that some network measures, including the mean degree, the number of edges, and the number of nodes can all be scaled by a factor of the sampling fraction to estimate their values in a larger network, as long as nodes of the larger network are missing at random. [193] However, it is important to note that not all network statistics are affected in the same way by missing data: some are more robust to missing network elements than others. [194-197] This should be considered in any attempt to make inference about the structure of a network when a significant portion of nodes is missing.

Unsurprisingly, the effect of missingness in a network also depend on the overall size and characteristics of the network. Larger, more centralized networks with positively-skewed degree distributions and higher clustering have been shown to be less affected by missing data. [195, 198]

Characteristics of missing nodes are also important: highly central missing nodes can have a pronounced effect on certain network measures, but can have little to no effect on others; missing nodes that are less central have been shown to result in few, if any, network statistics being severely affected. [198] Of course, the extent of missing data is also relevant for understanding the implications of missingness. Generally, the less data that are missing in a network, the less severe the measurement bias for a range of measures. [196-198] Collectively, studies on missingness in networks indicate that the effects of missing data can vary substantially, suggesting that a thoughtful consideration of what, and how much, data is missing is a critical component of conducting and interpreting a network analysis.

1.5 Transmission networks in tuberculosis

To date, the application of networks and network modeling to the transmission of tuberculosis has been limited, and as a result, its potential utility in providing insights into TB epidemics has been largely unexplored. Many studies have mapped putative transmission connections between cases, using social contact, molecular data, or both. However, systematic analysis of these networks has been notably absent in TB, even as it has gained in popularity in studies of other infectious diseases. The previously described challenges of identifying a complete set of potential contacts relevant for TB transmission, as well as the difficulty in establishing the timing and source of TB infection, might be reasons for the lack of research into TB transmission networks. However, in light of the increasing availability of whole genome sequencing to more reliably establish transmission links between individuals, network analysis may hold new promise for identifying transmission patterns.

Most studies utilizing networks to represent TB transmission in low-incidence settings have done so primarily to visualize, rather than analyze, putative transmission links between cases. In low-incidence settings, which typically experience small outbreaks concentrated in time and space, mapping of transmission links is a relatively straightforward task. Indeed, it is often employed to some degree in the public health response to an outbreak. Gardy et al. mapped transmission events in an outbreak of 41 cases in British Columbia using social contact data. [137] Their analysis of the social network generated from this contact data was limited to determining the most probable source of each TB case and identifying 'superspreaders', or cases who appeared to be the source for the highest number of secondary cases. A more recent study of a 14-year outbreak of isoniazid-resistant TB in London used networks in a similar manner to show putative transmission links between 344 cases, with the degree of certainty of the link based on agreement of epidemiologic and whole genome sequencing information. [140] Both studies enjoyed the benefit of near-complete capture of TB cases involved in the outbreak.

Several recent studies have attempted the significantly more challenging task of constructing transmission networks in settings of high TB incidence. Guerra-Assuncao et al. identified the most likely source for 1471 TB cases in Malawi diagnosed over a 14-year period, using an algorithm minimizing genomic distance between source-secondary case pairs. [67] Other studies have used network-like structures to map specific sub-clusters of cases with genomic evidence of transmission, and to identify 'superspreaders', or cases who appear to be connected to a relatively large number of other cases with similar TB strains. [104, 199-201] Several studies on drug-resistant TB in South Africa have also used networks to map cases and visualize transmission clusters. One of these was the same study which provides the data for this dissertation. This study described epidemiologic links of 404 XDR TB cases in KwaZulu-Natal province, South Africa over a four-year period, and in particular mapped an 'illustrative' clusters showing cases with epidemiologic and genomic evidence of transmission. [12] Another recent study in Cape Town identified and mapped 16 small clusters associated with 141 untreatable XDR TB cases discharged from the hospital back into the community, to assess the likelihood of transmission events occurring after patients with incurable disease were released from care. [202] The use of networks in these studies served the primary purpose of visualizing potential transmission events, rather than defining the structure or properties of a transmission network comprised of TB cases.

To our knowledge, there are only three studies employing an analytic approach to assess TB transmission networks; all three studies were focused on TB outbreaks in the United States. The first study constructed a network based in person-to-person and location-based links, and used network metrics to indicate key persons and locations driving transmission. [168] The second and third studies were similar, and both showed that individuals occupying a more central position in a social network that included active TB cases were more likely to be TST-positive. [203, 204] (The social networks constructed in these two studies included both TB cases and their disease-free contacts.) Studies using network analysis have been restricted to low TB incidence settings, and have not yet been leveraged in areas with a high incidence of TB, in which there is an urgent need to identify drivers of transmission and design interventions targeted towards groups and locations most responsible for transmission. Further, these studies have used conventional molecular typing tools to define transmission. Network analysis has been notably absent from more recent studies using new molecular tools, such as whole genome sequencing, to identify transmission links. Combining the discrimination of whole genome sequencing approaches for identifying transmission events with statistical methods uniquely suited to network analysis may provide new insights into TB transmission patterns in settings with a high incidence of TB.

1.6 Sequencing-based networks, previous work and potential applications in TB

Although analysis of networks based primarily on genomic sequence data has not been explored in TB, they have been used in other infectious diseases to gain meaningful insight into disease epidemiology. HIV epidemiology in particular has embraced the use of sequencing data to better understand disease transmission dynamics in a wide range of populations. HIV sequence data is highly informative on epidemiological time scales due to the rapid evolutionary rate of HIV. Although *Mtb* exhibits comparatively slower rates of evolution, many of the approaches and techniques that have been applied to analyze HIV sequencing-based networks, with some key modifications, may also provide insight into TB epidemiology.

Networks created using HIV sequence data have been used to answer a wide range of scientific questions. They have been used to estimate basic epidemic parameters, including the distribution of sexual partners among HIV cases in the UK [205], and identify secular trends in transmission across decades. [206] HIV sequencing-based networks have been used to characterize the size of disease clusters and to estimate and predict rates of cluster growth over time. [207-209] Other studies have used sequencing-based networks to investigate the impact of sequence variants that may directly influence HIV fitness and transmission. [210] Lastly, they have successfully been used to link case characteristics with network position, in a similar manner to that proposed in this dissertation: a study in Italy that combined HIV sequence data with sexual contact data showed that more highly connected persons in a sequencing-based network reported longer untreated periods. [211]

While the evolutionary features of *Mtb* may present some challenges for using sequence data to define transmission events, there are also several strengths to using sequencing-based approaches in TB. First, as previously mentioned, *Mtb*, unlike most other bacterial species, undergoes relatively little horizontal gene transfer. As a result, nearly all genomic variation in *Mtb* arises from *de novo* mutation. This allows genetic distances to be used as reliable proxy for evolutionary distances, without the need for complicated analyses to account for additional sources of genomic variation. Second, while epidemiologic investigations are biased towards transmissions that are due to close contact, sequencing-based approaches to identifying transmission links can capture transmission events due to both close and casual contact. This is particularly important in settings with a high burden of TB, where casual contact is estimated to account for a majority of transmission events. Third, sequencing approaches provide a continuous measure of genomic similarity among isolates, which allows for sensitivity analyses in which the ‘threshold’ of genetic similarity used to define transmission can be modified. In contrast, traditional genotyping methods provide only a dichotomous measure of relatedness.

Nonetheless, sequencing-based approached have important limitations. Uncertainty still exists about the relative mutation rates of *Mtb* during latent infection and active disease, and this can complicate inference when using exclusively genomic data to define transmission events. Complementing genomic data with epidemiologic information can provide an understanding of epidemiologic context in which an isolate of *Mtb* was sampled, as has been shown in applications to HIV. [209, 212] For example, using dates of diagnosis to suggest the directionality of transmission events suggested by genomic information may be a useful tool for constructing sequencing-based networks that can be used to examine individual-level factors related to transmission. Devising innovative ways to combine genomic and epidemiologic data may prove to be a powerful approach to furthering our understanding of TB epidemiology in high-burden settings.

Motivated by the successful application of sequence-based networks in HIV, we will use the same general principles to construct sequencing-based networks using *Mtb* sequences from cases of XDR TB in South Africa. Extending and adapting these methods to the unique features and challenges of the natural history of TB is a novel and exciting approach to interrogating the underlying structure and dynamics of TB epidemics. This dissertation will aim to create sequencing-based networks of TB transmission in order to better understand individual-level factors driving TB transmission.

2. Rationale and Specific Aims

Tuberculosis (TB) is the leading infectious cause of disease worldwide, with 9 million total cases and nearly half a million cases of drug-resistant TB reported in 2015. Recent studies show that transmission plays a critical role in the spread of extensively drug-resistant (XDR) TB in South Africa and globally, underscoring the importance of preventing transmission to reduce morbidity and mortality from TB. Limited public health resources in settings with the highest burdens of TB necessitate careful design and prioritization of interventions; however, little information is available to guide targeting of transmission interventions. Though previous studies have examined individual-level clinical factors that may influence potential for transmission, these studies have important limitations. First, they employ analytic methods that do not account for relationships between cases linked through transmission. Second, they use molecular typing techniques that examine an exceedingly small proportion (<1%) of the TB genome, though recent advances in sequencing have significantly advanced our ability to characterize TB genomic variation. In addition to clinical factors, behavioral patterns of TB cases provide information about settings and activities permissive of transmission. To date, very few studies have examined social mixing patterns linked with transmission, even though this information has important implications for intervention design. Further, no studies have addressed the potential impacts of unsampled cases on studying TB transmission networks, which is critical to ensure that an incomplete transmission network can accurately reflect underlying TB transmission patterns.

To fill these gaps, the proposed research will investigate clinical and behavioral drivers of XDR TB transmission by combining network analysis methods and bacterial whole-genome sequencing to identify transmission events. Although relational network data violates statistical assumptions required for standard regression, network models provide a statistically valid alternative to examine associations in is network. In Aim 1, we will use network models to examine the role of clinical features, including sputum smear status and grade, cavitary disease, and cough duration, in transmission of XDR TB. In the preliminary manuscript, we detail work on the geospatial patterns associated with XDR TB transmission. The studies were critical in shaping the research questions posed in Aim 2, by suggesting that urban exposure may be a key driver of XDR TB transmission in South Africa. In Aim 2, we will use network models to examine three domains of social mixing – close contact, casual contact (represented by contact with urban areas), and hospital exposure - and their association with transmission. In Aim 3, we will use exponential random graph models (ERGMs) to assess the impact of missing XDR TB cases in our transmission network based on different assumptions about potential mechanisms of selection bias. Through characterizing XDR TB transmission patterns in settings of high TB incidence, we can inform prevention efforts where the benefits of interrupting transmission are greatest.

2.1 Specific Aims

Aim 1: Measure associations between clinical markers of infectiousness of XDR TB disease and transmission in the network.

Hypothesis: *Higher infectiousness, represented by a combining smear-positivity, cavitary disease, and cough duration into a cumulative score, will be associated with high centrality in the sequencing-based network.*

Aim 2: Measure associations between reported social mixing patterns and transmission in the network.

Aim 2a: Measure the association between reported number of close contacts and transmission.

Aim 2b: Measure the association between casual contact (contact with urban areas) and transmission.

Aim 2c: Measure the association between time spent in hospitals and transmission.

Hypothesis: *We hypothesize that contact with urban areas, through increasing likelihood of frequent, casual contact, is driving transmission in KwaZulu-Natal. Time spent in urban areas, reflective of casual contact, will be associated with high centrality (Aim 3b), whereas reported number of close contacts and hospital exposure, which account for relatively less transmission than casual contact, will be weakly associated with network centrality (Aims 3a and 3c).*

Aim 3: Characterize a transmission network of XDR TB cases and estimate the effects of missing data on the network.

Hypothesis: Structural features between the observed TRAX and simulated networks will be grossly different under selection bias mechanisms in which unsampled cases are much more likely to transmit TB, but not if unsampled cases are unlikely to transmit TB.

3. Data source

3.1 Transmission of HIV-associated XDR TB in South Africa (TRAX)

The Transmission of HIV-associated XDR TB in Rural South Africa (TRAX) study is a cross-sectional study of 404 XDR TB cases in KwaZulu-Natal, South Africa diagnosed from 2011 to 2014. The primary aim was to quantify the proportion of XDR TB cases resulting from transmission as compared to amplification of resistance through treatment. Cases were recruited by identifying isolates meeting the criteria for XDR TB at the provincial diagnostic laboratory and contacting cases to whom the isolates belonged. Consenting patients (next-of-kin for deceased patients) underwent an interview in which they reported demographic information and completed a social network questionnaire, which involved reporting number and type of close social contacts, congregate locations frequented, and inpatient hospital stays prior to XDR TB diagnosis. Clinical information on TB disease was provided by patients and supplemented with information from the medical record.

Mtb was isolated from patient sputum samples, re-cultured and DNA was extracted. Isolates were sequenced on the Illumina (MiSeq) platform, aligned to the H37Rv reference genome (NC_000962.2) using the Burrows-Wheeler Aligner [213], and single nucleotide polymorphisms (SNPs) were detected using pairwise resequencing techniques against the reference using Samtools v0.1.19. [214]

3.2 Study Population

Cases were enrolled from KwaZulu-Natal, a province of 10.3 million people in eastern South Africa that is predominantly rural but home to the city of Durban, which has a population of about 4 million. South Africa generally [215] and KZN specifically [216] have high burdens of both HIV (prevalence of 16.9% in South Africa) and TB (1076 per 100,000 in South Africa).

The TRAX study enrolled 404 XDR TB cases, who accounted for 38% of all XDR TB cases diagnosed in KZN from 2011 to 2014. For this study, eligible cases were those that had a sequenced XDR TB isolate. 396 (98%) cases in the TRAX study had isolates available for whole genome sequencing and 344 (87% of 396) passed all sequencing quality filters. For XDR cases diagnosed in KZN but not recruited into TRAX (n=659), information on case age, sex, and facility of diagnosis is available through the provincial TB laboratory. Based on these characteristics, enrolled cases were not significantly different from diagnosed but unenrolled cases.

4. Preliminary Work

4.1. Preliminary manuscript

[This chapter appears as accepted for publication in the Journal for Infectious Diseases.]

CITATION: Nelson KN, Shah NS, Mathema B, Ismail N, Brust JCM, Brown TS, Auld SC, Omar SV, Morris N, Campbell A, Allana S, Moodley P, Mlisana K, Gandhi NR. Spatial patterns of extensively drug-resistant tuberculosis (XDR TB) transmission in KwaZulu-Natal, South Africa. Journal of Infectious Diseases. 2018; Advance online publication. doi: 10.1093/infdis/jiy394.

Spatial patterns of extensively drug-resistant tuberculosis (XDR TB) transmission in KwaZulu-Natal, South Africa.

Kristin N. Nelson,[1] N. Sarita Shah,[1,2] Barun Mathema,[3] Nazir Ismail,[4,5] James C.M. Brust,[6] Tyler S. Brown,[7] Sara C. Auld,[1,8] Shaheed Valley Omar,[4] Natashia Morris,[9] Angie Campbell,[1] Salim Allana,[1] Pravi Moodley,[10,11] Koleka Mlisana,[10,11] Neel R. Gandhi[1,8]

1. Emory University Rollins School of Public Health, Atlanta, GA, USA
2. Centers for Disease Control and Prevention, Atlanta, GA, USA
3. Columbia University Mailman School of Public Health, New York, NY, USA
4. National Institute for Communicable Diseases, Johannesburg, South Africa
5. University of Pretoria, Pretoria, South Africa
6. Albert Einstein College of Medicine and Montefiore Medical Center, Bronx, NY, USA
7. Massachusetts General Hospital, Infectious Diseases Division, Boston, MA, USA

8. Emory University School of Medicine, Atlanta, GA, USA
9. Environment and Health Research Unit, South African Medical Research Council, Johannesburg, South Africa
10. National Health Laboratory Service, Durban, South Africa
11. School of Laboratory Medicine and Medical Sciences, University of KwaZulu-Natal, Durban, South Africa

Abstract:

Background: Transmission is driving the global drug-resistant tuberculosis epidemic; nearly three-quarters of drug-resistant tuberculosis cases are attributable to transmission. Geographic patterns of disease incidence, combined with information on probable transmission links, can define the spatial scale of transmission and generate hypotheses about factors driving transmission patterns.

Methods: We combined whole-genome sequencing data with home GPS coordinates from 344 participants with extensively drug-resistant (XDR) tuberculosis in KwaZulu-Natal, South Africa diagnosed from 2011-2014. We aimed to determine if genomically linked (≤5 single nucleotide polymorphisms [SNP] differences) cases lived close to one another, which would suggest a role for local community settings in transmission.

Results: 182 study participants were genomically linked, comprising 1084 case-pairs. The median distance between case-pairs' homes was 108 km (IQR: 64-162 km). Between-district, as compared to within-district, links accounted for the majority (912/1084, 84%) of genomic links. Half (526, 49%) of genomic links involved a case from Durban, the urban center of KwaZulu-Natal.

Conclusions: The high proportions of between-district links with Durban provide insight into possible drivers of province-wide XDR TB transmission, including urban-rural migration. Further research should focus on characterizing the contribution of these drivers to overall XDR TB transmission in KwaZulu-Natal to inform design of targeted strategies to curb the drug-resistant tuberculosis epidemic.

Key words: tuberculosis, drug-resistance; extensively drug-resistant tuberculosis; molecular epidemiology; whole genome sequencing; transmission; geospatial analysis

Introduction

Drug-resistant tuberculosis is a global crisis, causing an estimated 1.2 million cases each year.[5] Extensively drug-resistant (XDR) tuberculosis has now been reported from 123 countries and is associated with mortality rates from 50-90%.[10, 15, 217] Although drug-resistant tuberculosis strains are initially created by selection of drug-resistant mutants during treatment (acquired resistance), recent studies show that the majority of drug-resistant tuberculosis cases now arise due to transmission of already drug-resistant strains.[12, 104] This shift makes clear the urgent need for interventions to prevent transmission.

Molecular epidemiology studies have consistently shown that close contacts account for only a minority of secondary tuberculosis cases in settings with high tuberculosis incidence, suggesting that a substantial proportion of transmission may occur as a result of 'casual' contact in the community. [69, 125-127, 218] Although modeling and social mixing studies support this hypothesis and point to public transportation, schools and workplaces as likely transmission sites in high tuberculosis incidence settings, this has not been demonstrated directly.[71, 73, 129] Understanding the role of contacts proximate to or distant from the home can generate hypotheses about the modes of contact driving transmission. The advent of bacterial whole-genome sequencing (WGS) offers new opportunities to identify tuberculosis cases that are likely to be linked through transmission, by discriminating between TB isolates at the level of single nucleotide polymorphisms (SNPs). Isolates from different patients that differ by small numbers of SNPs are considered likely to represent a transmission event. Recent studies have employed WGS to identify probable transmission events, map chains of transmission in tuberculosis outbreaks, and describe the burden of tuberculosis disease due to recent infection as compared to reactivation.[127, 137, 138, 141, 200, 219, 220] However, WGS has been underutilized to describe broader, population-level patterns of transmission in tuberculosis-endemic settings.

The spatial scale of disease transmission can provide insight into the settings and, by extension, the modes of contact that contribute to transmission. Tuberculosis transmission requires air exchange—and therefore close proximity—between an infectious and susceptible person. The nature and location of these interactions define the relevant geographic scale for person-to-person interactions resulting in transmission.[221, 222] For example, short distances between transmission-linked cases may indicate that local contacts in, or close to, the household are most important in transmission. Alternatively, transmission links found across longer distances may indicate that long-distance contacts, and perhaps migration, may play an important role in disseminating disease. Previous geospatial analyses in tuberculosis have focused on the spatial distribution of *cases*, rather than the spatial scale of transmission *links*. Combining geospatial analysis with WGS data has the potential to provide more comprehensive information about the dynamic process of disease transmission.

We combined *Mycobacterium tuberculosis* (*Mtb*) whole-genome sequencing and geographic data to, first, evaluate the spatial scale of XDR TB transmission in KwaZulu-Natal, South Africa, and second, quantify the proportion of transmission occurring within and between municipal districts in KwaZulu-Natal. Understanding the spatial scale and patterns of transmission can identify specific geographic areas and demographic groups that contribute to ongoing transmission and towards which interventions can be targeted.

Methods

Setting

South Africa has among the highest rates of tuberculosis globally, with 59% of tuberculosis patients co-infected with HIV. [5, 223] KwaZulu-Natal province, which comprises 11 districts and has a population of 10.3 million persons, has the highest tuberculosis and XDR TB burden (3 per 100,000) in South Africa.[215, 224] [225] The most populous district, eThekwini, is home to the city of Durban, a common destination for employment and educational opportunities. The population in KwaZulu-Natal is highly mobile— a recent study found that over a third of the population had changed residence in the past two years.[226]

Study design and procedures

The Transmission of HIV-Associated XDR TB (TRAX) study is a cross-sectional study that enrolled culture-confirmed XDR TB patients diagnosed from 2011 to 2014 in KwaZulu-Natal. Detailed methods of the TRAX study have been previously published.[12] Briefly, we identified XDR TB cases through the single referral laboratory that conducts drug-susceptibility testing (DST) for all public healthcare facilities in KwaZulu-Natal. All participants provided written informed consent; for deceased or severely ill participants, consent was obtained from next-of-kin.

We interviewed participants and performed medical record review to collect demographic information and medical history. Participants reported the locations of residences, schools, employment, hospital admissions and other congregate locations frequented in the five years preceding XDR TB diagnosis. A global position system (GPS) coordinate location was collected at the location of each participant's home residence.

Whole genome sequencing

The diagnostic XDR *Mtb* isolate was obtained for all participants and re-cultured on Löwenstein-Jensen slants. We conducted population sweeps, extracted genomic DNA, and prepared sequencing libraries using Nextera DNA kits (Illumina, San Diego, CA). Raw paired-end sequencing reads were generated on the Illumina (MiSeq) platform and aligned to the H37Rv reference genome (NC_000962.3) using the Burrows-Wheeler Aligner. All isolates had reads covering >99% of the reference genome, and the lowest mean coverage depth for any isolate was 15X. SNPs were detected using standard pairwise resequencing techniques (Samtools v0.1.19) against the reference and filtered for quality, read consensus (>75% reads for the alternate allele) and proximity to indels (>50 base-pairs from any indel). SNPs in or within 50 base pairs of hypervariable PPE/PE gene families, repeat regions, and mobile elements were excluded.[227] Alignment files can be found at NCBI Bioproject PRJNA476470.

Analysis

We defined a genomic link as a pair of XDR TB cases ('case-pair') with 5 or fewer SNP differences between their *Mtb* sequences.[67, 140, 200] We mapped and calculated median geographic distance between the home residences of genomically linked cases using the *sp* and *geosphere* packages in R 3.4.1.[228] [229]

We stratified distances between genomically linked cases by sex, given historically distinct migratory behavior among male and females in sub-Saharan Africa. We also stratified by HIV coinfection, since the influence of HIV on the susceptibility, progression, and transmissibility of tuberculosis remains uncertain.[67, 126, 230-232] Lastly, we stratified by strain type, by comparing pairs of the most common *Mtb* strain type in KwaZulu-Natal, LAM4, with other strain types. We conducted our analysis at varying SNP thresholds (≤3 SNPs, ≤1 SNP) to assess the robustness of results to this choice.

To describe patterns of transmission by district of residence, we classified each case according to the district of their home residence and calculated the proportion of between- and within-district genomic links for all districts. We also calculated the proportion of pairs in each district with links to the urban district of eThekwini.

Sensitivity analysis of differential enrollment in TRAX by district

To assess whether our results were sensitive to differential enrollment of XDR cases by district, we compared our results to those we might have observed had we enrolled all cases. We used the complete register of diagnosed XDR TB cases from the referral laboratory to calculate the fraction of diagnosed cases from each district that participated in TRAX (enrollment fraction). For within-district links, we adjusted the number of genomic links by a factor of the inverse enrollment fraction. For between-district links, we adjusted the number of links using the mean of the inverse enrollment fractions for both districts. We compared the proportions of within- and between-district links calculated using these enrollment fractions to the proportions we observed.

As cases from rural areas may have reduced access to high-quality healthcare services, we hypothesized they may be underdiagnosed, and thus included in TRAX, compared to cases from urban eThekwini district.[233, 234] To examine the effect of this potential source of bias, we varied our assumptions about the extent of this over-enrollment (assuming the enrollment fraction was anywhere from 20-40% higher in eThekwini than in other districts) and repeated our analysis of between and within-district links.

Ethical Considerations

The study was approved by the Institutional Review Boards of Emory University, Albert Einstein College of Medicine, and the University of KwaZulu-Natal, and by CDC's National Center for HIV, Hepatitis, STDs and Tuberculosis.

Results

Between 2011 and 2014, we screened 521 (51%) of 1027 culture-confirmed XDR TB patients diagnosed in KwaZulu-Natal and enrolled 404 (78% of screened) (Figure 1). TRAX participants were similar to all diagnosed XDR TB cases in terms of age (p=0.52), sex (p=0.76), and district of diagnosing facility (p=0.70). Among the 404 participants, 234 (58%) were female, with a median age of 34 years (interquartile range [IQR]: 28-43). Three hundred eleven (77%) participants were HIV-positive, of whom 236 (76%) were on antiretroviral therapy and 155 (50%) were virologically suppressed at enrollment (viral load < 400 copies/mL) (Table 4-1). Half (n = 204, 50%) of participants reported living in urban sub-districts, and 133 (33%) participants lived in eThekwini district. Mobility of TRAX participants was high, with 89 (22%) participants reporting living at a different residence than their current residence in the previous five years; 41 (46%) of those residences were in a district other than their current residence. Inter-district movement was also common—of those participants that reported spending >2 hours per week at congregate locations (n=254), 93 (37%) named a congregate location in a different district than their current residence.

Mtb isolates from 344 (85%) participants passed all sequencing quality filters and were available for analysis, creating a total of 58,996 unique case-pairs. Cases with WGS were similar to all enrolled cases (Table 4-1). Among these case-pairs, 1084 (1.8%) differed by 5 or fewer SNPs, indicating a genomic link; these case-pairs involved 182 unique participants (Figure 1). Among these 182 cases, the median number of genomic links per case was 6 (IQR: 2-17), with 63 (35%) participants having greater than 10 genomic links (Supplemental Figure 1). These 182 participants reported residences across all eleven districts in KwaZulu-Natal province, and were demographically similar to non-linked cases (Table 4-1 and 4-2, Figure 1).

Geographic distance between genomically linked participants

Among the 1084 genomically linked case-pairs, the homes of 3 (0.3%) case-pairs were within 1 km of one another, 12 (1%) were within 5 km of one another, and 29 (3%) were within 10 km. The majority of case-pairs' homes (871, 80%) were ≥ 50 km apart, and the homes of over half (589, 54%) of case-pairs were more than 100km apart. The median distance between the home residences of genomically linked cases was 108 km (IQR: 64-162 km). This distance was similar when we increased the stringency of the threshold for genomic links: among pairs with fewer than 3 SNPs, the median distance was 117 km (IQR 67-162); among pairs with fewer than 1 SNP difference, the median distance was 127 km (IQR 59-152) (Figure 3). The median distance between case-pairs homes' was >95 km for all strata of sex, HIV status, and strain type. (Supplemental Table 4-1).

Since some cases had multiple genomic links, we wanted to determine whether cases with distant links also had links close to home. We selected the geographically closest link for each case. Among the 182 cases involved in genomically linked case-pairs, 20 (11%) cases lived within 5 km of their closest link, 40 (22%) lived within 10 km; 68 (37%) lived more than 50 km from their closest geographic link, and 22 (12%) of cases lived over 100km from their closest link. The median distance to the closest geographic link was 32 km (Supplemental Figure 2).

Within- and between-district links

Overall, 16% of genomic links were among case-pairs residing within the same district (172/1084), while 84% of genomically linked case-pairs lived in different districts of KwaZulu-Natal province (912/1084)

(Figure 4-4, Table 4-3). Three districts had no within-district genomic links (Amajuba, iLembe, and Sisonke) and eThekwini had the highest proportion of within-district links (17%). Proportions of within- and between-district links were similar when the SNP threshold was reduced to fewer than 3 SNPs and fewer than 1 SNP (Supplemental Table 4-2).

Approximately half (n=526, 49%) of all case-pairs were linked to the urban district of eThekwini. In every district except for two (Sisonke and Amajuba), the plurality of genomic links included a case that lived in eThekwini (Figure 4-4, Table 4-3, Supplemental Table 4-3). eThekwini district had the highest proportion (20%) of links with Umzinyathi.

At the individual case level, nearly a third of genomically linked cases (53, 29%) lived in the metropolitan district of eThekwini. Of note, 37 (70%) of these 53 cases were genomically linked to at least one other case within eThekwini, and nearly all (n=51, 96%) were genomically linked to at least one case outside of eThekwini. Among the 129 cases who lived outside of eThekwini, approximately half (n=59, 46%) had at least one genomic link within their home district. Nearly all (n=127, 96%) had at least one genomic link outside their home district, and 76 (61%) of those cases had at least one genomic link with a case in eThekwini.

Adjustment for differential enrollment by district

Enrollment fractions, based on the total number of diagnosed cases in each district, ranged from 0.22 in Sisonke and Amajuba to 0.50 in Umkhanyakude. Adjusting for enrollment, the proportion of within- and between-district links were 15% and 85%, respectively, which is nearly identical to the proportions in the unadjusted analysis. District-specific proportions of within- and between-district links were also similar to the unadjusted proportions (Supplemental Table 4-4). When we varied the proportion of cases enrolled in eThekwini relative to other districts (assuming enrollment was up to 40% higher in eThekwini than in other districts), eThekwini still accounted for the plurality of links in all but two districts.

Discussion

We aimed to define the spatial scale and identify geographic patterns of XDR TB transmission in KwaZulu-Natal, South Africa. We found that genomically linked pairs of XDR TB cases generally lived far apart, and that the majority (84%) of genomic links were between cases who lived in different districts. Nearly half of all genomically linked case-pairs involved a case in eThekwini district. Taken together, this evidence suggests that movement across districts, as well as into and out of eThekwini, may play a central role in the dissemination of XDR TB across the province.

The median geographic distance between genomically liked cases was 108 km, which is remarkably high considering that tuberculosis cases with genetically similar strains have been found to be geographically clustered in other settings.[235, 236] We found similarly high geographic distances at more stringent thresholds of 3 and 1 SNP. Although there is no universal SNP threshold for defining a direct transmission link, there is general agreement that the threshold should be tailored to local tuberculosis epidemiology.[151, 237] Further, we also examined median distance by strain type, given that the genomic epidemiology of XDR TB in KwaZulu-Natal is dominated by a single, highly clonal strain (LAM4).[238] The median distance between genomically linked cases was similarly high among pairs of cases with the LAM4 strain and among non-LAM4 pairs. Although the LAM4 strain accounted for the majority of genomic links in our study, the phenomenon of the predominance of an individual clone is common in other settings with a high prevalence of drug-resistant tuberculosis.[148, 239].

The high proportions of between-district links and links with eThekwini suggest that cross-district movement, and perhaps eThekwini, plays a central role in patterns of XDR TB transmission in KwaZulu-Natal. While previous studies have shown concentrations of tuberculosis cases in urban areas, suggesting that these settings are conducive to transmission, they have not examined the role of urban settings in driving transmission patterns and incidence in broader geographic areas.[71, 240] Although our convenience sample of XDR TB cases diagnosed during the study period (n=404, 39%) does not provide a complete set of transmission links, we performed several analyses to assess whether our results are robust to potential selection bias. First, the demographic characteristics of TRAX cases were similar to all diagnosed cases in terms of age, sex, and the district of diagnosing facility. Second, our bias analysis showed that the proportions of between-district links and links with eThekwini remained high under scenarios of differential enrollment by district. Lastly, given that most cases of TB progress to active disease within two years of infection, it is likely that we captured the majority of relevant transmission links among TRAX cases, and that these links reflect larger transmission patterns in KwaZulu-Natal.[241]

Collectively, these findings provide insight into possible drivers of XDR TB transmission in KwaZulu-Natal. Human movement and migration can transport pathogens across long distances, resulting in transmission that occurs far from an individual’s home. Cyclical migration between rural and urban areas for employment is common in South Africa and in other rapidly developing countries, and effectively creates ‘bridge’ populations between urban and rural areas. This type of migration, which has previously been linked to HIV transmission, could also be driving tuberculosis transmission.[230] As such, it could explain both the large distances between the homes of genomically linked cases and that cases were more likely to be linked to eThekwini district than to another case in their home district.

In addition to migration for employment, individuals may move between districts for other reasons. A previous analysis of TRAX participants showed that 36% of cases who were diagnosed with XDR TB in eThekwini lived in a different district, indicating that travel from rural to urban areas for healthcare is common.[242] Importantly, travel to seek tuberculosis diagnosis and treatment is likely to coincide with an individual’s infectious period, potentially providing abundant opportunities for transmission. Inter-district travel, be it for employment, healthcare, or other reasons, expands the geographic range of settings that are relevant for transmission. Indeed, almost a quarter of congregate locations reported by TRAX participants were outside of their home district, further suggesting that many locations that are potential settings of exposure or transmission may be distant from home.

There are several limitations to this study. Underdiagnosis of XDR TB remains a challenge in resource-limited settings where insensitive diagnostic tools are commonly used and limited laboratory capacity curbs access to comprehensive drug susceptibility testing. As a result, transmission patterns observed among diagnosed cases provide only a limited characterization of province-wide patterns. In this study, however, we employed WGS to identify case-pairs with a high likelihood of transmission based on stringent SNP thresholds. The spatial scale we observed suggests an important role of migration, even if intermediate cases in the transmission chain were not diagnosed or enrolled in TRAX. Second, we captured participant's homes as only one location. In a setting like KwaZulu-Natal where migration is common, individuals may have multiple 'current' or recent residences, all of which may be possible locations of tuberculosis exposure and transmission. Thus, the 22% of cases that reported living in a different residence in the past five years may represent a lower bound on the proportion of cases that occupy multiple residences throughout the year. Future studies should aim to understand the role of cyclical migratory patterns and multiple residences in defining the settings relevant for tuberculosis exposure and transmission. Lastly, 'mixed' infections, or genetically distinct populations within the same host, present potential challenges for inferring transmission based on a single *Mtb* isolate.[243] Yet, we do not expect mixed infections to be differential with respect to participants' homes, suggesting that our results are robust to the potential effects of within-host bacterial heterogeneity.

Evidence that the drug-resistant tuberculosis epidemic is increasingly attributable to transmission of drug-resistant strains has highlighted the importance of understanding transmission patterns in order to prevent incident cases.[12, 104, 244, 245] Despite the challenges of measuring transmission, the use of next-generation bacterial sequencing technologies brings us a step closer to understanding the settings and modes of contact sustaining tuberculosis transmission in high-burden settings. By defining the spatial scale of transmission, we provide preliminary data about transmission patterns and lay the foundation for further studies that more explicitly examine associations between casual contact in urban settings, migratory behavior, and the ongoing spread of XDR TB. Ultimately, this knowledge can inform the development of tailored prevention strategies that target geographic areas and demographic groups that contribute disproportionately to transmission.

Acknowledgments: *We are grateful to the study team at the University of KwaZulu-Natal for their tireless efforts in data collection, record abstraction, participant recruitment, and interviews. We thank the participants and their families who consented to participate in this study.*

Footnotes

Conflicts of Interest: *The authors have no conflicts of interests to declare.*

Disclaimer: *The findings and conclusions in this manuscript are those of the authors and do not necessarily represent the official position of the funding agencies. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.*

Financial Support: *This study was primarily funded by grants from the US National Institute of Allergy and Infectious Diseases (NIAID), National Institutes of Health (NIH): R01AI089349 (PI Gandhi and R01AI087465 (PI Gandhi). It was also supported in part by NIH/NIAID grants: K23AI083088 (PI Brust), K24AI114444 (PI Gandhi), K23AI134182 (PI Auld), Emory CFAR P30AI050409 (PI Curran), Einstein CFAR P30AI051519 (PI Goldstein), by Einstein/Montefiore ICTR UL1 TR001073 (PI Shamoon).*

Meeting presentation: *Findings from this study have been presented at the Conference on Retroviruses and Opportunistic Infections (CROI) in February 2017 in Seattle, Washington (abstract number 660).*

Contact Information*: Correspondence and requests for reprints may be directed to Neel Gandhi (neel.r.gandh@emory.edu) at 1518 Clifton Road NE, CNR 3031, Atlanta, GA 30322.*

4.2 Figures and Tables

Table 4-1. Characteristics of participants in TRAX cohort, and comparison to subset with Whole Genome Sequencing (WGS) results and with genomic links – KwaZulu-Natal Province, South Africa.

Characteristic	TRAX cohort, n=404 n (%)	Cases with WGS, n=344 n (%)	p-value[1]	Genomically linked cases (≤5 SNPs), n=182 n (%)	p-value[2]
Demographic					
Female	234 (58)	202 (59)	0.44	111 (61)	0.37
Age, median (IQR)	34 (28-43)	34 (29-43)	0.19	34 (29-44)	0.97
0-15 yr	16 (4)	12 (3)	0.21	9 (5)	0.47
16-34 yr	207 (51)	171 (50)		88 (48)	
35-54 yr	150 (37)	134 (39)		71 (39)	
≥55 yr	31 (8)	27 (8)		14 (8)	
Monthly household income					
<R500	139 (34)	120 (35)	0.36	64 (35)	0.27
R500-R2,500	186 (46)	153 (44)		83 (46)	
>R2,500	79 (20)	71 (21)		35 (19)	
Clinical					
Current or former smoker	39 (10)	35 (10)	0.47	18 (10)	0.98
Diabetes	23 (6)	22 (6)	0.15	10 (5)	0.47
HIV positive	311 (77)	266 (77)	0.70	145 (80)	0.27
Receiving antiretroviral therapy	236 (76)	204 (77)	0.49	108/145	0.32
CD4 cell count (median, IQR)	340 (117–431)	240 (111-425)	0.26	233 (104-316)	0.54
Virologic suppression (<400 copies/mL)	155 (50)	134 (39)	0.56	74 (41)	0.49
Cough					
Patients with cough	333 (82)	284 (83)	0.87	147 (81)	0.35

Median duration of cough	8 (4-12)	8 (4-12)	0.22	8 (4-12)	0.39
Sputum smear positive for acid-fast bacilli	270 (67)	235 (68)	0.31	118 (65)	0.16
Previous treatment for any tuberculosis	291 (72)	247 (72)	0.81	127 (70)	0.38
Previous treatment for multidrug-resistant tuberculosis	124 (31)	105 (31)	0.86	45 (25)	0.01

[1] p-values compare cases with WGS (n=344) to all TRAX participants (n=404)

[2] p-values compare linked cases (n=182) to all cases with WGS (n=344)

Figure 4-1. Selection of study participants and identification of genomic links using whole genome sequencing (WGS).

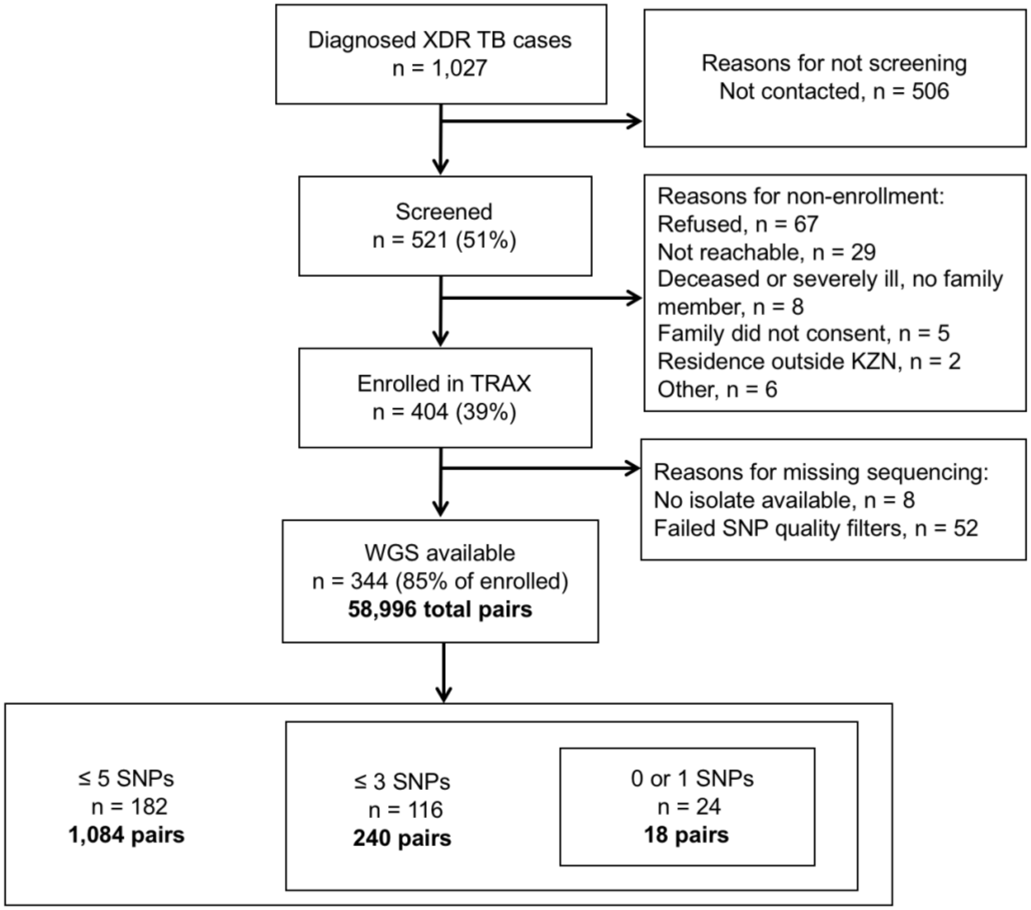
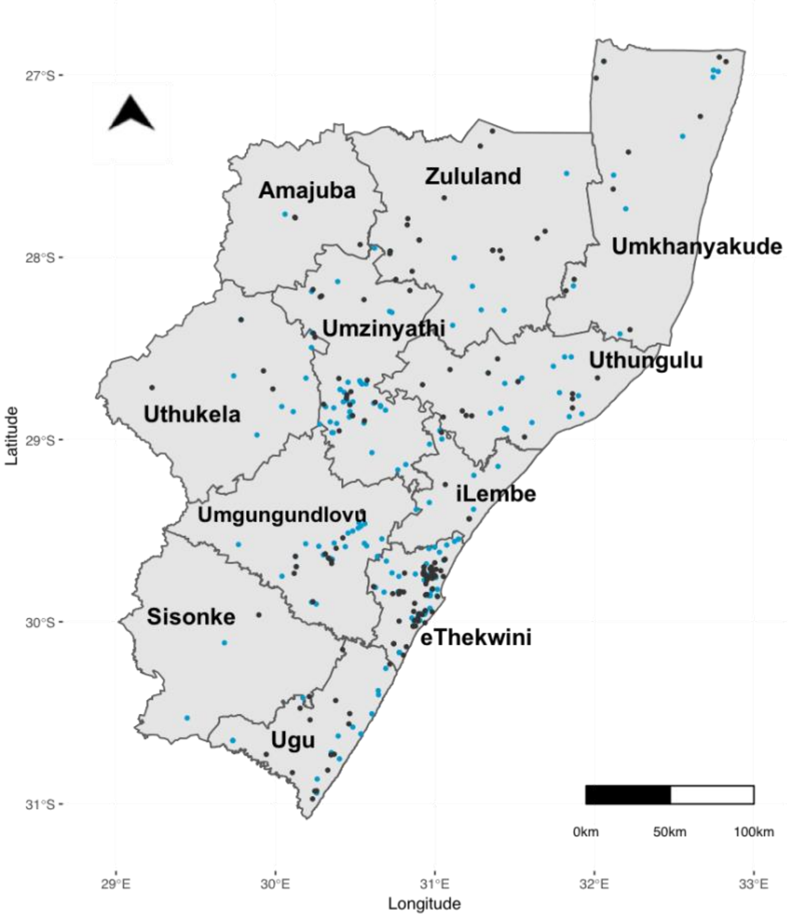

Diagnosed XDR TB cases
n = 1,027
Reasons for not screening
Not contacted, n = 506
Screened
n = 521 (51%)
Reasons for non-enrollment:
Refused, n = 67
Not reachable, n = 29
Deceased or severely ill, no family member, n = 8
Family did not consent, n = 5
Residence outside KZN, n = 2
Other, n = 6
Enrolled in TRAX
n = 404 (39%)
Reasons for missing sequencing:
No isolate available, n = 8
Failed SNP quality filters, n = 52
WGS available
n = 344 (85% of enrolled)
58,996 total pairs
≤ 5 SNPs
n = 182
1,084 pairs
≤ 3 SNPs
n = 116
240 pairs
0 or 1 SNPs
n = 24
18 pairs

Figure 4-2. Geographic distribution of XDR TB cases with genomic links in KwaZulu-Natal province, South Africa**.**



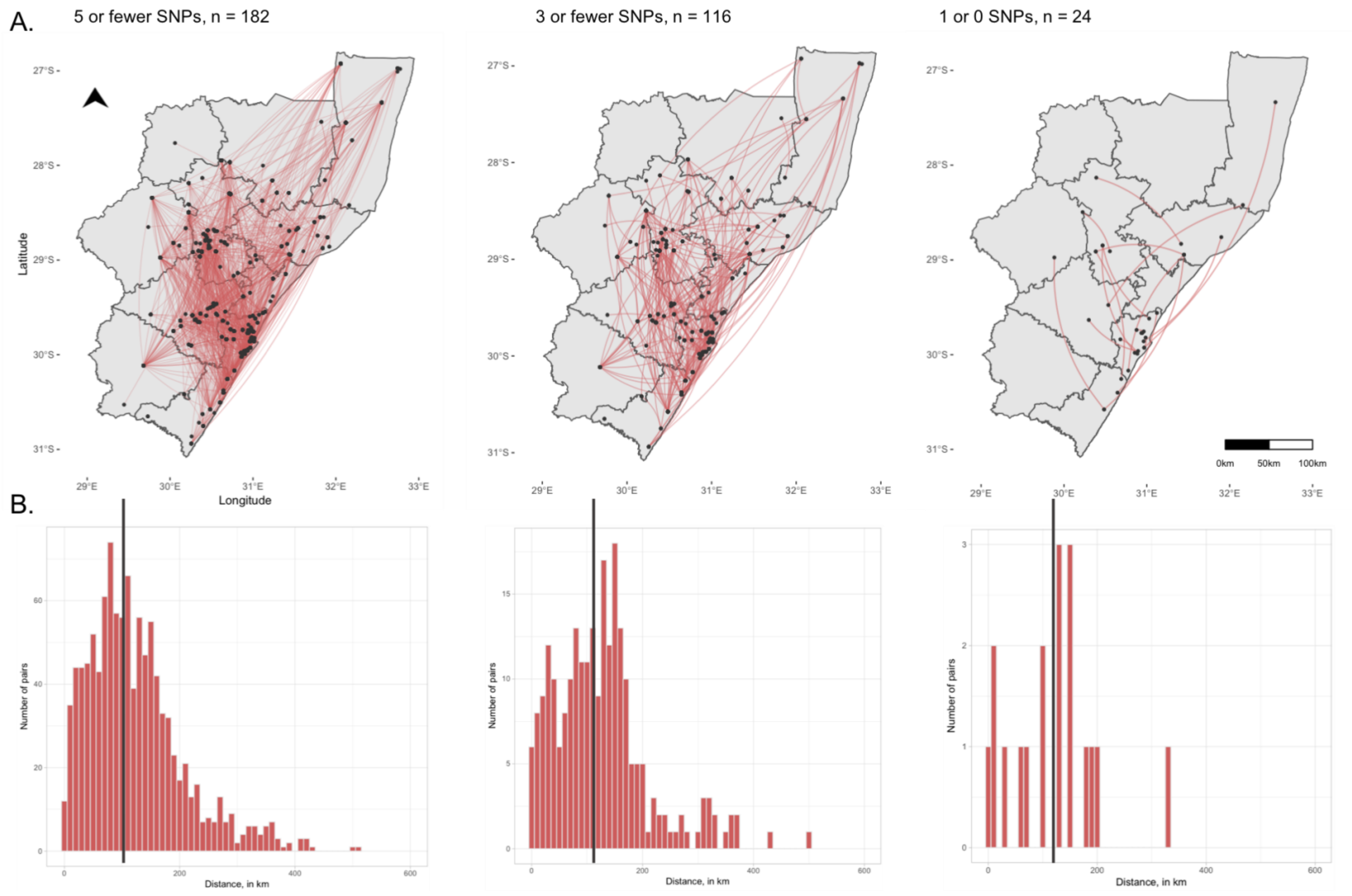
Blue dots indicate georeferenced locations of reported home residences of TRAX cases who are genomically linked; black dots indicate those not genomically linked. The eleven districts of KwaZulu-Natal are labeled. The most populous district in KwaZulu-Natal is eThekwini, which includes the city of Durban. Note: As of 2015, Sisonke district is known as Harry Gwala district and as of 2016, Uthungulu district is known as King Cetshwayo district.

Table 4-2. Geographic distribution of XDR TB cases by district.

District	n (% of total)	Population (thousands)	Genomically linked (% of total)
Amajuba	4 (1.2)	500 (4.9)	1 (0.5)
eThekwini	115 (33)	3,400 (33)	53 (29)
iLembe	11 (3.2)	607 (5.9)	7 (4)
Sisonke	4 (1.2)	461 (4.5)	3 (2)
Ugu	32 (9.3)	722 (7.0)	14 (8)
UMgungundlovu	37 (10.8)	1,018 (10)	26 (14)
Umkhanyakude	19 (5.5)	626 (6.1)	9 (5)
Umzinyathi	53 (15.4)	510 (5.0)	37 (20)
Uthukela	15 (4.4)	669 (6.5)	9 (5)
Uthungulu	30 (8.7)	908 (8.8)	16 (9)
Zululand	24 (7.0)	840 (8.2)	7 (4)
Total	344	10,261	182

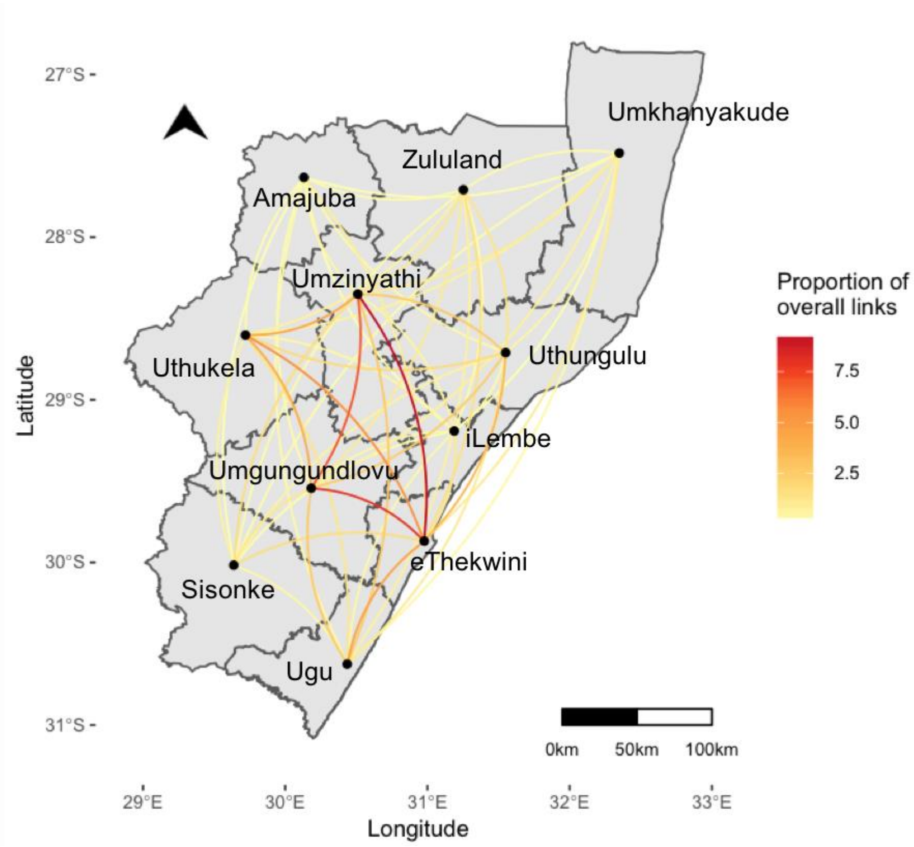
Population by district and percent of cases in each district with at least one genomic link. Population statistics sourced from the Statistics South Africa 2011 Census (http://www.statssa.gov.za/)

Figure 4-3. Map and distribution of geographic distances between home residences of genomically linked case-pairs (≤5 SNPs) in KwaZulu-Natal.



A. Black dots indicate home residences of XDR TB cases; red lines represent genomic links between cases. B. Black lines on histograms indicate the median distance between homes of genomically linked case-pairs at each SNP threshold. Note differences in y axis range across plots

Figure 4-4. Genomic links (≤5 SNPs) within and between districts in KwaZulu-Natal.



The proportion of genomic links occurring between each district out of the total number of links involving that district is represented by the color of the line. Amajuba district, which had only one genomic link, was excluded from this analysis.

Table 4-3. Proportions of within- and between-district genomic links (≤5 SNPs) in KwaZulu-Natal.

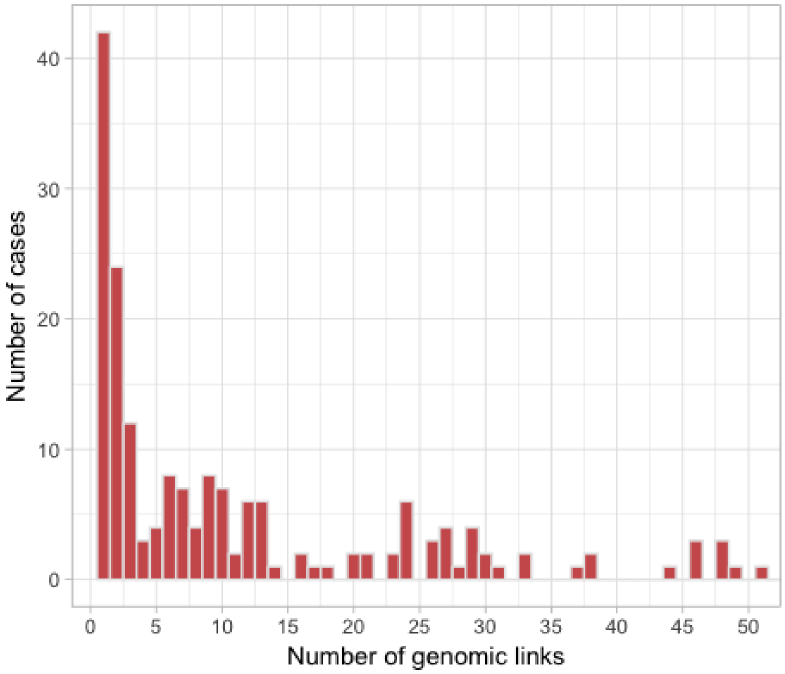
District	Total links	Within-district links (%)	Between-district links (%)	Links with eThekwini (%)
Amajuba	1	0 (0)	1 (100)	0 (0)
eThekwini	526	91 (17)	435 (83)	--
iLembe	32	0 (0)	32 (100)	10 (31)
Sisonke	61	0 (0)	61 (100)	12 (20)
Ugu	236	12 (5)	224 (95)	75 (32)
UMgungundlovu	313	23 (7)	290 (93)	100 (32)
Umkhanyakude	97	1 (1)	96 (99)	25 (26)
Umzinyathi	334	32 (10)	302 (90)	104 (31)
Uthukela	160	7 (4)	153 (96)	45 (28)
Uthungulu	171	5 (3)	166 (97)	50 (29)
Zululand	65	1(2)	64 (99)	14 (22)

4.3 Supplemental Results

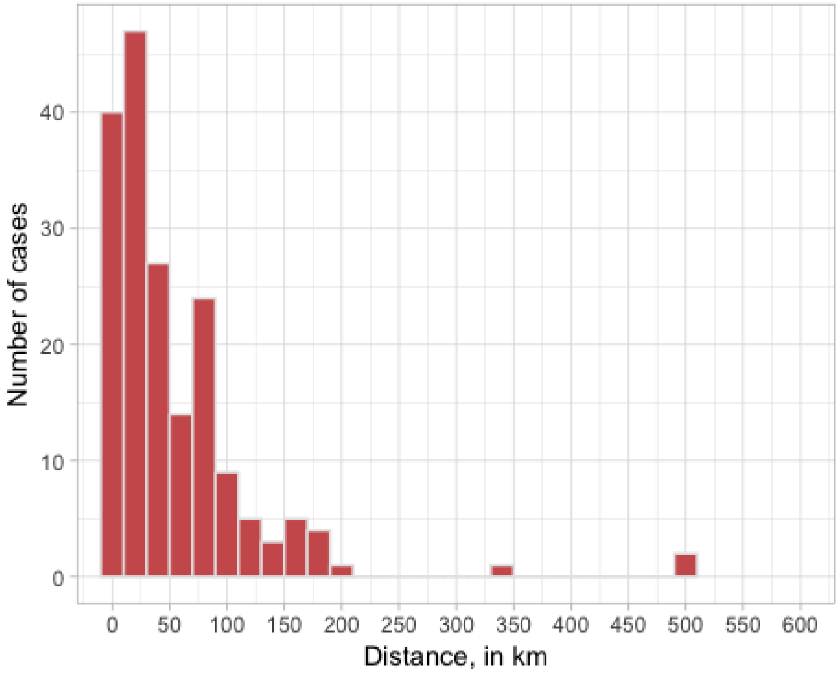
Supplemental Table 4-1. Genomic links (≤5 SNPs) by sex, HIV status, and strain type.

Pair	Total links (% of total)	Median distance, in km (IQR)
Sex		
Female / Female	382 (35)	96 (49 – 150)
Female / Male	534 (49)	107 (69 – 165)
Male / Male	168 (15)	131 (84 – 150)
HIV status		
HIV+ / HIV+	654 (60)	104 (58 – 155)
HIV+ / HIV-	377 (35)	117 (70 – 166)
HIV- / HIV-	53 (5)	136 (80 – 197)
Strain type		
LAM4 / LAM4	1075 (99)	127 (68 – 147)
Non-LAM4 / Non-LAM4	9 (1)	108 (64 – 162)

Supplemental Figure 4-1. Number of genomic links (≤5 SNPs) per case.



Supplemental Figure 4-2. Shortest geographic link among genomic links (≤5 SNPs) for each case.



Supplemental Table 4-2. Within and between-district links at ≤3 and ≤1 SNPs.

	≤ 5 SNPs				≤ 3 SNPs				≤ 1 SNP			
District	Total links	Within-district (%)	Between-district (%)	With eThekwini (%)	Total links	Within-district (%)	Between-district (%)	With eThekwini (%)	Total links	Within-district (%)	Between-district (%)	With eThekwini (%)
Amajuba	1	0 (0)	1 (100)	0 (0)	0	-	-	-	0	-	-	-
eThekwini	526	91 (17)	435 (83)	--	115	24 (21)	91 (79)	--	11	4 (36)	7 (64)	--
iLembe	32	0 (0)	32 (100)	10 (31)	10	(0)	10 (100)	3 (30)	0	-	-	-
Sisonke	61	0 (0)	61 (100)	12 (20)	15	(0)	15 (100)	4 (27)	0	-	-	-
Ugu	236	12 (5)	224 (95)	75 (32)	74	50 (7)	24 (93)	25 (34)	4	0 (0)	4 (100)	1 (25)
UMgungundlovu	313	23 (7)	290 (93)	100 (32)	47	1 (2)	46 (98)	9 (19)	2	0 (0)	2 (100)	1 (50)
Umkhanyakude	97	1 (1)	96 (99)	25 (26)	20	1 (5)	19 (95)	6 (30)	2	0 (0)	2 (100)	1 (50)
Umzinyathi	334	32 (10)	302 (90)	104 (31)	64	50 (8)	14 (92)	21 (33)	6	0 (0)	6 (100)	2 (33)
Uthukela	160	7 (4)	153 (96)	45 (28)	37	3 (8)	34 (92)	9 (24)	2	0 (0)	2 (100)	0 (0)
Uthungulu	171	5 (3)	166 (97)	50 (29)	39	1 (3)	38 (97)	12 (31)	5	0 (0)	5 (100)	2 (40)
Zululand	65	1(2)	64 (99)	14 (22)	18	1 (6)	17 (94)	2 (11)	0	-	-	-

Supplemental Table 4-3. Within and between-district genomic links (≤5 SNPs).

	Amajuba	eThekwini	iLembe	Sisonke	Ugu	UMgungundlovu	Umkhanyakude	Umzinyathi	Uthukela	Uthungulu	Zululand	Total number of links involving each district	Within-district links (%)	Between-district links (%)	Links with eThekwini (%)
Amajuba	0	0	0	0	0	0	0	0	0	1	0	1	0 (0)	1 (100)	0 (0)
eThekwini		91	10	12	75	100	25	104	45	50	14	526	91 (17)	435 (83)	--
iLembe			0	2	3	3	2	6	2	4	0	32	0 (0)	32 (100)	10 (31)
Sisonke				0	6	9	4	15	6	5	2	61	0 (0)	61 (100)	12 (20)
Ugu					12	45	13	39	16	17	10	236	12 (5)	224 (95)	75 (32)
UMgungundlovu						23	16	54	27	27	9	313	23 (7)	290 (93)	100 (32)
Umkhanyakude							1	10	11	11	4	97	1 (1)	96 (99)	25 (26)
Umzinyathi								32	32	29	13	334	32 (10)	302 (90)	104 (31)
Uthukela									7	12	2	160	7 (4)	153 (96)	45 (28)
Uthungulu										5	10	171	5 (3)	166 (97)	50 (29)
Zululand											1	65	1(2)	64 (99)	14 (22)

Supplemental Table 4-4. Enrollment fraction bias analysis of between- and within-district genomic links (≤5 SNPs).

District	Sampling fraction	Inverse sampling fraction	Total number of links involving each district	Within-district links (%)	Between-district links (%)	Links with eThekwini (%)
Amajuba	0.22	4.50	24.2	2.3 (9.3)	22.0 (90.7)	1.8 (7.3)
eThekwini	0.39	2.54	1377.9	230.7 (16.7)	1147.2 (83.3)	230.7 (16.7)
iLembe	0.28	3.57	106.0	1.8 (1.7)	104.2 (98.3)	30.5 (28.8)
Sisonke	0.22	4.50	223.6	2.3 (1.0)	221.3 (99.0)	42.2 (18.9)
Ugu	0.42	2.37	598.0	2.3 (4.8)	595.8 (95.3)	183.3 (30.7)
UMgungundlovu	0.37	2.68	839.4	28.4 (7.3)	811.0 (92.7)	260.6 (31.0)
Umkhanyakude	0.50	2.00	232.1	61.6 (0.9)	170.5 (99.1)	56.7 (24.4)
Umzinyathi	0.36	2.80	915.6	2.0 (9.8)	913.6 (90.2)	277.3 (30.3)
Uthukela	0.30	3.38	554.6	89.5 (4.3)	465.1 (95.7)	133.0 (24.0)
Uthungulu	0.40	2.48	499.6	23.6 (2.5)	476.0 (97.5)	125.4 (25.1)
Zululand	0.38	2.61	206.3	12.4 (1.3)	193.9 (98.7)	36.0 (17.4)

Total inflated links = 3017.2

5. Aims 1 and 2

5.1 Manuscript 1

Clinical markers and social mixing patterns associated with transmission of extensively drug-resistant (XDR) tuberculosis.

Abstract: Tuberculosis (TB) is the leading infectious cause of death globally, and drug-resistant TB strains pose a serious threat to controlling the global TB epidemic. Interventions to reduce transmission of drug-resistant TB in high-burden settings require a clear understanding of the clinical features, types of social contact, and settings driving the spread of TB. We constructed a network of genomic links using *Mtb* whole genome sequences and found that XDR TB cases with cavitary disease, reporting 2-3 months of cough, or had more extensive contact with urban settings were more highly connected in the network, while those with smear-positive disease were less likely to be linked. These associations persisted in networks using different SNP thresholds to define genomic links, using undirected networks in which we did not make assumptions about direction of transmission, and in analyses using conventional regression methods. Understanding the factors driving local TB epidemics can aid in tailoring TB control efforts; however, further analyses should explicitly consider the role of missing cases in transmission networks and the extent to which this may affect conclusions about transmission patterns.

Introduction

Tuberculosis (TB) is the leading infectious cause of death globally, and drug-resistant TB strains pose a serious threat to controlling the global TB epidemic. [5] Recent studies have shown that the majority of drug-resistant TB cases in settings with a high burden of TB are due to transmission of drug-resistant strains, rather than acquisition of resistance through inadequate or incomplete TB treatment. [12, 104] Interventions to reduce the burden of drug-resistant TB in high-burden settings must include efforts to reduce transmission, which will require a clear understanding of the clinical features, types of social contact, and settings driving the spread of TB.

Recent work has suggested that transmission heterogeneity, or the notion that there is variation in the number of secondary infections caused by individual cases, may play a critical role in shaping TB epidemiology in high-burden settings. [41, 130, 246, 247] On a population level, transmission heterogeneity arises as a result of inter-individual differences in three domains: the extent and duration of infectiousness, rate of contact with susceptible individuals, and the susceptibility of exposed individuals. The first two factors can be considered functions of the biologic features of disease; the third is related to behavioral patterns that define where and with whom infected persons spend time. [36] Identifying individual-level risk factors associated with a high number of secondary cases can provide insight into the clinical features that may identify individuals at high risk of transmitting disease, as well as the settings or types of contact that should be prioritized for interventions. Indeed, studies have suggested that control measures targeted towards individuals more likely to transmit disease, or implemented in specific geographic areas, may outperform broader control measures. [41, 248]

However, the factors driving TB transmission in high-burden settings remain poorly understood. Clinical characteristics associated with transmission are markers of high bacterial burden, including sputum smear status and cavitary disease on chest x-ray. However, recent studies have suggested that smear-negative cases may be responsible for a larger proportion of transmission in high-burden settings than previously thought, calling into question the ‘classical’ picture of infectious TB. [69, 126] Coughing aids in expelling *Mtb* from the respiratory tract and into the surrounding environment; as such, length of cough symptoms may also be associated with transmission. However, there are few studies that examine duration of cough as a predictor of transmission. [55] An individual's likelihood of transmitting TB is also influenced by social behaviors, or mixing patterns, that enable person-to-person contact permissive of transmission. Although studies suggest that over 80% of transmission may occur outside of the home, there is limited understanding of the settings and behaviors that drive transmission in community settings and among non-close contacts, or persons who do not know one another. Although institutional (i.e., hospitals) and congregate community settings have previously been implicated as venues of transmission, the relative importance of these settings in driving TB spread is not well understood. [69, 125-127]

The advent of bacterial whole genome sequencing (WGS) to characterize *Mtb* patient isolates provides new opportunities to study transmission, especially in areas with a high burden of TB. Relative to less sensitive genotyping methods, WGS allows for improved precision in identifying cases with highly similar *Mtb* sequences, thereby generating hypotheses about putative transmission events between cases. WGS may prove especially useful in settings with a high burden of TB, where the majority of transmission is expected to occur in community settings among patients who may not know one another and could therefore not be identified through investigation of close contacts. *Mtb* sequencing data can be used to generate a network comprised of putative transmission links between cases, and analytic methods specific to networks provide a framework for studying case characteristics associated with transmission. Network analysis has been successfully used in combination with whole genome sequences to identify clinical and social drivers of HIV transmission [249, 250]; however, such methods have been not been utilized to analyze networks of TB transmission and provide insight into potentially important patterns of TB spread. [168, 251]

Identifying individual-level factors associated with transmission can aid in developing interventions that will reduce disease spread. In this study, we used bacterial whole genome sequencing to define plausible transmission links between extensively drug-resistant TB cases in KwaZulu-Natal, South Africa. Using the resulting 'sequencing-based network', we measured associations between clinical and social mixing factors and transmission using network analysis methods.

Methods

Study design and procedures

The Transmission of HIV-Associated XDR TB (TRAX) study was a cross-sectional study that enrolled culture-confirmed XDR TB patients diagnosed from May 2011 through August 2014 in KwaZulu-Natal, South Africa. [12] The primary aim of the study was to determine the proportion of XDR TB cases that develop due to transmission, as compared to acquired drug resistance resulting from inadequate treatment. Detailed methods of the TRAX study have been previously published. [12] We identified XDR TB cases through the single referral laboratory that conducts drug-susceptibility testing (DST) for all public healthcare facilities in the province. All participants provided written informed consent. For deceased or severely ill participants, consent was obtained from next-of-kin.

We interviewed participants and collected demographic information, including age, sex, occupation, education and income level. Clinical information on the participants' medical history, including previous tuberculosis disease and HIV status, was provided by participants and supplemented with information from the medical record. Participants reported the locations of residences, schools, employment, and other congregate locations frequented in the five years prior to their XDR TB diagnosis. Participants also reported the location and duration of hospital admissions in the five years prior to diagnosis.

Whole genome sequencing

The diagnostic XDR TB isolate was obtained for all enrolled participants and re-cultured on Löwenstein-Jensen slants. We conducted population sweeps, extracted genomic DNA, and prepared sequencing libraries using Nextera DNA kits (Illumina, San Diego, CA). Raw paired-end sequencing reads were generated on the Illumina (MiSeq) platform and aligned to the H37Rv reference genome (NC_000962.3) using the Burrows-Wheeler Aligner. All isolates had reads covering >99% of the reference genome, and the lowest mean coverage depth for any isolate was 15X. SNPs were detected using standard pairwise resequencing techniques (Samtools v0.1.19) against the reference and filtered for quality, read consensus (>75% reads for the alternate allele) and proximity to indels (>50 base-pairs from any indel). SNPs in or within 50 base pairs of hypervariable PPE/PE gene families, repeat regions, and mobile elements were excluded. [227]

Defining transmission links

We defined a directed genomic link as a pair of XDR tuberculosis cases with 5 or fewer SNP differences between their *Mtb* sequences and considered the case with the earliest diagnosis date as the primary case. We constructed a network comprised of directed genomic links between cases, in which each node in the network represented a case and each edge, or link, in the network represented a genomic link. For each case in the network, we calculated the number of genomic links in which they were the primary case, reflecting the number of putative forward transmission events for which the case was responsible.

Clinical markers of infectiousness

We considered sputum smear, grade, cavitary disease, and duration of cough symptoms as markers of infectiousness. Sputum smear grade and status for each patient was collected from the diagnostic XDR TB sputum sample. Chest x-ray results at time of XDR TB diagnosis were abstracted from patient medical charts. Cough duration prior to diagnosis was reported by each patient at the time of enrollment into the study.

Social mixing measures

We constructed measures of social mixing using information using information reported by cases on their close contacts, as well as activities and locations visited prior to XDR TB diagnosis. We constructed a measure of hospital contact by using patient-reported locations and durations of hospitals stays in the five years prior to diagnosis, by summing the total number of months that each case had spent in hospital during this period. To define urban contact, we used information on current and previous residences as well as congregate locations at which patients spent more than two hours per week. The urban exposure measure included three components: whether a case reported ever living in eThekwini, whether they had been admitted to a hospital in eThekwini in the previous five years, and whether they reported spending time at a congregate location in eThekwini (theoretical range of this variable: 0-3). A second urban exposure measure was constructed representing the *number* of urban locations (residential, healthcare, or other congregate) at which a case spent time. Participants also completed a social network questionnaire, in which they reported the number of close contacts with whom they spent more than two hours per week. We used the number of contacts reported in this questionnaire as a measure of close contact.

Exponential random graph models

While generalized linear models require the assumption that the values of the dependent variable are independent across subjects, this assumption is not met by the relational data represented in a network. Specifically, node attributes and therefore network position may be correlated across nodes. To address this, we used exponential random graph models (ERGMs) to test associations using between node attributes and their connectivity, or degree, in the network. These models account for the statistical dependencies in network data and produce valid estimates of associations between case attributes and network position.

We fit ERGMs to test associations between individual case attributes (clinical markers of infectiousness and social mixing) and their position in the network. In addition to the primary predictors of interest, our main models included terms for sex, age category, and HIV status (HIV-negative, HIV-positive with an undetectable viral load, HIV-positive with an undetectable viral load).

Alternate models and sensitivity analyses

We tested associations using two alternative SNP thresholds to define genomic links: one more stringent (≤ 3 SNPs) and one less stringent (≤ 10 SNPs). We also tested a definition for genomic links that combined whole genome sequencing with conventional genotyping (restriction fragment length polymorphism typing, or RFLP) results. In addition, we tested associations in ‘undirected’ networks, in which we did not assume a direction of transmission for genomic links.

In addition to ERGMs, we used standard regression methods to ensure our results were robust to analytic method. We used negative binomial regression models and zero-inflated negative binomial models to test associations between case attributes (infectiousness and social mixing variables) and number of directed genomic links. (Zero-inflated models account for the high proportion of individuals with no links in the network.) We performed negative binomial regression and calculated robust standard errors using the glm.nb function from the MASS package in R and zero-inflated negative binomial models using the COUNTREG procedure in SAS.

We constructed parsimonious models, which excluded terms that did not change main effect estimates by more than 10% when removed from the model, and fully-specified models, which included additional terms for year of study enrollment and *Mtb* strain type, both of which may be artefactually related to the number of genomic links per case. The predominant strain of XDR TB circulating in KwaZulu-Natal, the LAM4 KZN strain, is highly clonal and therefore a lower SNP threshold may be required to define transmission links as compared to other strains. As such, cases with the LAM4 strain may be artificially linked with larger number of other cases than cases with other strains; controlling for strain type removes this effect. Second, year of enrollment in the study was associated with the likelihood of observing transmission links for that case. We observed more transmission links from cases enrolled earlier in the study because cases whom they infected could be enrolled in the later years of our study; we were unable to capture as many forward transmission from cases enrolled in later years of the study. To account for these relationships, we assessed whether models adjusted for *Mtb* strain type and year of study enrollment produced similar results to our main model.

Results

TRAX cohort

Between 2011 and 2014, we screened 521 (51%) of 1027 culture-confirmed XDR TB patients diagnosed in KwaZulu-Natal and enrolled 404 (78% of screened). Among the 404 participants, 234 (58%) were female, with a median age of 34 years (interquartile range [IQR]: 28-43). Three hundred eleven (77%) participants were HIV-positive, of whom 236 (76%) were on antiretroviral therapy and 155 (50%) were virologically suppressed at enrollment (viral load < 400 copies/mL) (Table 1). *Mtb* isolates from 344 (85%) participants passed all sequencing quality filters and were available for analysis.

Genomic links

Among 344 cases comprising the sequencing-based network (threshold of ≤ 5 SNPs), there were a total of 740 genomic links. 125 cases (36%) had at least one genomic link. Among those cases with genomic links, the number of links ranges from 1 to 28; 38 cases (30%) had between 1 and 5 genomic links, 69 cases (55%) had 6 to 9 genomic links and 18 cases (14%) had more than 10 genomic links (Figure 5-1). When the SNP threshold was increased to ≤ 10 SNPs, there were 181 (53%) cases with at least one link; at a lower threshold of ≤ 3 SNPs, there were 116 (34%) cases with at least one link (Supplemental Table 1).

Infectiousness measures

In our primary model, reporting 2 or 3 months of cough was associated with being more highly linked the network than those reporting no cough: the odds of a genomic link among cases with 2 and 3 months of cough was 2.7 times (95% CI: 2.18, 3.26) higher and 2.4 times higher (95% CI: 1.94, 2.85) than those with no cough, respectively (Table 5-2, Figure 5-2). However, this trend did not continue in the highest category of cough; those who reported more than 4 months of cough were less likely to be linked.

Irrespective of smear grade, smear-positive cases were less likely to be linked than smear-negative cases. Cases with the highest smear grade of 3+ were the least likely to be linked (OR: 0.55, 95% CI: 0.44, 0.68) (Table 5-2, Figure 5-2). When we ignored smear grade and considered only smear status (smear-negative and smear-positive), smear-positive cases were 0.43 times less likely (95% CI: 0.15, 1.24) to have a genomic link than smear-negative cases (Supplemental Table 2). Cavitary disease was associated with a higher likelihood of genomic links: the odds of a link among cases with cavitary disease was 1.5 times higher (95% CI: 1.27, 1.85) than among those with no cavitary disease.

In networks defined using a more stringent SNP threshold (≤3 SNPs) and in which we did not assume a direction of transmission, the direction and magnitude of associations with infectiousness were generally similar (Supplemental Table 3). Associations with smear status were inconsistent across networks, but higher smear grade was generally associated with a lower number of genomic links. In the full models that included terms for *Mtb* strain type and year of enrollment, results were very similar to those from the primary model (Supplemental Table 4). Conventional regression models showed results similar in direction and magnitude to exponential random graph models; however, some associations were weaker in zero-inflated negative binomials models (Supplemental Table 5).

Social mixing measures

In our primary model, cases reporting contact with 1 or more urban settings (residential, healthcare, or other congregate) were more highly connected in the network than those whom reported no contact. Compared to those reporting no contact with urban settings, cases reporting contact with 1 urban setting had 2.6 times the odds of a genomic link (95% CI: 2.19, 3.06); those reporting 2 or more urban settings had 1.7 times the odds of a link (95% CI: 1.33, 2.24) (Table 5-3, Figure 5-3). When we deconstructed the variable describing contact with urban settings to determine specifically which component was most strongly associated with being highly linked, we found that reporting a stay in an urban hospital most associated with a high number of genomic links (IRR: 2.65, 95% CI: (1.60, 4.39). However, cases with very long lengths of stay in a hospital (irrespective of location) were less likely to be linked in the network than those who spent less time. The odds of a genomic link among cases who spent 3-5 months in hospital was 0.82 times that of those who spent ≤ 2 months (95% CI: 0.69, 0.98); those who spent 5 or more months had 0.4 times the odds of a link (95% CI: 0.28, 0.48). Close contact was moderately associated with being linked in the network. Cases who reported 5-10 or more than 10 contacts were more likely to have a link than cases who reported fewer than four close contacts (Table 5-3, Figure 5-3). Notably, associations with infectiousness measures (smear status, cavitary disease, cough duration) persisted and were similar in magnitude in models including all social mixing predictors.

Age, HIV status, and all infectiousness measures were important confounders and could not be dropped from the model without altering model coefficients substantially; the only predictor that could be dropped was sex. Models that excluded sex as a predictor produced very similar results to those models including sex, with little improvement in precision. In models that also included terms for *Mtb* strain type and year enrolled, associations were similar in direction and magnitude to results from the primary model, with the exception that associations with urban contact were weaker (Supplemental Table 5-7).

In networks with alternative definitions of genomic links, directions of associations with urban contact, hospital contact, and close contacts were generally consistent, but the strength of associations were model-dependent (Supplemental Table 5-8). In conventional regression models, results were consistent with our main model (Supplemental Table 5-9).

Discussion

We constructed a network of genomic links using *Mtb* whole genome sequences and found that XDR TB cases with cavitary disease, reporting 2-3 months of cough, or had more extensive contact with urban settings were more highly connected in the network, while those with smear-positive disease were less likely to be linked. These associations persisted in networks using different SNP thresholds to define genomic links, using undirected networks in which we did not make assumptions about direction of transmission, and in analyses using conventional regression methods.

These associations are largely consistent with prior studies of TB transmission. Cough duration and cavitary disease have previously been associated with being genotypically-linked to other cases in drug-susceptible and multidrug-resistant TB. [51, 55, 57] These studies employed conventional genotyping methods to define genotypic links; our study shows that these associations are robust to more precise molecular characterization methods (WGS) with higher specificity to define transmission links between cases. Although the trend between cough and network position did not persist in the group with the longest duration of cough symptoms; this may reflect an upper bound on the effect of cough duration on transmission. Indeed, previous research has suggested that ‘saturation’ of contacts may occur over the course of a long infectious period, and long infectious periods may be common in settings with high TB burden and among drug-resistant TB patients. [130, 252] Surprisingly, we found a negative association between smear status and position in the network. There may be several reasons for this: first, we only considered a single sputum result at the time of diagnosis, which may fail to fully represent smear status over the course of TB infection. This association may also be due to a relationship between cavitary disease and smear status, whereby increased bacterial burden in the lungs leads to higher levels of bacteria in sputum. We included both variables in our models, which might attenuate a true positive association between smear status and transmission. Another potential explanation for this paradoxical finding is that patients with smear-negative disease may experience diagnostic delays, leading to longer infectious periods and thus more opportunities for transmission. [28]This is consistent with recent evidence suggesting that smear-negative cases may be important in driving transmission in settings with high TB incidence. [69, 126]

We found associations with several social settings that may be associated with transmission. Contact with urban settings prior to diagnosis was related to being more highly connected in the network. Urban areas are known to have higher incidence of TB, because they tend to provide ideal conditions for disease spread by generally increasing person-to-person contact rates, but they have also been hypothesized to drive disease incidence in wider geographic areas. Our findings support previous findings from this same cohort suggesting that rural-urban migration may be driving transmission of XDR TB in KwaZulu-Natal. [253] We also report the number of months spent in the hospital in the five years prior to diagnosis was negatively associated with connectivity in the network. This was a peculiar finding, as healthcare facilities and other institutional settings are generally considered to be ‘amplifiers’ of transmission; as such, TB outbreaks in these settings are known to drive transmission in the larger community. [120] The lack of an association between hospital stays and transmission may be obscured by the granularity at which hospital stay data was collected, which was in months. Measuring duration of stay in weeks may be more relevant for assessing transmission risk. However, there are also potential explanations for our negative findings: individuals that spent time in hospital during their infectious periods may have encountered fewer susceptible individuals than those who spent their infectious periods living, working, and socializing in their communities.

We found a weakly positive association with the number of close contacts named by participants and network connectivity. We hypothesized that reporting more close contacts would reflect a higher level of engagement in person-to-person contact (either through social activities, employment, school, or home life) and thus more opportunities for transmission. This trend did not persist among those reporting the highest numbers of close contacts, but numbers were small in these groups. As with cough duration, this may reflect a threshold above which having more contacts does not necessarily lead to additional transmission events, a notion supported by previous modeling studies. [130]

This analysis has several limitations. First, we enrolled 40% of all diagnosed XDR TB cases in KwaZulu-Natal during this time period; therefore, there are missing cases and genomic links in this network. If missing cases are intermediates in the transmission chain between sampled, genomically linked cases, this would likely lead to larger genomic differences between sampled cases. However, reducing the SNP threshold showed similar results to our primary analysis, suggesting that these findings may be robust to the effects of missing cases. Although cases enrolled in the study were demographically similar to all diagnosed cases in terms of age (p = 0.52), sex (p = 0.76), and district of diagnosing facility (p = 0.70), the extent to which enrolled cases are representative of undiagnosed cases in not clear. Second, our assumptions about the direction of genomic links using diagnosis dates may have led to incorrect classification of the directionality of links. Although we also conducted our analyses using a network in which we did not assume directionality and found largely similar results, we still cannot distinguish between individual-level factors that increase risk of infection from those that increase risk of transmission. However, it is useful to note that, theoretically, that the source of infection should only account for one link per case; the remainder of links are due to forward transmissions. Therefore, this limitation likely has little effect on our results. Lastly, the appropriate threshold genetic distance for defining transmission remains uncertain, and may depend on local epidemiologic context. [46, 151] However, we conducted our analyses using several different thresholds for genomic links and found similar results, suggesting that our main results are not sensitive to this threshold.

Identifying individual-level factors driving tuberculosis transmission can inform development of prevention strategies, if such factors can be easily identified and targeted with effective interventions. Although network analysis has potential to enhance understanding of drug-resistant TB transmission patterns, further analyses should explicitly consider the role of missing cases in analyzing partial transmission networks. Evidence-based interventions will be critical to reduce the burden of TB in countries with high levels of ongoing transmission.

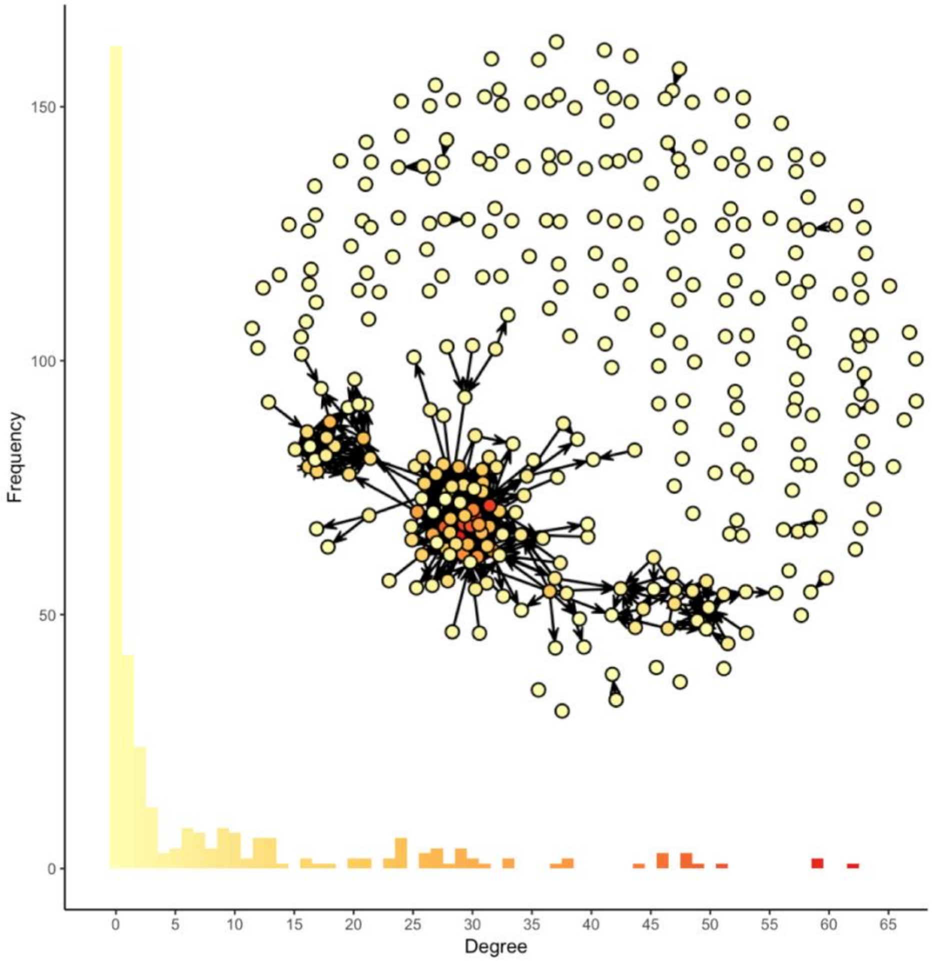
5.2 Figures and Tables

Table 5-1. Demographic, clinical and social mixing characteristics of the TRAX cohort.

Characteristic	n (%), unless otherwise noted
Demographic	
Female	202 (59)
Age, median (IQR)	34 (29-43)
0-15	12 (3)
16-34	171 (50)
35-54	134 (39)
≥ 55	27 (8)
Monthly household income	
< R500	120 (35)
R500-R2,500	153 (44)
> R2,500	71 (21)
Clinical characteristics	
Current or former smoker	35 (10)
Diabetes	22 (6)
HIV positive	266 (77)
Receiving antiretroviral therapy	204 (77)
CD4 cell count (median, IQR)	240 (111-425)
Virologic suppression (<400 copies/mL)	134 (39)
Cough	
Patients with cough	284 (83)
Median duration of cough	8 (4-12)
Sputum smear negative for acid-fast bacilli	235 (68)
Scanty-positive	
Smear-positive, grade 1	59 (17)
Smear-positive, grade 2	51 (15)
Smear-positive, grade 3 +	
Cavitary disease	60 (17)
Previous treatment for any tuberculosis	247 (72)
Previous treatment for multidrug-resistant tuberculosis	105 (31)

Social mixing characteristics	
Number of reported close contacts, median (IQR)	7 (4, 10)
Number of months in hospital, median (IQR)	3 (2, 5)
Previous stay at urban hospital	175 (51)
Current or previous urban residence	39 (11)
Reported visiting urban congregate setting	38 (11)

Figure 5-1. Variation in the number of links among TRAX cases in the sequencing-based network.



The number of links per case, or the degree distribution, of the sequencing-based network of TRAX cases. In this network, edges are defined by genomic links of ≤ 5 SNPs. Inset: The sequencing-based network. Each dot represents a TRAX cases and each arrow a genomic link. In both figures, cases with a lower network degree are in yellow and higher degree in red.

Table 5-2. Associations between infectiousness and network position. [1]

	n (%)	Odds Ratio	95% CI	p
Cough duration				
No cough reported	128 (37)	Ref	-	-
1 month	60 (17)	0.51	(0.37, 0.71)	< 0.01
2 months	51 (14)	2.66	(2.18, 3.26)	< 0.01
3 months	72 (21)	2.35	(1.94, 2.85)	< 0.01
≥ 4 months	33 (10)	1.06	(0.78, 1.45)	0.72
Smear status				
Smear -	109 (32)	Ref	-	-
Smear +, scanty +	37 (11)	0.97	(0.76, 1.24)	0.86
Smear +, grade 1	59 (17)	0.65	(0.52, 0.82)	< 0.01
Smear +, grade 2	51 (15)	0.70	(0.56, 0.88)	< 0.01
Smear +, grade 3+	88 (26)	0.55	(0.44, 0.68)	< 0.01
Cavitary disease				
No cavitary disease	284 (83)	Ref	-	-
Cavitary disease	60 (17)	1.53	(1.27, 1.85)	< 0.01

[1] Model also includes terms for HIV status, sex, and age

Figure 5-2. Associations between infectiousness and network position.

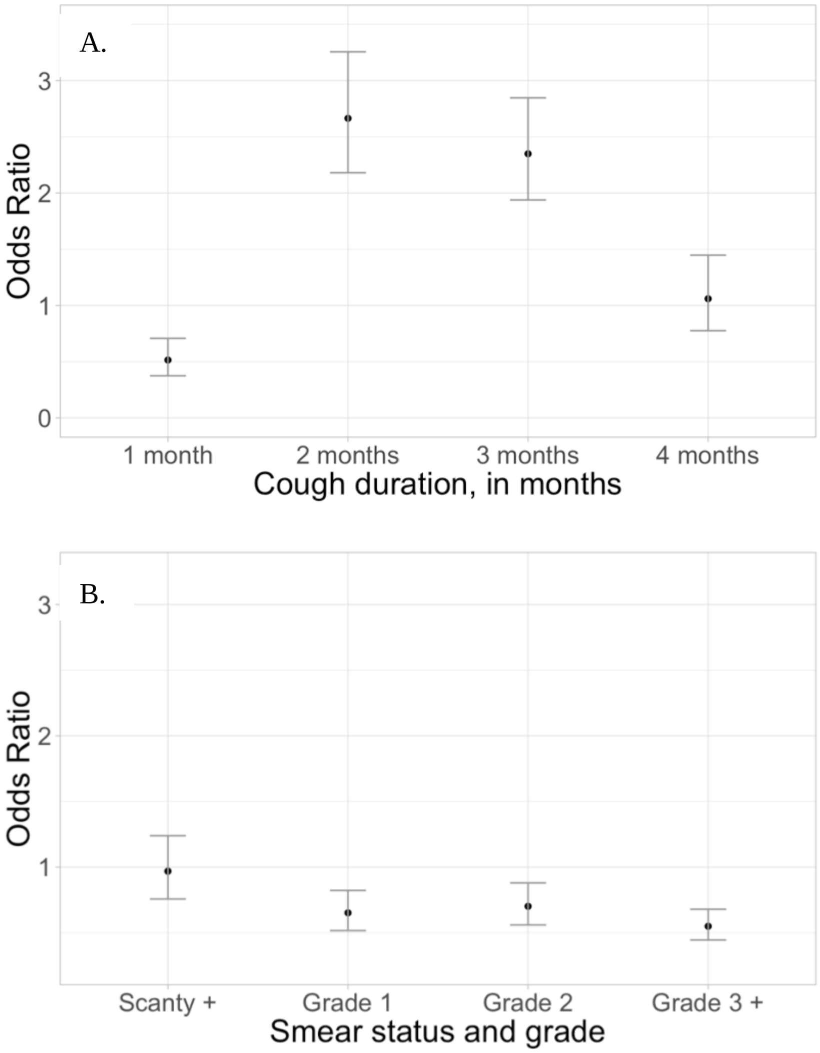


Figure 2. Dots represent odds ratio estimate and bars represent 95% confidence interval. A. Reference category for cough duration is no cough reported. B. Reference category for smear status and grade is smear-negative. (Cavitary disease not shown)

Table 5-3. Associations between social mixing measures and network position. [1]

	n (%)	Odds Ratio	95% CI	p
Contact with urban areas				
0 urban settings	149 (43)	Ref	-	-
1 urban settings	147 (43)	2.58	(2.18, 3.06)	< 0.01
≥ 2 urban settings	48 (20)	1.56	(1.33, 2.24)	< 0.01
Duration in hospital				
0 - 2 months	113 (33)	Ref	-	-
3 - 5 months	81 (24)	0.82	(0.69, 0.98)	0.03
> 5 months	59 (17)	0.37	(0.28, 0.48)	< 0.01
Named close contacts				
0 - 4 contacts	108 (31)	Ref	-	-
5 - 10 contacts	144 (42)	1.19	(0.99, 1.43)	0.05
> 10 contacts	85 (25)	1.42	(1.16, 1.73)	< 0.01

[1] Model also includes terms for HIV status, sex, and age, smear status and grade, cavitary disease, and cough duration

Figure 5-3. Associations between social mixing measures and network position.

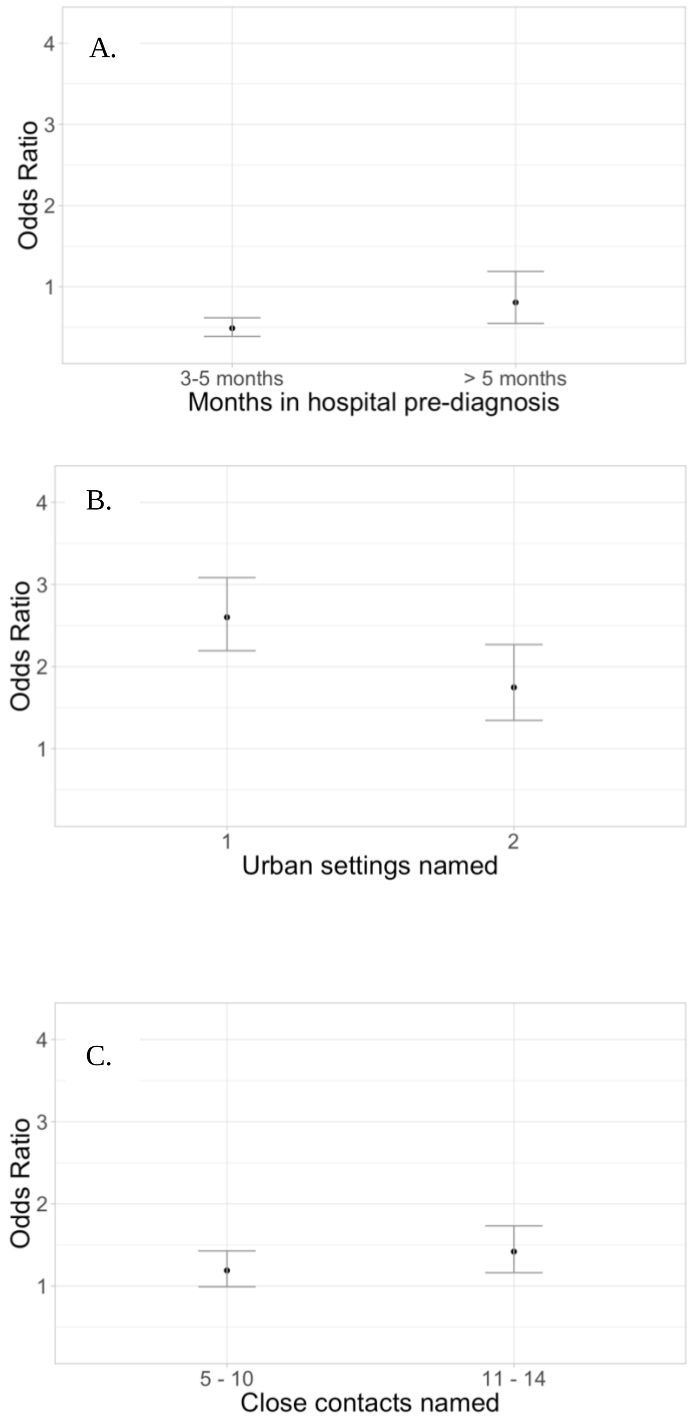


Figure 3. Dots represent odds ratio estimate and bars represent 95% confidence interval. A. Reference category for urban settings is zero. B. Reference category for months in hospital pre-diagnosis is 0-2 months. C. Reference category for close contacts is 0-5.

5.3 Supplemental Results

Supplemental Table 5-1. Characteristics of networks with alternate SNP thresholds and genomic link definitions.

Genomic link definition	Number of edges (total links) in network	Cases with at least one link (%)	Mean links per case	Median links per case	Maximum number of links
≤ 3 pairwise SNPs, undirected	240	116 (34)	1.4	0.0	22
Undirected RFLP/WGS (≤ 3 pairwise SNPs) combined	353	100 (29)	2.1	7.50	20
≤ 10 pairwise SNPs, directed	4704	181 (53)	13.7	2.0	95

Supplemental Table 5-2. Associations between infectiousness and network position in models with collapsed variable for smear status.

	n (%) Total n = 344	Odds Ratio	95% CI		p
Cough duration					
No cough reported	128 (37)	Ref	-	-	-
1 month	60 (17)	0.51	0.37	0.70	<0.01
2 months	51 (14)	2.59	2.12	3.16	<0.01
3 months	72 (21)	2.28	1.88	2.76	<0.01
≥ 4 months	33 (10)	1.06	0.78	1.45	0.72
Smear status					
Smear -	109 (32)	Ref	-	-	-
Smear +	235 (68)	0.65	0.55	0.76	<0.01
Cavitary disease					
No cavitary disease	284 (83)	Ref	-	-	-
Cavitary disease	60 (17)	1.50	1.25	1.81	<0.01
HIV status					
HIV-negative	78 (23)	Ref	-	-	-
HIV-positive, undetectable VL	133 (39)	1.17	0.95	1.42	0.14
HIV-positive, detectable VL	133 (39)	1.02	0.83	1.25	0.87
Sex					
Male	142 (41)	Ref	-	-	-
Female	202 (59)	0.92	0.80	1.07	0.28
Age category					
≤ 15	12 (3)	Ref	-	-	-

16-34	171 (50)	0.76	0.53	1.09	0.14
35-54	134 (39)	0.84	0.58	1.21	0.36
≥ 55	27 (8)	1.25	0.84	1.86	0.27

Supplemental Table 5-3. Associations between infectiousness and network position in networks with alternate genomic link definitions.

	Undirected ≤ 5 SNPs				Directed ≤ 10 SNPs				Undirected RFLP/WGS (≤ 3 SNPs) combined			
	Odds Ratio	95% CI		p	Odds Ratio	95% CI		p	Odds Ratio	95% CI		p
Cough duration												
No cough reported	Ref	-	-	-	Ref	-	-	-	Ref	-	-	-
1 month	1.27	1.12	1.45	<0.01	0.51	0.45	0.57	<0.01	0.70	0.54	0.89	<0.01
2 months	1.67	1.47	1.90	<0.01	1.92	1.77	2.08	<0.01	1.53	1.24	1.88	<0.01
3 months	1.69	1.51	1.90	<0.01	1.84	1.71	1.99	<0.01	1.37	1.13	1.67	<0.01
4 months	0.88	0.68	1.12	0.29	0.63	0.52	0.77	<0.01	0.49	0.30	0.81	0.01
≥ 5 months	0.80	0.62	1.04	0.10	1.11	0.95	1.29	0.20	0.82	0.54	1.24	0.35
Smear status												
Smear - negative	Ref	-	-	-	Ref	-	-	-	Ref	-	-	-
Smear +, scanty +	1.26	1.10	1.45	<0.01	0.78	0.70	0.87	<0.01	1.51	1.16	1.95	<0.01
Smear +, grade 1	0.66	0.57	0.76	<0.01	0.63	0.57	0.69	<0.01	1.17	0.92	1.48	0.20
Smear +, grade 2	0.89	0.78	1.02	0.09	0.65	0.60	0.72	<0.01	1.02	0.79	1.32	0.87
Smear +, grade 3+	0.80	0.71	0.90	<0.01	0.61	0.56	0.66	<0.01	1.35	1.10	1.67	<0.01
Cavitary disease												
No cavitary disease	Ref	-	-	-	Ref	-	-	-	Ref	-	-	-
Cavitary disease	1.31	1.17	1.46	<0.01	1.50	1.39	1.61	<0.01	1.46	1.21	1.75	<0.01
HIV status												
HIV-negative	Ref	-	-	-	Ref	-	-	-	Ref	-	-	-
HIV-positive, undetectable VL	1.20	1.06	1.35	<0.01	1.41	1.30	1.54	<0.01	1.29	1.04	1.61	0.02
HIV-positive, detectable VL	1.07	0.95	1.21	0.26	1.19	1.10	1.30	<0.01	1.45	1.17	1.80	0.00

Sex												
Male	Ref	-	-	-	-	-	-	-	Ref	-	-	-
Female	1.01	0.93	1.11	0.75	1.12	1.05	1.19	<0.01	1.00	0.86	1.16	0.98
Age category												
≤ 15	Ref	-	-	-	Ref	-	-	-	Ref	-	-	-
16-34	0.77	0.62	0.97	0.02	0.66	0.57	0.76	<0.01	1.15	0.72	1.82	0.57
35-54	0.82	0.65	1.03	0.08	0.75	0.65	0.86	<0.01	1.02	0.64	1.64	0.93
≥ 55	1.08	0.84	1.39	0.53	1.14	0.97	1.33	0.11	1.63	0.97	2.72	0.06

Supplemental Table 5-4. Associations between infectiousness and network position in full and parsimonious models.

		Parsimonious model (excluding sex)				**Full model (including TB strain and year)**			
	n (%) Total n = 344	Odds Ratio	95% CI		p	Odds Ratio	95% CI		p
Cough duration									
No cough reported	128 (37)	Ref	-	-	-	Ref	-	-	-
1 month	60 (17)	0.51	0.37	0.71	<0.01	0.74	0.53	1.04	0.08
2 months	51 (14)	2.67	2.18	3.26	<0.01	2.15	1.75	2.64	<0.01
3 months	72 (21)	2.35	1.94	2.85	<0.01	1.96	1.61	2.39	<0.01
4 months	16 (5)	0.97	0.63	1.48	0.89	1.25	0.81	1.93	0.32
≥ 5 months	17 (5)	1.17	0.77	1.78	0.46	0.82	0.54	1.24	0.35
Smear status									
Smear -	109 (32)	Ref	-	-	-	Ref	-	-	-
Smear +, scanty +	37 (11)	0.96	0.75	1.22	0.72	1.13	0.87	1.47	0.37
Smear +, grade 1	59 (17)	0.65	0.51	0.81	<0.01	0.80	0.62	1.01	0.06
Smear +, grade 2	51 (15)	0.70	0.56	0.88	<0.01	0.73	0.57	0.93	0.01
Smear +, grade 3+	88 (26)	0.55	0.44	0.68	<0.01	0.59	0.47	0.74	<0.01
Cavitary disease									
No cavitary disease	284 (83)	Ref	-	-	-	Ref	-	-	-
Cavitary disease	60 (17)	1.52	1.26	1.83	<0.01	1.72	1.41	2.09	<0.01
HIV status									
HIV-negative	78 (23)	Ref	-	-	-	Ref	-	-	-
HIV-positive, undetectable VL	133 (39)	1.19	0.97	1.45	0.10	1.32	1.06	1.63	0.01
HIV-positive, detectable VL	133 (39)	1.00	0.81	1.23	0.98	1.06	0.86	1.32	0.57

Sex									
Male	142 (41)	-	-	-	-	Ref	-	-	-
Female	202 (59)	-	-	-	-	0.83	0.71	0.97	0.02
Age category									
≤ 15	12 (3)	Ref				Ref			
16-34	171 (50)	0.81	0.56	1.17	0.27	0.67	0.47	0.97	0.04
35-54	134 (39)	0.91	0.62	1.31	0.60	0.63	0.44	0.92	0.02
≥ 55	27 (8)	1.37	0.92	2.04	0.12	1.03	0.68	1.55	0.89
Strain									
HP	259 (75)	-	-	-	-	Ref	-	-	-
Non-HP	85 (25)	-	-	-	-	37.76	17.89	79.68	<0.01
Year									
2011	58 (17)	-	-	-	-	Ref	-	-	-
2012	107 (31)	-	-	-	-	0.43	0.36	0.51	<0.01
2013	82 (24)	-	-	-	-	0.36	0.29	0.44	<0.01
2014	97 (28)	-	-	-	-	0.14	0.10	0.20	<0.01

Supplemental Table 5-5. Associations between infectiousness and network position using conventional regression methods.

	n (%), Total n = 344	Negative binomial					Zero-inflated Negative binomial		
		Estimate	IRR	95% CI		p	Mean difference[1]	95% CI	
Cough duration									
No cough reported	128 (37)	Ref	-	-	-	-	Ref	-	-
1 month	60 (17)	-0.80	0.45	0.43	0.87	0.02	-0.99	-2.32	0.35
2 months	51 (14)	1.08	2.94	2.92	5.98	<0.01	2.9	-0.59	6.38
3 months	72 (21)	1.05	2.85	2.85	5.34	<0.01	2.78	-0.29	5.85
4 months	16 (5)	-0.66	0.52	0.25	2.24	0.38	-0.77	-3.42	1.88
≥ 5 months	17 (5)	0.10	1.11	0.21	3.51	0.86	0.47	-2.66	3.6
Smear status									
Smear - negative	109 (32)	Ref	-	-	-	-	Ref	-	-
Smear +, scanty +	37 (11)	0.55	1.73	1.41	3.33	0.10	0.82	-1.25	2.89
Smear +, grade 1	59 (17)	-0.58	0.56	0.47	1.10	0.09	-0.8	-2.02	0.43
Smear +, grade 2	51 (15)	-0.22	0.81	0.23	1.94	0.63	1.43	-1.55	4.42
Smear +, grade 3+	88 (26)	-0.38	0.68	0.46	1.22	0.20	-0.7	-1.74	0.35
Cavitary disease									
No cavitary disease	284 (83)	Ref	-	-	-	-	Ref	-	-
Cavitary disease	60 (17)	0.31	1.36	0.72	2.53	0.33	0.78	-0.9	2.46
HIV status									
HIV-negative	78 (23)	Ref	-	-	-	-	Ref	-	-
HIV-positive, undetectable VL	133 (39)	0.35	1.42	0.81	2.75	0.29	0.16	-1.15	1.48
HIV-positive, detectable VL	133 (39)	0.06	1.06	0.20	1.98	0.85	-0.05	-1.3	1.19
Sex									
Male	142 (41)	Ref	-	-	-	-	Ref	-	-
Female	202 (59)	-0.08	0.93	0.21	1.48	0.75	-0.09	-0.97	0.79
Age category									
≤15	12 (3)	Ref	-	-	-	-	Ref	-	-
16-34	171 (50)	-0.48	0.62	0.33	1.62	0.33	-0.7	-2.73	1.33
35-54	134 (39)	-0.23	0.80	0.23	1.99	0.63	0.56	-1.85	2.96
≥ 55	27 (8)	0.52	1.69	0.81	5.37	0.37	0.54	-2.55	3.64

[1] Zero-inflated negative binomial results are expressed as the mean difference in log risk relative to an individual with characteristics corresponding to all reference categories.

Supplemental Table 5-6. Deconstructing urban contact variable. [1]

		Negative binomial				
	n (%), Total n = 344	Estimate	IRR	95% CI		p
Named urban congregate setting						
No	306 (89)					
Yes	38 (11)	-0.27	0.77	0.38	1.54	0.36
Named urban hospital setting						
No	169 (49)					
Yes	175 (51)	0.98	2.65	1.60	4.39	0.26
Named (current or former) urban residence						
No	305 (89)					
Yes	39 (11)	0.10	1.11	0.56	2.19	0.77

[1] Model also adjusted for all other social mixing, clinical infectiousness, and demographic variables.

Supplemental Table 5-7. Associations between social mixing measures and network position in alternate models.

		Parsimonious model (excluding sex)				**Full model (including TB strain and year)**			
	n (%) Total n = 344	Odds Ratio	95% CI		p	Odds Ratio	95% CI		p
Contact with urban areas									
0 urban settings	149 (43)	Ref	-	-	-	Ref	-	-	-
1 urban settings	147 (43)	2.66	2.23	3.16	<0.01	2.13	1.78	2.56	<0.0001
2 urban settings	39 (11)	1.65	1.21	2.25	<0.01	1.07	0.78	1.46	0.68
3 urban settings	9 (3)	3.85	2.53	5.87	<0.01	2.88	1.78	4.64	<0.0001
Duration in hospital									
0 - 2 months	113 (33)	Ref	-	-	-	Ref	-	-	-
3 - 5 months	81 (24)	0.90	0.75	1.08	0.25	0.89	0.74	1.08	0.23
≥ 6 months	59 (17)	0.36	0.27	0.47	<0.01	0.43	0.33	0.58	<0.0001
Named close contacts									
0 - 4 contacts	108 (31)	Ref	-	-	-	Ref	-	-	-
5 - 9 contacts	144 (42)	1.19	0.99	1.44	0.07	0.97	0.80	1.18	0.74
10 - 14 contacts	72 (21)	1.49	1.21	1.83	<0.01	1.39	1.11	1.74	0.00
≥ 15 contacts	13 (4)	0.98	0.66	1.44	0.91	1.33	0.89	1.99	0.17
Cough duration									
No cough reported	128 (37)	Ref	-	-	-	Ref	-	-	-
1 month	60 (17)	0.43	0.31	0.59	<0.01	0.68	0.48	0.95	0.02
2 months	51 (14)	2.39	1.95	2.93	<0.01	1.82	1.47	2.25	<0.01
3 months	72 (21)	2.32	1.91	2.83	<0.01	1.88	1.53	2.31	<0.01
4 months	16 (5)	1.02	0.66	1.57	0.94	1.30	0.82	2.06	0.27

≥ 5 months	17 (5)	1.43	0.94	2.18	0.10	1.00	0.65	1.52	0.99
Smear status									
Smear - negative	109 (32)	Ref	-	-	-	Ref	-	-	-
Smear +, scanty +	37 (11)	0.89	0.69	1.14	0.34	1.04	0.80	1.36	0.75
Smear +, grade 1	59 (17)	0.63	0.50	0.81	<0.01	0.80	0.62	1.03	0.09
Smear +, grade 2	51 (15)	0.63	0.50	0.79	<0.01	0.70	0.55	0.89	<0.01
Smear +, grade 3+	88 (26)	0.59	0.48	0.74	<0.01	0.64	0.51	0.80	<0.01
Cavitary disease									
No cavitary disease	284 (83)	Ref	-	-	-	Ref	-	-	-
Cavitary disease	60 (17)	1.43	1.26	1.63	<0.01	1.48	1.29	1.69	<0.01
HIV status									
HIV-negative	78 (23)	Ref	-	-	-	Ref	-	-	-
HIV-positive, undetectable VL	133 (39)	1.22	1.00	1.49	0.05	1.35	1.09	1.68	0.01
HIV-positive, detectable VL	133 (39)	0.95	0.78	1.16	0.63	1.07	0.86	1.32	0.56
Sex									
Male	142 (41)	-	-	-	-	Ref	-	-	-
Female	202 (59)	-	-	-	-	0.79	0.68	0.93	0.01
Age category									
≤15	12 (3)	Ref	-	-	-	Ref	-	-	-
16-34	171 (50)	0.86	0.59	1.25	0.43	0.78	0.54	1.14	0.20
35-54	134 (39)	0.88	0.60	1.28	0.49	0.66	0.45	0.96	0.03
≥ 55	27 (8)	1.17	0.78	1.74	0.45	1.14	0.76	1.72	0.52
Strain									

HP	259 (75)	-	-	-	-	Ref	-	-	-
Non-HP	85 (25)	-	-	-	-	35.50	16.79	75.05	<0.01
Year									
2011	58 (17)	-	-	-	-	Ref	-	-	-
2012	107 (31)	-	-	-	-	0.52	0.43	0.63	<0.01
2013	82 (24)	-	-	-	-	0.47	0.37	0.60	<0.01
2014	97 (28)	-	-	-	-	0.16	0.11	0.22	<0.01

Supplemental Table 5-8. Associations between social mixing measures and node outdegree in networks with alternate SNP thresholds and genomic link definitions.

	Undirected ≤ 5 SNPs				**Directed ≤ 10 SNPs**				**Undirected RFLP/WGS (≤ 3 SNPs) combined**			
	Odds Ratio	95% CI		p	Odds Ratio	95% CI		p	Odds Ratio	95% CI		p
Contact with urban areas												
0 urban settings	Ref	-	-	-	Ref	-	-	-	Ref	-	-	-
1 urban settings	1.49	1.35	1.65	<0.01	1.83	1.71	1.96	<0.01	1.35	1.14	1.61	0.00
2 urban settings	1.77	1.52	2.06	<0.01	1.47	1.31	1.64	<0.01	1.11	0.84	1.48	0.46
3 urban settings	1.89	1.47	2.43	<0.01	1.16	0.93	1.44	0.19	1.24	0.77	1.99	0.38
Duration in hospital												
0 - 2 months	Ref	-	-	-	Ref	-	-	-	Ref	-	-	-
3 - 5 months	0.97	0.87	1.08	0.57	0.83	0.77	0.90	<0.01	0.95	0.79	1.15	0.61
≥ 6 months	0.41	0.35	0.48	<0.01	0.57	0.52	0.63	<0.01	0.71	0.56	0.90	0.00
Named close contacts												
0 - 4 contacts	Ref	-	-	-	Ref	-	-	-	Ref	-	-	-
5 - 9 contacts	1.10	0.98	1.22	0.09	1.26	1.17	1.36	<0.01	1.18	0.98	1.43	0.08
10 - 14 contacts	1.42	1.25	1.60	<0.01	1.32	1.21	1.44	<0.01	1.59	1.29	1.96	<0.01
≥ 15 contacts	0.70	0.55	0.88	<0.01	0.80	0.68	0.94	0.01	0.74	0.48	1.15	0.18
Cough duration												
No cough reported	Ref	-	-	-	Ref	-	-	-	Ref	-	-	-
1 month	1.10	0.96	1.25	0.17	0.86	0.81	0.92	<0.01	0.68	0.53	0.88	<0.01
2 months	1.57	1.38	1.78	<0.01	1.47	1.38	1.56	<0.01	1.46	1.18	1.80	<0.01
3 months	1.68	1.49	1.89	<0.01	1.18	1.11	1.25	<0.01	1.38	1.13	1.68	<0.01
4 months	0.81	0.63	1.04	0.09	0.84	0.75	0.95	<0.01	0.49	0.29	0.81	<0.01

≥ 5 months	0.86	0.66	1.12	0.27	0.82	0.73	0.92	<0.01	0.84	0.55	1.28	0.43
Smear status												
Smear - negative	Ref	-	-	-	Ref	-	-	-	Ref	-	-	-
Smear +, scanty +	1.23	1.07	1.42	<0.01	0.97	0.90	1.04	0.35	1.56	1.20	2.04	<0.01
Smear +, grade 1	0.67	0.58	0.78	<0.01	0.60	0.56	0.65	<0.01	1.18	0.92	1.50	0.19
Smear +, grade 2	0.86	0.75	0.98	0.03	0.78	0.73	0.83	<0.01	1.00	0.77	1.30	0.98
Smear +, grade 3+	0.83	0.73	0.93	<0.01	0.81	0.77	0.86	<0.01	1.37	1.11	1.70	<0.01
Cavitary disease												
No cavitary disease	Ref	-	-	-	Ref	-	-	-	Ref	-	-	-
Cavitary disease	1.41	1.26	1.57	<0.01	1.27	1.20	1.34	<0.01	1.51	1.25	1.82	<0.01
HIV status												
HIV-negative	Ref	-	-	-	Ref	-	-	-	Ref	-	-	-
HIV-positive, undetectable VL	1.25	1.11	1.41	<0.01	1.18	1.11	1.25	<0.01	1.30	1.05	1.62	0.02
HIV-positive, detectable VL	1.07	0.95	1.21	0.26	0.97	0.92	1.03	0.35	1.46	1.17	1.81	<0.01
Sex												
Male	Ref	-	-	-	Ref	-	-	-	Ref	-	-	-
Female	0.91	0.83	1.00	0.05	1.13	1.08	1.18	<0.01	0.92	0.79	1.08	0.31
Age category												
≤ 15	Ref	-	-	-	Ref	-	-	-	Ref	-	-	-
16-34	0.78	0.62	0.98	0.03	0.82	0.73	0.91	<0.01	1.17	0.73	1.86	0.52
35-54	0.81	0.64	1.02	0.07	0.98	0.87	1.09	0.66	1.02	0.64	1.65	0.92
≥ 55	1.01	0.78	1.30	0.97	1.07	0.95	1.21	0.28	1.57	0.94	2.64	0.09

Supplemental Table 5-9. Associations between social mixing measures and node outdegree using conventional regression methods.

	Negative Binomial					**Zero-inflated Negative Binomial**		
	Estimate	IRR	95% CI		p	Mean difference[1]	95% CI	
Contact with urban areas								
0 urban settings	Ref	-	-	-	-	-	-	-
1 urban settings	0.94	2.55	1.54	4.24	<0.01	2.78	-1.39	6.96
2 urban settings	0.87	2.40	1.20	4.77	0.01	2.43	-0.8	5.65
3 urban settings	0.32	1.38	0.25	7.58	0.71	0.79	-3.37	4.96
Duration in hospital								
0 - 2 months	Ref	-	-	-	-	-	-	-
3 - 5 months	-0.01	0.99	0.55	1.78	0.96	-0.5	-1.03	0.03
≥ 6 months	-1.01	0.36	0.19	0.69	<0.01	-0.68	-1.31	-0.05
Named close contacts								
0 - 4 contacts	Ref	-	-	-	-	-	-	-
5 - 9 contacts	0.27	1.31	0.78	2.22	0.31	-0.06	-0.61	0.49
10 - 14 contacts	0.44	1.56	0.84	2.88	0.16	0.21	-0.52	0.95
≥ 15 contacts	-0.48	0.62	0.16	2.35	0.48	-0.29	-1.11	0.54
Cough duration								
No cough reported	Ref	-	-	-	-	-	-	-
1 month	-0.81	0.45	0.22	0.93	0.03	-0.47	-1.07	0.14
2 months	1.14	3.11	1.59	6.11	<0.01	1.86	-0.08	3.81
3 months	1.14	3.12	1.70	5.71	<0.01	2.15	0.36	3.95
4 months	-0.63	0.53	0.15	1.90	0.33	0.37	-2.18	2.92
≥ 5 months	0.13	1.14	0.41	3.18	0.80	0.74	-1.16	2.64
Smear status								
Smear - negative	Ref	-	-	-	-	-	-	-
Smear +, scanty +	0.69	2.00	1.00	3.99	0.05	0.12	-0.68	0.92
Smear +, grade 1	-0.49	0.61	0.30	1.22	0.16	0.01	-1.45	1.48
Smear +, grade 2	-0.51	0.60	0.29	1.24	0.17	0.85	-0.54	2.24
Smear +, grade 3+	-0.46	0.63	0.38	1.06	0.08	-0.48	-1	0.04
No cavitary disease	Ref	-	-	-	-	-	-	-
Cavitary disease	0.28	1.33	0.75	2.34	0.33	-0.12	-0.66	0.42

HIV status								
HIV-negative	Ref	-	-	-	-	-	-	-
HIV-positive, undetectable VL	0.49	1.63	0.92	2.91	0.09	0.02	-0.63	0.67
HIV-positive, detectable VL	0.29	1.33	0.76	2.32	0.31	0.06	-0.61	0.73
Sex								
Male	Ref	-	-	-	-	-	-	-
Female	-0.19	0.83	0.52	1.31	0.42	0.02	-0.51	0.56
Age Category								
≤15	Ref	-	-	-	-	-	-	-
16-34	-0.67	0.51	0.20	1.30	0.16	-0.55	-1.52	0.42
35-54	-0.46	0.63	0.24	1.66	0.35	0.6	-2.88	4.09
≥ 55	0.14	1.15	0.37	3.56	0.81	0.07	-1.28	1.42

[1] Zero-inflated negative binomial results are expressed as the mean difference in log risk relative to an individual with characteristics corresponding to all reference category.

6. Aim 3

6.1 Manuscript 2

<u>Modeling missing cases in transmission networks of extensively drug-resistant (XDR) tuberculosis.</u>

Abstract

Tuberculosis (TB) is the leading infectious cause of death worldwide, and in 2016, there were over half a million new cases of drug-resistant TB. Analysis of transmission networks may be useful to understand drivers of transmission, however, constructing transmission network using data collected in settings with a high burden of TB will often be incomplete. The absence of cases from a network is concerning, because it may pose problems for making inferences about larger, complete transmission networks. We conducted a simulation study to examine the impact of different missing scenarios on the structure of partial transmission networks. We found that no single scenario we tested could account for the missingness in the partial network constructed using data from our TB transmission study. However, we found that missingness was unlikely to be random; rather, the most likely scenarios or combination of scenarios involved oversampling of low-transmitting cases or omission of a factor strongly related to transmission from our models. Our results were strongly influenced by several key assumptions, including the genomic threshold for transmission in the empirical network, the average number of transmissions per case (mean degree) in the full network, and the number of cases involved in the complete transmission network. Therefore, this analysis also highlights the uncertainties around the aspects of our model, and parameters of TB transmission more broadly, that preclude more exact inference regarding missing cases and thus underlying transmission patterns.

Introduction

Tuberculosis (TB) is the leading infectious cause of death worldwide, and in 2016, there were over half a million new cases of drug-resistant TB. [5] Extensively drug-resistant (XDR) tuberculosis is resistant to both first and second-line anti-TB drugs and accounts for 10% of all drug-resistant TB cases, representing an emergent but severe threat to global public health. Recent studies have found that the majority of drug-resistant TB cases are due to transmission of already-resistant strains, rather than inadequate treatment. This finding has underscored the importance of efforts to understand TB transmission patterns, as these patterns can inform the development of strategic interventions to reduce transmission.

Our ability to identify discrete TB transmission events has improved dramatically in light of the increasing availability of whole genome sequencing, which can resolve variation between *Mtb* sequences at the level of individual base pairs. Cases with similar *Mtb* sequences are likely to be linked through transmission; collectively, such links can be used to create networks of putative transmission events. Previous studies have descriptively mapped potential transmission links between cases using social contact, molecular data, or both, but there has been limited effort to systematically analyze and model TB transmission networks, which may provide critical new insight into local TB epidemics. [137, 140] Indeed, network-based approaches have been successfully used to enhance our understanding of the transmission dynamics of other infectious diseases, including HIV. [251, 254]

Analyzing transmission networks constructed using *Mtb* sequencing and epidemiologic data may be especially useful to understand drivers of transmission, but data missing from these networks may pose serious challenges for making conclusions about transmission patterns. In settings with a high burden of TB, collecting information on all cases would require immense resources, and as a result is unrealistic if not impossible. Therefore, networks constructed using empirical data will be ‘partial’, representing only a subset of cases and links present in the true, but unobservable, transmission network. Cases may be unreachable, and therefore missing from empirical networks, for many reasons. For example, poor survival after diagnosis may lead to challenges in ascertaining cases. This may be particularly important for studies of XDR TB, since the survival rate with treatment is low (28%). [16] Even if all transmission events between diagnosed cases can be captured, little will be known about transmission events among undiagnosed cases. Barriers to diagnosis of XDR TB, including the requirement of culture-based drug susceptibility testing, contribute to the inability to collect information on undiagnosed cases and transmission events that occur among them. If missing data causes the empirical network to poorly resemble to the true transmission network, patterns detected in empirical networks may not accurately represent underlying transmission patterns.

Thoughtful consideration of the nature and extent of missing data in a measured transmission network can provide insight into what a complete network, had it been measured, may have looked like. If cases are missing randomly, inference from a partial transmission network may be feasible. Previous studies have suggested that nodes or cases, missing at random from a network can have a minimal effects on network structure, and can often allow inference to be made about the complete network. [193] On the other hand, if missing cases are systematically different from sampled cases, this may pose a more serious challenge. The absence of nodes that are highly connected in a network can have a pronounced effect on network structure, which may reduce the validity of conclusions one can make from a partial network. [198] Systematic missingness may be especially severe if there are differences between sampled and unsampled cases with respect to transmission. For example, if undiagnosed cases are likely to have forms of TB that are traditionally more difficult to detect (e.g., smear-negative disease), they may experience diagnostic delays and thereby longer infectious periods, and be responsible for more transmission than sampled cases. [28] In addition to nature of missing data, the extent of missingness is also important for understanding the ability of a partial network to represent a larger one. Unsurprisingly, inference becomes increasingly challenging as more cases are missing from network. [196-198] Making explicit hypotheses about the mechanisms by which cases may be missing from a transmission network can help us to anticipate, and potentially obviate, the effects of missing data. Understanding these effects can inform inference from partial transmission networks, either by cautioning generalization of findings to broader, unsampled populations or by providing reassurance about the robustness of findings from a partial network to missing data.

In this study, we used data from a transmission study of extensively drug-resistant (XDR) TB cases in KwaZulu-Natal, South Africa to create an empirical transmission network based on *Mtb* sequence data. We simulated hypothetical, ‘complete’ transmission networks based on several different assumptions about how data was missing from the empirical network. We aimed to determine the type of missingness that was most consistent with the network we observed, in an effort to understand the extent to which the empirical network may reflect underlying XDR TB transmission patterns.

Methods

Study design and procedures

The Transmission of HIV-Associated XDR TB (TRAX) study is a cross-sectional study that enrolled culture-confirmed XDR TB patients diagnosed from 2011 to 2014 in KwaZulu-Natal province, South Africa. South Africa has among the highest rates of TB globally, driven in part by high HIV prevalence: 59% of TB patients are co-infected with HIV. [223] KwaZulu-Natal province has the highest TB and XDR TB burden (3 per 100,000) in South Africa. [215, 224, 225]

Detailed methods of the TRAX study have been previously published. [12] Briefly, we identified XDR TB cases through the single referral laboratory that conducts drug-susceptibility testing (DST) for all public healthcare facilities in the province. All participants provided written informed consent; for deceased or severely ill participants, consent was obtained from next-of-kin. We interviewed participants and performed medical record review to collect demographic and clinical information. The interview included a social network questionnaire that elicited information on locations frequented and close contacts prior to diagnosis.

Ethical Considerations

The study was approved by the Institutional Review Boards of Emory University, Albert Einstein College of Medicine, and the University of KwaZulu-Natal, and by CDC's National Center for HIV, Hepatitis, STDs and Tuberculosis.

Whole genome sequencing

The diagnostic XDR TB isolate was obtained for all enrolled participants and re-cultured on Löwenstein-Jensen slants. We conducted population sweeps, extracted genomic DNA, and prepared sequencing libraries using Nextera DNA kits (Illumina, San Diego, CA). Raw paired-end sequencing reads were generated on the Illumina (MiSeq) platform and aligned to the H37Rv reference genome (NC_000962.3) using the Burrows-Wheeler Aligner. All isolates had reads covering >99% of the reference genome, and the lowest mean coverage depth for any isolate was 15X. Single nucleotide polymorphisms (SNPs) were detected using standard pairwise resequencing techniques (Samtools v0.1.19) against the reference and filtered for quality, read consensus (>75% reads for the alternate allele) and proximity to indels (>50 base-pairs from any indel). SNPs in or within 50 base pairs of hypervariable PPE/PE gene families, repeat regions, and mobile elements were excluded. [227]

Constructing networks using Mtb *sequence data*

We defined a genomic link as a pair of XDR TB cases with 5 or fewer SNP differences between their *Mtb* sequences. We constructed sequencing-based transmission networks of TRAX cases, in which each node in the network represents an XDR TB case and each edge, or connection, between nodes in the network represents a genomic link. We calculated the degree of each node, or the number of genomic links per case, in the network; this also referred to as the degree distribution. All network analysis was completed using the *sna* and *statnet* packages in R.

We considered the sequencing-based network comprised of TRAX cases as our empirical, or measured, network. This empirical network is a partial network, in that it represents a subset of cases and transmission events sampled from the unobserved, complete transmission network that includes all XDR TB cases and transmission events in KwaZulu-Natal during the study period.

Defining missing data scenarios

We hypothesized three different scenarios under which cases may be missing from the empirical network. In the simplest scenario, we assumed that cases were missing at random (Scenario 1, Table 6-1). If cases are missing at random, we can reasonably confident about inferences from the empirical network. In the second scenario, we considered the possibility that we systematically oversampled cases that were either involved in many transmission events ('high-transmitters') or few transmission events ('low-transmitters') (Scenario 2, Table 6-1). This scenario encompasses a range of more specific hypotheses about *why* cases who were sampled may be more or less likely to transmit TB. For example, if unsampled, undiagnosed cases tended to have longer infectious periods, they may be responsible for more transmission events. Conversely, unsampled, undiagnosed cases may have been more likely to live in rural areas, have lower contact rates, and thus be responsible for less transmission. This scenario would be problematic, because it would indicate that the structure of the empirical network may be meaningfully different from the true transmission network. In the third scenario, we made the more specific hypotheses that cases were sampled differentially by HIV or smear status (Scenario 3, Table 6-1). For example, HIV-positive cases may be more likely to have smear-negative disease, which is difficult to diagnose and may thereby lead to diagnostic delays and longer infectious periods. Conversely, HIV-positive cases may be more closely linked with the healthcare system, and therefore more likely to be promptly diagnosed with TB. The implications of this scenario are similar to those of Scenario 2 but may provide additional insight as to the potential characteristics of missing cases.

In the fourth and last scenario, we did not make any assumptions about how cases were sampled, but rather accounted for the possibility that we failed to measure a factor that was strongly related to likelihood of transmission (Scenario 4, Table 6-1). Previous models included only characteristics of cases that we measured in TRAX, which we recognize may not sufficiently capture factors important for explaining variation in transmission potential across individuals. Thus, scenario 4 reflects the hypothesis that the variables we included in models to generate transmission networks were missing a component that meaningfully impacts transmission.

Fitting exponential random graph models to missing data scenarios

We used information collected in the TRAX study on demographic and clinical case characteristics considered to be related to transmission, including age, sex, HIV status, and clinical markers of infectiousness (cough duration and sputum smear status) to parameterize exponential random graph models (ERGMs) representing complete XDR TB transmission networks (Table 6-2, Technical Appendix). Using exponential random graph models, we modeled the probability of links, or transmissions, between the cases in the network as a function of demographic and clinical characteristics. We specified these models under each missing data scenario. We tested each of these scenarios across a range of values for the average number of transmissions per case, or the overall mean degree in the complete network, from 2 to 20. ERGMs were constructed using the *ergm* package in R, which is a part of the *statnet* suite of software.

From each model, we simulated 1,000 theoretical, complete transmission networks. To model complete transmission networks, we needed to estimate their size, which we assumed was the total number of diagnosed and undiagnosed XDR TB cases in KwaZulu-Natal during the study period (2011-2014). We used data from the South Africa Tuberculosis Drug Resistance Survey to estimate the number of diagnosed XDR TB cases and active case-finding studies to estimate the number of additional, undiagnosed cases (see Technical Appendix). [13] These sources led us to estimate a complete transmission network size of 2000 cases for our primary analyses.

Sampling from simulated networks

We mimicked our transmission study by sampling a similar number of cases (350) as in our empirical network from each simulated, complete network.

To determine which missing data scenario was most likely, we aimed to determine which scenario produced simulated, sampled networks that closely matched the empirical network. To compare the empirical network with the simulated, sampled networks, we used three key structural features from the empirical network as ‘target statistics’: i) the proportion of nodes with a degree of zero, or the number of unlinked cases in the network, ii) the maximum degree, or the number of links of the most highly linked case, and iii) the proportion of nodes with a degree greater than 10, or the proportion of cases with more than 10 links. Note that the first target statistic reflects the number of poorly connected cases and thus the left side of the degree distribution of the empirical network, while the second two target statistics capture highly connected cases, or the right side of the degree distribution. For each model, we calculated the proportion of simulated, sampled networks (out of 1,000 simulations) that matched each target statistic (Figure 6-1).

Sensitivity Analyses

The genomic threshold defining a direct TB transmission event is uncertain, so we constructed a second empirical, sequencing-based network using a more stringent SNP threshold (3 SNPs). We modified the target statistics accordingly and compared our results using this empirical network to our primary analysis. Additionally, we tested the sensitivity of our results to assumptions about the size of the complete transmission network. We examined the effect of decreasing (n = 1500) and increasing (n = 4000) the size of the simulated, complete transmission networks.

Results

Structure of the empirical, sequencing-based network

The empirical sequencing-based network comprised of 344 TRAX cases contained 1084 total genomic links. Each case had an average of 6.3 links (the overall network mean degree), and 182 (53%) cases in the network had at least one genomic link. The mean degree of cases with key clinical and demographic characteristics is shown in Table 6-1; corresponding characteristics of the empirical network defined by a 3 SNP threshold are shown in Supplemental Table 6-1.

We defined target statistics based on the following features of the empirical network: 162 (47%) cases were unlinked (i.e., degree of 0), the most highly linked case had 62 links (i.e., maximum degree), and 62 (18%) cases had 10 or more links (i.e., degree ≥ 10) (Table 6-1).

Missing case scenarios

The assumption that cases were randomly sampled from the complete network (Scenario 1) was generally inconsistent with the empirical network (Figure 6-2A). Several models that we tested under this scenario could reproduce the first target statistic: for example, the number of unlinked cases in the network was 48%, roughly matching the empirical network, when we assumed 5 transmissions per case (mean degree of 5) in the complete network (Table 6-3). However, none of the models assuming random sampling reproduced the second two target statistics: the number of links of the most highly linked case (maximum degree), or proportion of cases with more than 10 links (Table 6-3). In other words, models assuming random sampling could not account for the cases in the empirical network that were highly connected. To determine what a complete transmission network would need to look like to account for these highly connected cases under random sampling, we increased the average number of transmissions per case in the simulated, complete network until sampled networks reproduced these highly connected cases. We found that the average number of transmissions per case in the complete network needed to be unreasonably high (200) to reflect this feature of the empirical network (Supplemental Figure 6-1).

Oversampling of high- or low-transmitters (Scenario 2) significantly changed the structure of simulated, sampled networks, but still could not produce networks similar to the empirical network (Figure 6-2A). If high-transmitters were oversampled, sampled networks reached a maximum degree of 9.3, which is higher than under the assumption of random sampling (7.0) but still much lower than in the empirical network (62) (Table 6-3). The general structure of sampled networks under this assumption was also dissimilar to that of the empirical network: the assumption that high-transmitters were oversampled shifted the peak of the distributions higher than in the empirical network, which has a peak at 0 (Figure 6-2A). Ultimately, these models failed to reproduce all three target statistics simultaneously (Table 6-3). When we assumed that low-transmitters were oversampled, the overall shape of the degree distribution of simulated, sampled networks was more consistent with the empirical network, with its peak at 0. (Figure 6-2B). However, these models failed to reproduce any of the three target statistics from the empirical network: the proportion of unlinked cases was too high, ranging from 86-99%, and the maximum degree for all models was 2.1, far lower than the target statistic of 62. (Table 6-3). In other words, these models produced far more unlinked cases and far fewer highly linked cases than we observed in the empirical network.

Sampling cases differentially by HIV and smear status (Scenario 3) yielded few changes in the degree distributions of simulated, sampled networks (Figure 6-3). These models could reproduce the target statistic for the number of unlinked cases: the median proportion of unlinked cases across simulated, sampled networks was 26-38%, which approached the target statistic of 47% unlinked cases in the empirical network. However, models under the assumption of differential sampling by HIV or smear status failed to match the other two target statistics, demonstrating their inability to account for highly linked cases in the network (Table 6-4).

None of the first three scenarios, in which we made assumptions about the manner in which cases were sampled, produced networks similar to the empirical network. In Scenario 4, we instead made the assumption that our initial model was missing one or more individual-level case characteristics that might explain variation in the number of links per case that we observed in the empirical network. When we added an unmeasured variable strongly linked with transmission to the model, the resulting networks most closely reproduced the empirical distribution. Although these models could replicate only one of three target statistics (the number of unlinked cases), models including a characteristic very strongly linked with transmission (x40) resulted in networks with a higher maximum degree than any other model (9.8). (Table 6-4, Figure 6-3). Ultimately, however, this approach still could not produce networks that replicated the maximum degree in the empirical network (62) or the proportion of cases with more than 10 links (18%) (Table 6-4).

Sensitivity analyses

Larger complete transmission networks (n=4000 cases) yielded simulated, sampled networks that were more sparse with a lower mean degree, a higher proportion of unlinked cases, and fewer highly linked cases. Under the assumption of random sampling, models assuming 4000 total XDR TB cases were generally unable to reproduce the highly connected cases we observed in the empirical network. Complete transmission networks that were smaller than our primary models (n=1500) resulted in sampled networks that were more dense, with fewer unlinked cases and more highly linked cases. Simulated, sampled networks from these models could not reproduce all three target statistics simultaneously. However, if we assumed a high average number of transmissions per case (20), the resulting networks could reproduce the third target statistic, or the proportion of cases with more than 10 links in the empirical network (Figure 6-4, Supplemental Table 6-2). Thus, assuming fewer total XDR TB cases in the complete network resulted in networks more similar to the empirical network than assuming more XDR TB cases.

We also assessed the robustness of our results to the SNP threshold used to define a genomic link in the empirical network. Since a 3 SNP threshold required cases' isolates to be more closely related to define a link in the empirical network, the proportion of unlinked cases was higher, and both the maximum degree and the proportion of cases with more than 10 links was lower (Figure 6-5). Using a 3 SNP empirical network, some models were able to reproduce the target statistic for the number of cases with greater than 10 links under the assumption of random sampling, which was not possible using the 5 SNP empirical network (Supplemental Table 6-3). However, random sampling still could not simultaneously reproduce all three target statistics of the 3 SNP empirical network. To reproduce the maximum degree in the 3 SNP empirical network, the number of average transmissions per case in the complete network still needed to be unreasonably high (50) (Supplemental Figure 6-2). When we accounted for an unmeasured factor strongly associated with transmission, the resulting sampled networks could meet two of three target statistics of the empirical 3 SNP network (Supplemental Table 6-4). However, even in the empirical 3 SNP network, all models we tested still failed to reach the target statistic for maximum degree. In other words, even at a more stringent SNP threshold for transmission in the empirical network, our models could still not account for the highly linked cases we observed in our transmission study.

Discussion

Studies of endemic disease transmission will yield only partial transmission networks, from which we aim to draw conclusions about population-level transmission patterns. We found that no single missing data scenario we tested could account for the missingness in the partial network constructed using data from our TB transmission study. However, we found that missingness was unlikely to be random; rather, the most likely scenarios or combination of scenarios involved oversampling of low-transmitting cases or omission of a factor strongly related to transmission from our models. Although our initial goal was to estimate the relative plausibility of missing data scenarios, our results were strongly influenced by several key assumptions. These assumptions included the genomic threshold for transmission in the empirical network, the average number of transmissions per case (mean degree) in the complete network, and the number of cases involved in the complete transmission network. Therefore, this analysis also highlights the uncertainties around the aspects of our model, and parameters of TB transmission more broadly, that preclude more exact inference regarding missing cases and thus underlying transmission patterns.

The first scenario, in which cases were missing at random, was unlikely based on our models. This finding was robust to our choice of SNP threshold for the empirical network, assumptions about the size of the complete network, and across a range of values for the average number of transmissions per case in the complete network. Assuming that our models are correct, this finding suggests that inferences about transmission patterns from the empirical network of TRAX cases should be made with caution, as there may be important structural differences between the complete transmission network and the measured network.

In the second and third scenarios, we aimed to identify plausible reasons for systematic missingness that would provide insight into the characteristics of missing cases. In Scenario 3, we hypothesized that there may be differences with respect to the ability of the healthcare system to effectively diagnose cases with atypical clinical presentations (e.g., HIV-positive or smear-negative cases), and that as a result, these cases may be over- and under-represented in our study. However, when we examined whether this assumption alone could explain the partial network we observed, we found that it did not. When we assumed that we preferentially sampled high- or low-transmitting cases, regardless of their specific demographic or clinical characteristics, we found that this was also unlikely to be the only mechanism of missingness. However, the assumption that low-transmitting cases were more likely to be sampled than high-transmitting cases could produce networks similar to the observed network. Ultimately, these findings do not provide support for a specific mechanism by which cases are missing from the network, but do suggest the possibility that low-transmitting cases were oversampled in our transmission study (in other words, high-transmitters were undersampled). There are various reasons why this may have occurred: cases who are undiagnosed and thus untreated may have longer infectious periods, leading to more transmission. Alternatively, lifestyles or specific behaviors that lead to higher contact rates, including frequent cross-province travel for short-term employment opportunities, may also lead to a lower likelihood of diagnosis and retention in care. Previous research has suggested that this may be an important driver of both TB and HIV transmission in KwaZulu-Natal. [230, 255-257]

Although these other factors may play an important role in transmission, our network models only included clinical (cough duration, smear status, HIV status) and demographic factors (age, sex) related to transmission. In the fourth scenario, we tested the hypothesis that our models were missing a factor strongly related to transmission. We found that networks including this factor were most consistent with the empirical network, suggesting that our initial models may have been missing important variables that would explain sources of inter-individual variation in transmission among cases. This ‘unmeasured’ factor may reflect strongly suspected but yet understudied sociocultural or behavioral factors driving transmission in high-incidence settings, including cross-province travel for employment or frequent use of public transport. [129, 253, 258] However, recent research has also suggested there may be previously unrecognized, inherent biological features critical in explaining interindividual variation in transmission; for example, that certain individuals may be more able to generate infectious, *Mtb*-containing aerosols independent of their clinical disease presentation. [42, 44]

Our results were strongly influenced by factors about which there is a substantial degree of uncertainty in this setting, including the SNP threshold used to define a direct transmission event between two cases as well as key TB transmission parameters. We used a 5 SNP threshold for our primary analysis, as it has been used to define transmission in previous studies, but recognize that this threshold is not universal and likely depends on local TB epidemiology. [151] An empirical sequencing-based network based on a 3 SNP threshold contained many fewer genomic links and was consistent with a wider range of tested models than the empirical network based on a threshold of 5 SNPs. This result emphasizes the challenge of relying upon pairwise genomic distances to define transmission events: conclusions regarding transmission can be vastly different based on the threshold being used. The recent development of probabilistic methods to identify transmission is a promising step towards being able to construct more accurate empirical transmission networks based on genomic evidence from *Mtb* patient samples. [259] Future research should utilize these new approaches to more accurately define transmission events between cases.

Our results also varied substantially based on key transmission parameters. Our primary models varied the average number of transmissions per case (mean degree) in the complete network from 2 to 20. This range was selected after considering the range of previous estimates of the effective reproduction number of TB (R_{eff}), as there is evidence that this number may vary by setting and we did not find any studies estimating this parameter specifically for XDR TB in South Africa. [33] Interestingly, the models most consistent with the empirical network had a mean degree of 10 and above, which is substantially higher than most previous estimates of the R_{eff} of TB. Further studies on the transmission dynamics of drug-resistant TB in high-incidence settings should better characterize these foundational epidemiologic parameters to benefit future quantitative modeling studies. Lastly, changing our assumptions about the size of the complete transmission network dramatically changed the structure of sampled networks. Underdiagnosis of TB is a persistent challenge in low-resource settings and is even more difficult for XDR TB, which requires culture-based drug susceptibility testing. As a result, it is difficult to know the true number of XDR TB cases involved in transmission during the time period of our study. Our finding that larger complete networks were less likely to match the empirical network suggests that it is unlikely we significantly underestimated the number of XDR TB cases in KwaZulu-Natal in our primary analysis. However, the results from this sensitivity analysis underscore the broader challenge of diagnosing drug-resistant TB in low-resource settings, understanding the true magnitude of disease burden, and using this information to accurately model population-level transmission dynamics.

Ultimately, we could not make definitive conclusions about the type of missing data most relevant for our transmission study and none of our models reproduced all target statistics of the empirical network. This could be due to a failure to accurately define the empirical transmission network, and more sophisticated methods to define genomic transmission links, as described above, may be warranted in similar future analyses. However, our inability to match the empirical network may also be due to the assumptions of the models we used. ERGMs utilize Poisson distributions to describe the number of links per network node, and this distribution may fail to capture inherent properties of TB transmission. Indeed, recent studies have shown that the R_{eff} of TB, represented in our models by the mean degree in the network, may be best represented by a negative binomial distribution. This distribution may better fit the structure of the empirical network, and specifically, better account for highly linked cases. We could not force our network models to produce sampled networks that followed a negative binomial distribution, but further work could investigate the fit of models with different, or fewer, distributional assumptions.

This study has two key limitations. We assumed that the relative mean degree of sampled cases with specific attributes was similar to that of unsampled cases; for example, the relative mean degree of HIV-positive cases and HIV-negative cases that we measured in our study was equivalent to the ratio among unsampled cases. This may not be true if sampled cases are systematically different from unsampled cases with respect to their relative transmission potential. For example, this may have occurred if sampled HIV-positive cases were more likely virologically suppressed and therefore had similar transmission potential to HIV-negative cases, relative to unsampled HIV-positive cases. Second, we did not distinguish the direction of transmission in modeled or empirical networks. Incorporating directionality of transmission would have complicated network models and required specification of many additional parameters about which we had low certainty. It is important to note that for a given case, every link except one (corresponding to the source case) should theoretically correspond to forward transmission. This decision to model undirected networks was made to maintain the simplicity of models while ensuring that we didn’t need to specify many parameters about which we were uncertain; we believe that it does not reduce the validity of our models.

Constructing and analyzing transmission networks can provide critical insight into disease epidemiology and suggest potential avenues for intervention to reduce the spread of disease. Ironically, while a clearer understanding of transmission is perhaps most important in settings with a high burden of disease, sparse data in these settings also poses serious challenges for interpretation of transmission studies. This is, to our knowledge, the first study to use network modeling approaches to understand TB transmission, and the first to explicitly define and assess the support for different mechanisms of missingness in a study of TB transmission. We hope that this analysis lays the foundation for future efforts to better understand the important and complex role of missing data in TB transmission networks. Further, we hope that it has highlighted gaps in our understanding of TB transmission that at present hinder modeling efforts, but provided improved estimates can be generated, would enhance our ability to build models of TB transmission and use them to further understand disease dynamics.

6.2 Figures and Tables

Table 6-1. Missing case scenarios.

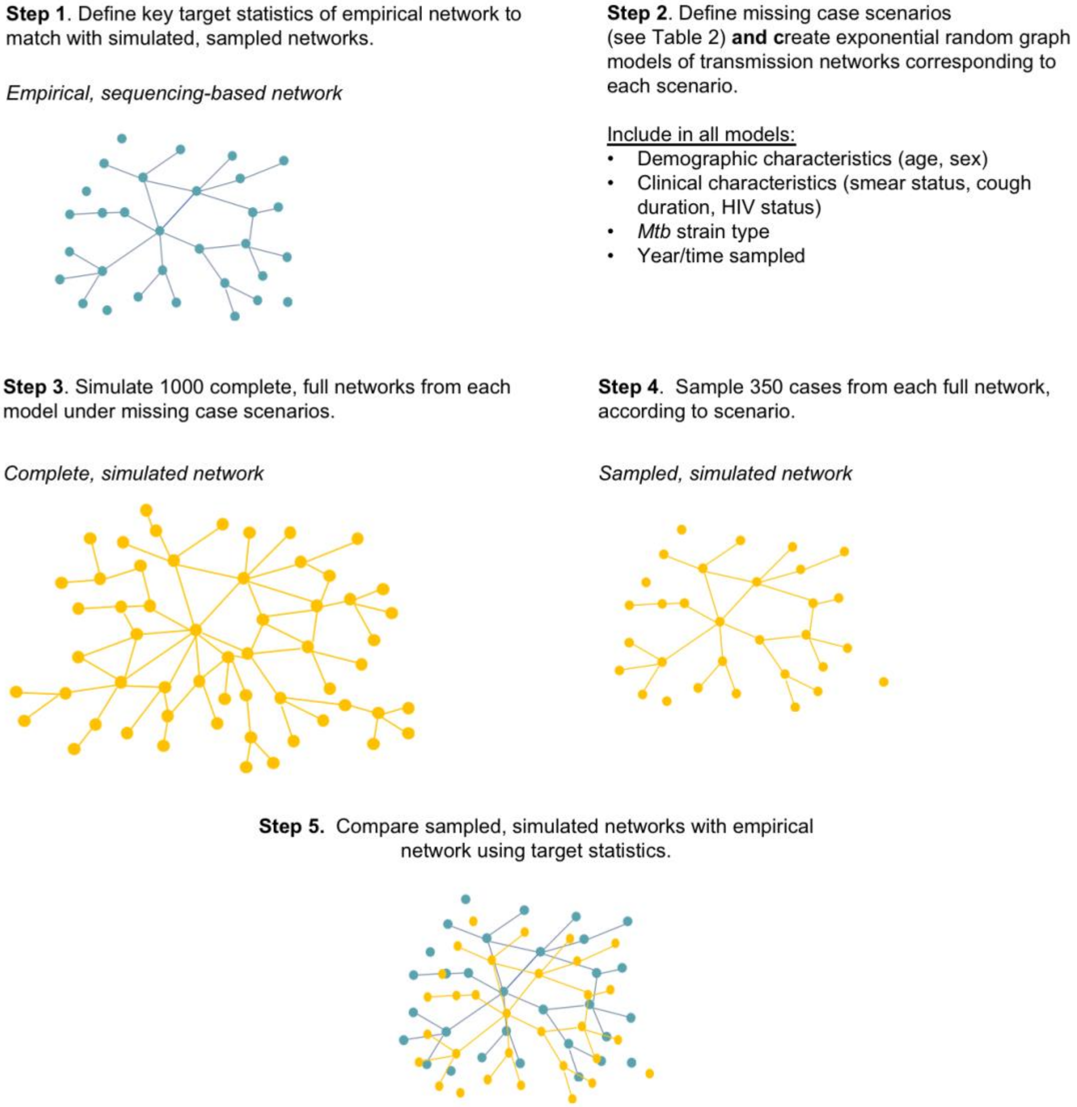
Scenario	**Complete transmission network model**
1. Cases missing at random	No changes to model terms.
2. Cases missing by connectivity	No changes to model terms. Sample from complete network nonrandomly using degree to define sampling weights. **a**. Highly connected cases ('high-transmitters') more likely to be sampled: sampling weighted by degree **b**. Poorly connected cases ('low-transmitters') more likely to be sampled: sampling weighted by inverse degree
3. Cases missing by attribute	No changes to model terms. Vary distribution of cases in complete network with attribute relative to empiric network. **a**. Vary distribution of HIV status **b.** Vary distribution of smear status
4. Unmeasured factor	Add model term corresponding to strong, unmeasured factor in a minority of cases. Vary strength and prevalence of factor. **a**. latent factor that increases transmission by factor of 10 (prevalence: 10%, 20%, 30%) **b**. latent factor that increases transmission by factor of 20 (prevalence: 10%, 20%, 30%) **c**. latent factor that increases transmission by factor of 40 (prevalence: 10%)

Table 6-2. Descriptive characteristics of the empirical sequencing-based network of XDR TB cases from the TRAX study.

	n (%)	Mean
Overall		
Edges (genomic links)	1084	-
Isolates (unlinked cases)	162 (47)	-
Overall mean degree	-	6.3
Maximum degree	62	-
Nodes with degree ≥ 10	62 (18)	-
By attribute		
HIV status		
HIV - negative	78 (23)	6.2
HIV - positive, undetectable viral load	133 (39)	6.7
HIV - positive, detectable viral load	133 (39)	5.9
Cough duration		
No cough	128 (37)	5.0
1mo	60 (17)	6.5
2mo	51 (15)	8.1
3mo	72 (21)	8.1
4mo	16 (5)	4.6
≥ 5mo	17 (5)	3.7
Smear status/grade		
Negative	109 (32)	6.9
Scanty positive	37 (11)	8.4
Positive, grade 1	59 (17)	4.6
Positive, grade 2	51 (15)	6.4
Positive, grade 3+	88 (26)	5.7
Sex		
Female	202 (59)	6.1
Male	142 (41)	6.4
Age category		
≤ 15	12 (3)	7.8
16-34	171 (50)	5.9
35-54	134 (39)	6.4

≥ 55	27 (8)	7.9
TB Strain		
LAM4	259 (75)	8.3
Other	85 (25)	0.2
Year		
2011	58 (17)	8.1
2012	107 (31)	5.6
2013	82 (24)	5.8
2014	97 (28)	6.4

Figure 6-1. Schematic representation of simulation and sampling methods.



Step 1. Define key target statistics of empirical network to match with simulated, sampled networks.

Empirical, sequencing-based network

Step 2. Define missing case scenarios (see Table 2) **and c**reate exponential random graph models of transmission networks corresponding to each scenario.

Include in all models:

- Demographic characteristics (age, sex)
- Clinical characteristics (smear status, cough duration, HIV status)
- *Mtb* strain type
- Year/time sampled

Step 3. Simulate 1000 complete, full networks from each model under missing case scenarios.

Complete, simulated network

Step 4. Sample 350 cases from each full network, according to scenario.

Sampled, simulated network

Step 5. Compare sampled, simulated networks with empirical network using target statistics.

Figure 6-2. Degree distributions of simulated, sampled networks under scenarios (1) and (2) compared to the empirical network.

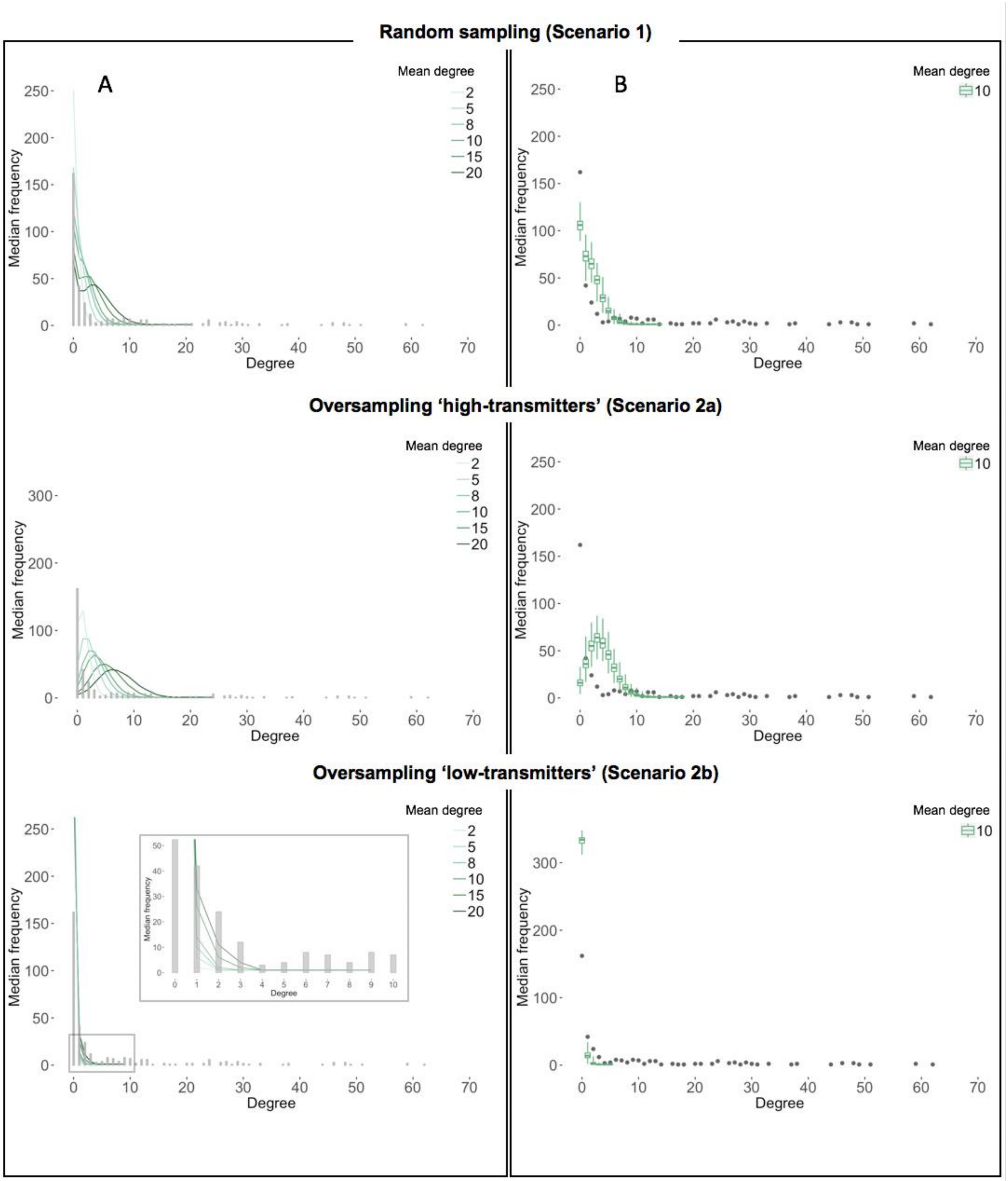


Figure 6-2. Degree distributions of empirical (≤ 5 SNPs) and simulated, sampled networks under scenarios 1 and 2. **A**. Grey bars show the distribution of the number of links per case, or the degree distribution, of the empirical network (≤ 5 SNPs) from the TRAX transmission study. Each colored line shows the median degree distribution across 1000 simulated, sampled networks for the corresponding model. Line color indicates the mean degree, or the average number of transmissions per case, assumed in the complete, simulated network. **B**: Range of the degree distributions of the simulated, sampled networks for one model (mean degree = 10). Grey dots show the degree distribution of the empirical network (≤ 5 SNPs) from the TRAX transmission study and are equivalent to the distribution shown by the grey bars in panel A. Colored boxplots show the median, interquartile range, minimum, and maximum frequencies for each degree in the distribution across 1000 simulated, sampled networks.

Table 6-3. Target statistics of simulated, sampled networks under scenarios (1) and (2) compared to the empirical network.

	Target statistic 1		**Target statistic 2**		**Target statistic 3**	
Mean degree	Average proportion[1] of isolates in sampled networks	Proportion of sampled networks with 40-60% isolates	Average maximum degree in sampled networks	Proportion of sampled networks with maximum degree > 40	Average proportion of nodes with degree > 10 in sampled networks	Proportion of sampled networks with > 10% of nodes with degree > 10
Random sampling (Scenario 1)						
2	0.72	0	1.74	0	0	0
5	0.48	0.99	2.82	0	0	0
8	0.35	0.02	3.77	0	0	0
10	0.30	0	4.41	0	0	0
15	0.23	0	5.44	0	0.003	0
20	0.20	0	7.03	0	0.018	0
Preferential sampling of high transmitters (Scenario 2a)						
2	0.32	0.01	2.69	0	0	0
5	0.14	0	4.09	0	< 0.001	0
8	0.07	0	5.29	0	0.002	0
10	0.05	0	6.01	0	0.007	0
15	0.02	0	7.59	0	0.048	0
20	0.01	0	9.29	0	0.154	0.988
Preferential sampling of poor transmitters (Scenario 2b)						
2	0.99	0	0.37	0	0	0
5	0.98	0	0.63	0	0	0
8	0.97	0	0.87	0	0	0
10	0.95	0	1.09	0	0	0
15	0.90	0	1.62	0	0	0
20	0.86	0	2.08	0	< 0.001	0

[1] 1,000 networks were simulated from each model, each simulated network was sampled once.

Figure 6-3. Degree distributions of simulated, sampled networks under scenarios (3) and (4) compared to the empirical network.

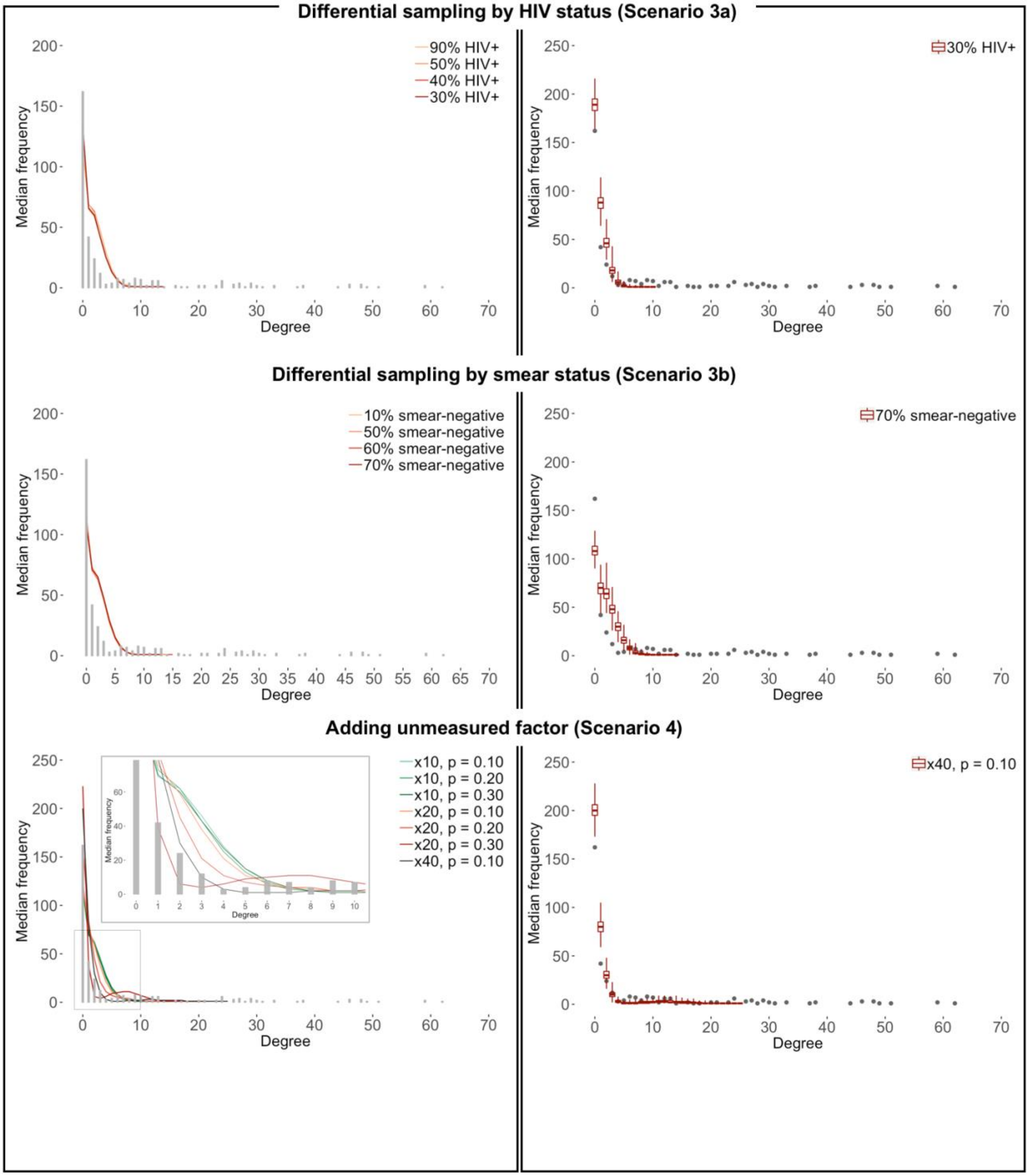


Figure 6-3. Degree distributions of empirical (≤ 5 SNPs) and simulated, sampled networks under scenarios 3 and 4. All models shown assume an average mean degree in the complete network of 10. **A**. Grey bars show the distribution of the number of links per case, or the degree distribution, of the empirical network (≤ 5 SNPs) from the TRAX transmission study. Each colored line shows the median degree distribution across 1000 simulated, sampled networks for the corresponding model. Line color indicates the distribution of HIV and smear status (Scenario 3) or the strength and prevalence of the unmeasured factor (Scenario 4) assumed in the complete, simulated network. **B**: Range of the degree distributions of the simulated, sampled networks for an individual model. Grey dots show the degree distribution of the empirical network (≤ 5 SNPs) from the TRAX transmission study and are equivalent to the distribution shown by the grey bars in panel A. Colored boxplots show the median, interquartile range, minimum, and maximum frequencies for each degree in the distribution across 1000 simulated, sampled networks.

Table 6-4. Target statistics of simulated, sampled networks under scenarios (3) and (4) compared to the empirical network.

Model [1]	**Target statistic 1**		**Target statistic 2**		**Target statistic 3**	
	Average proportion [2] of isolates in sampled networks	Proportion of sampled networks with 40-60% isolates	Average maximum degree in sampled networks	Proportion of sampled networks with maximum degree > 40	Average proportion of nodes with degree > 10 in sampled networks	Proportion of sampled networks with > 10% nodes with degree > 10
Cases sampled preferentially by HIV status (Scenario 3a) [3]						
10/90 HIV - / +	0.33	0	4.31	0	< 0.001	0
50/50 HIV - / +	0.38	0.17	4.13	0	< 0.001	0
60/40 HIV - / +	0.26	0	5.74	0	0.004	0
70/30 HIV - / +	0.38	0.15	4.15	0	< 0.001	0
Cases sampled preferentially by smear status (Scenario 3b) [4]						
30/70 smear - / +	0.31	0.31	4.24	0	< 0.001	0
50/50 smear - / +	0.33	0.33	4.43	0	0	0
70/30 smear - / +	0.31	0.31	4.46	0	0	0
90/10 smear - / +	0.32	0.32	4.22	0	< 0.001	0
Unmeasured factor (Scenario 4)						
10x, p = 0.10	0.33	0	4.72	0	< 0.001	0
10x, p = 0.20	0.35	0.001	4.53	0	< 0.001	0
10x, p = 0.30	0.34	0.002	4.76	0	< 0.001	0
20x, p = 0.10	0.36	0.017	6.38	0	0.006	0
20x, p = 0.20	0.46	0.995	6.04	0	0.005	0
20x, p = 0.30	0.64	0.017	7.54	0	0.039	0
40x, p = 0.10	0.57	0.868	9.76	0	0.005	0

[1] All scenario 3 and 4 models shown assume a mean degree in the complete network of 10.

[2] 1,000 networks were simulated from each model, each simulated network was sampled once.

[3] HIV distribution among TRAX cases (in empirical network): 23% HIV-negative, 77% HIV-positive.

[4] Smear distribution among TRAX cases (in empirical network): 32% smear-negative, 68% smear-positive.

Figure 6-4. Effect of modifying complete network size on network models under random sampling.

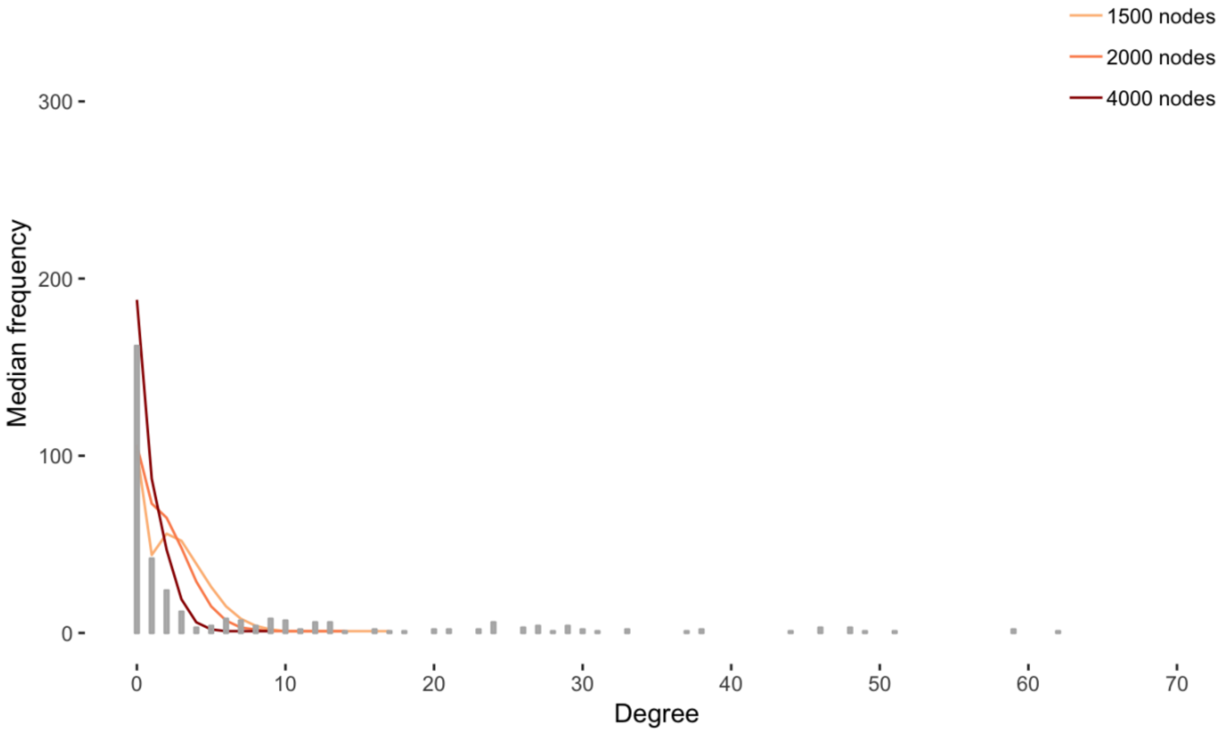


Figure 6-4. Degree distributions of empirical (≤ 5 SNPs) and simulated, sampled networks under different scenarios Grey bars show the distribution of the number of links per case, or the degree distribution, of the empirical network (≤ 5 SNPs) from the TRAX transmission study; colored line shows the median degree distribution across 1000 simulated, sampled networks for the corresponding model. Each model makes a different assumption about the total number of XDR TB cases involved in the transmission network during the time period of our transmission study (2011-2014), or the size of the simulated, complete transmission network. The model shown has a mean degree in the complete network of 10.

Figure 6-5. Effect of reducing SNP threshold (≤ 3 SNPs) on the empirical network.

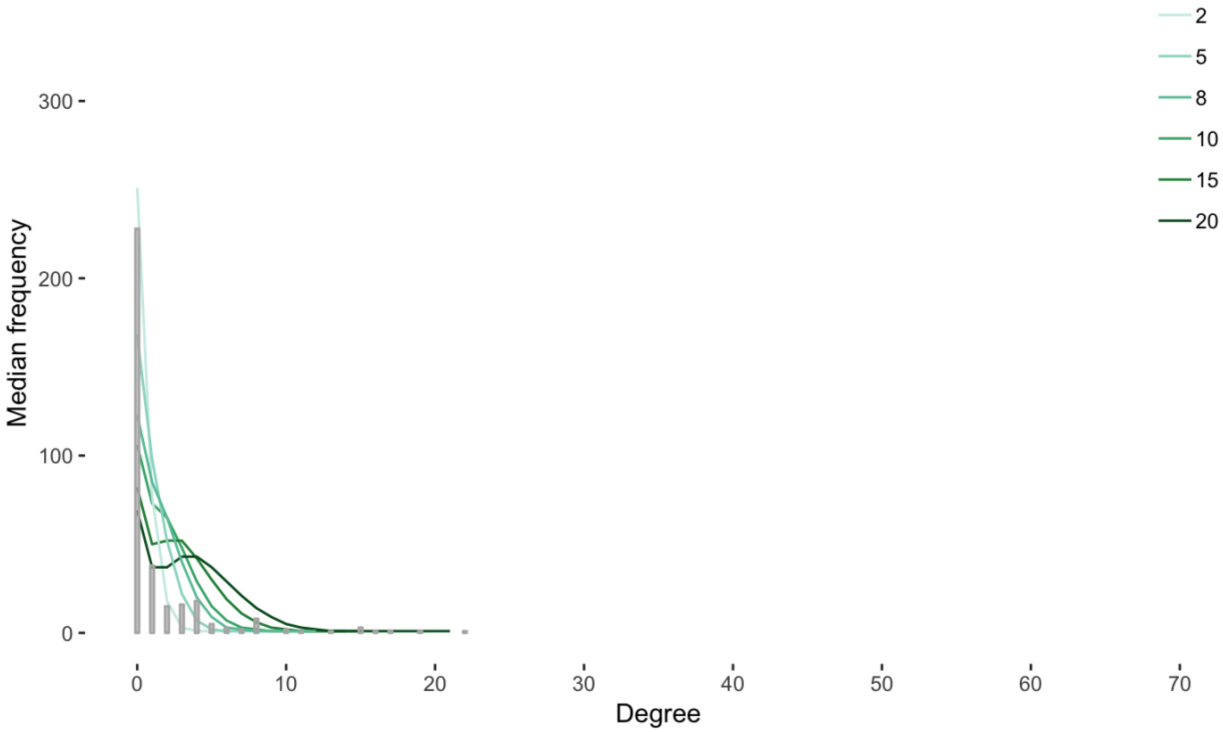


Figure 6-5. Degree distributions of empirical (≤ 3 SNPs) and simulated, sampled networks under scenarios 1 and 2. Grey bars show the distribution of the number of links per case, or the degree distribution, of the empirical network (≤ 3 SNPs) from the TRAX transmission study. Each colored line shows the median degree distribution across 1000 simulated, sampled networks for the corresponding model. Line color indicates the mean degree, or the average number of transmissions per case, assumed in the complete, simulated network.

6.3 Technical Appendix

Technical Appendix Table of Contents

1 Introduction

1.1 Model framework

1.2 Model software

2 Empirical Data

3 Defining models using missing case assumptions

3.1 Cases missing at random

3.2 Cases missing by level of connectivity

3.3 Cases missing by HIV/smear status

3.4 Unmeasured factor contributing to transmission

4

5 Clinical measures

5.1 Cough duration

5.2 Smear status

5.3 HIV

5.4 Mtb *strain type*

6 Demographic measures

6.1 Age

7 Joint distributions

7.1 Age and smear status

7.2 Age and HIV

7.3 Other (Smear status and HIV)

8 Simulation and sampling methods

9 Sensitivity analyses

9.1 Genomic threshold for transmission

9.2 Complete network size

--

1 Introduction

This technical appendix describes the models used in the associated manuscript, including their conceptual basis and parameterization as well as simulation procedures and statistical analysis.

1.1 Model framework

The network models in this study were used to represent and simulate transmission networks of active tuberculosis (TB) cases. Links, or edges, in modeled networks represent a transmission event that occurred between two cases in the network.

Modeled networks do not involve individuals (1) *infected* with TB but whom did not progress to active disease or (2) exposed contacts of TB cases. Rather, modeled networks reflect all transmission events observed from a sample of cases enrolled in a study over a defined four-year period.

Cases in each modeled network were assigned specific attributes according to a defined distribution. (For example, we specified that 70% of cases in the network were HIV-positive.) Assigned attributes of each case influence the number of other cases to whom that case is connected in the network. (More detail is provided in the Empirical Data section.)

We included each attribute as a 'nodefactor' term in the network model. (Some nodefactor terms represented joint distributions of two attributes, see Joint Distributions section.) We defined the target statistics for each nodefactor term as the number of edges in the network involving nodes with that attribute. (For example, the target statistic for the HIV nodefactor term was the number of edges involving HIV-positive nodes, or the number of transmission events in the network involving HIV-positive cases.)

We calculated target statistics for each nodefactor term using empirical data from the TRAX network (see Empirical Data section below).

1.2 Model software

All models were programmed in R. The code used to create and analyze these models is available on GitHub (user: kbratnelson).

The modeling methods employed in this study utilized the *ergm* R package, which requires the *statnet* suite of software.

2 Empirical Data

The Transmission of XDR TB (TRAX) study is a cross-sectional study that enrolled 404 XDR TB cases from KwaZulu-Natal province, South Africa from 2011-2014. This study collected clinical, demographic and social network data.

The empiric transmission network was created from 344 cases with available whole genome sequencing results of their *Mtb* isolates. We created sequencing-based networks using pairwise differences between *Mtb* sequences. We considered fewer than 5 single nucleotide polymorphisms a transmission link and constructed an undirected network. We also considered a more stringent SNP threshold (see Sensitivity Analyses section below).

In this empiric, undirected network, the maximum degree was 62, there were 162 (47%) of cases with no links (degree = 0), and 62 (18%) of cases with 10 or more links (degree ≥ 10).

To define the likelihood of being linked in the modeled networks based on a particular attribute, we used empiric data from the TRAX sequencing-based network. Specifically, we calculated the relative mean degree of cases with each attribute as compared to a reference group. (For example, we calculated the mean degree of HIV-positive relative to HIV-negative cases.)

To calculate the target statistics for each term in the network model, we multiplied the relative mean degree by the overall mean degree specific to the modeled network and the number of nodes in that network with the attribute. (For example, to model a network with an overall mean degree of 5, we multiplied the mean degree of HIV-positive relative to HIV-negative cases by 5 and the total number of HIV-positive cases in the network, to give the total number of edges associated with HIV-positive cases in the network.) See example calculation below. Models with target statistics specified for all levels of every variable did not easily converge, so we reduced the number of target statistics for variables with more than four categories. For these variables, we used the 3-4 categories corresponding to the highest number of edges as target statistics to parameterize models.

Using these target statistics, we simulated complete transmission networks from each model.

3 Defining models using missing case assumptions

We defined models and simulated complete transmission networks under several scenarios which made different assumptions about cases missing from the partial, empirical TRAX network.

3.1 Cases missing at random

We assumed that cases missing at random would result in missing transmission links randomly across the network. Therefore, to simulate complete transmission networks based on the assumption cases were missing at random, we used the same model but modified the mean degree in the complete network. We simulated complete networks with mean degrees of 2, 5, 8, 10, 15, and 20.

The mean degree in the network roughly corresponds to the number of cases whom were infected by another TB case in the network and progressed to disease during the study period. It is important to note that the modeled and simulated networks are undirected, so the direction of transmission is not indicated. Parameterizing models to include both risk factors for infection and transmission was beyond the scope of this project. However, it is important to note that Theoretically, one link per case in modeled networks corresponds not to a forward transmission event, but to the source case.

3.2 Cases missing by level of connectivity

We assumed two opposing scenarios: that cases who were highly connected in the complete network were more likely to be sampled, and that cases that were poorly connected in the complete network were more likely to be sampled. To simulate the former scenario, we created and used sampling weights proportional to each cases' degree in the complete network; to simulate the latter scenario, we created and used sampling weights inversely proportional to degree in the complete network.

We sampled cases using this method in complete networks with various mean degrees (2, 5, 8, 10, 15, 20).

3.3 Cases missing by HIV/smear status

We assumed that cases were undersampled, or oversampled, systematically based on their HIV or smear status. We modified the distribution of HIV status in the complete network relative to the empirical TRAX network to reflect each hypothesis. For example, in the empirical network, 70% of cases were HIV-positive. We created scenarios in which the proportion of HIV-positive cases in the complete network was 10%, 40%, and 90%. A smaller proportion of cases in the complete network relative to the empiric TRAX network (10%, 40%) reflects the assumption that HIV-positive cases were oversampled; a larger proportion (90%) reflects the assumption HIV-positive cases were undersampled.

We sampled cases from complete networks with varying distributions of HIV and smear status with various mean degrees (2, 5, 8, 10, 15, 20).

3.4 Unmeasured factor contributing to transmission

We hypothesized that a factor contributing strongly to transmission risk but that was not accounted for in our model might have a substantial impact on network structure. We hypothesized that such a factor might increase transmission by at least 10 times, that is, cases with this factor would be responsible for 10 times as many transmission events than those lacking this factor. We created a 'nodefactor' term in the model for this unmeasured factor at various strengths (10x, 20x, 40x) and varied its prevalence in the population of cases from 10 to 30%.

4 Size of complete networks

To simulate complete networks, it was necessary to make assumptions about the number of cases involved in XDR TB transmission over the time period 2011-2014. We estimated the number of diagnosed and undiagnosed XDR TB cases in KwaZulu-Natal province contributing to transmission using data from the South African National Tuberculosis Drug Resistance Survey.[13] We used active case-finding studies to estimate the proportion of TB cases in South Africa that are undiagnosed. [28]

332, 783 TB cases in SA in 2014

Proportion of cases with pulmonary TB (infectious form) = 0.89

Proportion of cases in KwaZulu-Natal province (area of study) = 0.31

Proportion of cases with XDR = 0.005

*332, 783 * (0.31) * (0.89) * (0.005) * 4 yrs = 1836 cases (736 - 2572)*

Accounting for underdiagnosis of TB cases [28], multiply by factor of 2:

*1836 cases * 2 = 3672 cases (1472 - 5144)*

We simulated networks assuming a complete network size of n = 2000, but also explored the impact of changing network size (see Sensitivity Analyses section below).

5 Clinical measures

5.1 Cough duration

We categorized cough duration by month. The marginal distribution is below:

Technical Appendix Table 6-5. Mean degree by cough duration.

Cough duration	n (%)	Mean degree
No cough	128 (37)	5.0
1 month	60 (17)	6.5
2 months	51 (15)	8.1
3 months	72 (21)	8.1
4 months	16 (5)	4.6
5 months	17 (5)	3.7

We used target statistics for the largest categories ‘No cough’, ‘1 month’, ‘2 months’, and ‘3 months’ in network models.

5.2 Smear status

Although both smear status and grade were available, we used only smear status (smear-positive and smear-negative) to reduce the number of model parameters. The marginal distribution is below:

Technical Appendix Table 6-6. Mean degree by smear status.

Smear status	n (%)	Mean degree
Negative	109 (32)	6.9
Positive	235 (68)	6.0

We used the joint distribution of age and smear status for model target statistics; see Joint Distributions section.

5.3 HIV

Although both HIV status and information on virologic suppression were available, we used only HIV status (HIV-positive and HIV-negative) to reduce the number of model parameters. The marginal distribution is below:

Technical Appendix Table 6-7. Mean degree by smear status.

HIV status	n (%)	Mean degree
Negative	78 (23)	6.2
Positive	266 (77)	6.3

We used the joint distribution of age and HIV status for model target statistics; see Joint Distributions section.

5.4 Mtb *strain type*

The dominant strain of XDR TB in KwaZulu-Natal is the LAM4 strain. There is evidence that the phenotype of this strain may lead to differences in its transmission and evolutionary rate. [150, 260] We categorized *Mtb* strains into LAM4 or non-LAM4. The marginal distribution is below:

Technical Appendix Table 6-8. Mean degree by *Mtb* strain type.

Mtb strain	n (%)	Mean degree
LAM4	259 (23)	8.3
Non-LAM4	85 (77)	0.2

6 Demographic measures

6.1 Age

We categorized age into four groups: 0-15, 16-34, 35-54, >55. The marginal distribution is below:

Technical Appendix Table 6-9. Mean degree by age.

Age category	n (%)	Mean degree
0 - 15	12 (3)	7.8
16 - 34	171 (50)	5.9
35 - 54	134 (39)	6.4
> 55	27 (8)	7.9

We used the joint distribution of age/HIV status and age/smear status for model target statistics; see Joint Distributions section.

7 Joint distributions

7.1 Age and smear status

Technical Appendix Table 6-10. Mean degree by age and smear status.

Age category	Smear status	n (%)	Mean degree
0 - 15	Negative	7 (2)	10.1
16 - 34	Negative	44 (13)	7.3

35 - 54	Negative	40 (12)	4.4
> 55	Negative	18 (5)	10.3
0 - 15	Positive	5 (1)	4.6
16 - 34	Positive	107 (31)	5.4
35 - 54	Positive	79 (23)	7.2
> 55	Positive	7 (2)	3.0

We used target statistics for the largest categories, ’16-34, Smear-negative’, ’35-54, Smear-positive’, ’16-34, Smear-positive’, and ’35-54, Smear-positive’ as target statistics for network models.

7.2 Age and HIV

Technical Appendix Table 6-11. Mean degree by age and HIV status.

Age category	HIV status	n (%)	Mean degree
0 - 15	Negative	5 (1)	12.8
16 - 34	Negative	41 (12)	4.4
35 - 54	Negative	15 (14)	9.9
> 55	Negative	17 (5)	5.2

0 - 15	Positive	7 (2)	4.3
16 - 34	Positive	130 (38)	6.3
35 - 54	Positive	119 (35)	5.9
> 55	Positive	10 (3)	12.4

We used target statistics for the largest categories,’16-34, HIV-positive’, and ’35-54, HIV-positive’ as target statistics for network models.

7.3 Other (Smear status and HIV)

Although smear-negative disease tends to be more common among HIV-positive TB cases, this was not the case in the empirical data. The proportion of cases with HIV was nearly equivalent among smear-positive and smear-negative cases and the proportion of smear-positive cases was nearly equivalent among HIV-positive and HIV-negative cases. Thus, we chose not to represent the joint distribution of smear status and HIV in model target statistics.

8 Simulation and sampling methods

From each network model, we simulated 1000 networks. We specified the following parameters of the Markov Chain Monte Carlo (MCMC) algorithm: we set the number of burn-in simulations as 100000, the MCMC interval as 5000, and the MCMC sample size as 10000.

We ensured that the MCMC algorithm used to estimate parameters for each model converged appropriately by checking for adequate mixing of the MCMC chain and sufficient exploration of parameter space using the mcmc.diagnostics function in the *ergm* package.

We sampled 350 cases from each simulated, complete network, mimicking sampling 350 cases in our TRAX study from the larger population of XDR TB cases. We compared the degree distributions of simulated, sampled sampled networks to that of the empirical TRAX network.

We attempted to 'match' three key features of the empirical degree distribution: (1) the proportion of nodes, or cases, that were unlinked; (2) the maximum degree of the network; (3) the proportion of cases with degree ≥ 10. We calculated the proportion of simulated, sampled networks from each model that matched target statistics. We also calculated the average proportion isolates, average maximum degree, and average proportion of cases with degree ≥ 10 from the set of simulated networks for each model.

9 Sensitivity analyses

9.1 Genomic threshold for transmission

Since the threshold for defining genomic evidence of transmission is not well-defined, we also defined an empirical network using a more stringent threshold of 3 pairwise SNP differences. This resulted in no changes to modeled networks, but did change the target statistics we attempted to 'match' with simulated, sampled networks.

9.2 Complete network size

We considered several other sizes of the complete network. We assumed that the complete network may be larger than 2000 cases (n = 4000 cases), or that it may be smaller (n = 1500 cases). We compared the results from these networks to our main models, which assumed a complete network size of 2000 cases.

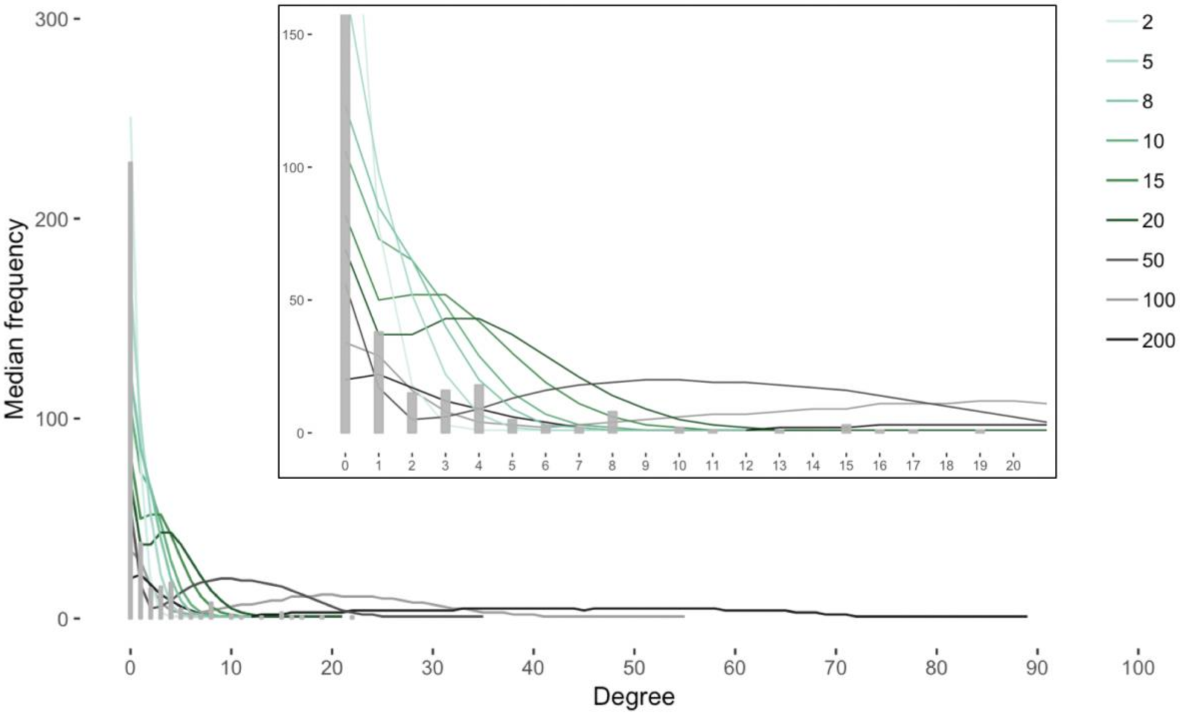
6.4 Supplemental Results

Supplemental Table 6-1. Descriptive characteristics, sequencing-based network of XDR TB cases in the TRAX study (≤ 3 SNP threshold).

	N (%)	Mean
Total network		
Edges	240	-
Isolates (Unlinked cases)	228	-
Degree		1.4
By attribute		
HIV status		
HIV-	78 (23)	1.15
HIV+, undetectable VL	133 (39)	1.50
HIV+, detectable VL	133 (39)	1.37
Cough duration		
No cough	128 (37)	1.30
1mo	60 (17)	1.35
2mo	51 (15)	1.75
3mo	72 (21)	1.69
4mo	16 (5)	0.06
5mo	17 (5)	0.71
Smear status/grade		
Negative	109 (32)	1.49
Scanty +	37 (11)	1.73
Positive, grade 1	59 (17)	1.05
Positive, grade 2	51 (15)	1.35
Positive, grade 3+	88 (26)	1.31
Sex		
Female	202 (59)	1.34
Male	142 (41)	1.42
Age category		
< 15	12 (3)	1.25
16-34	171 (50)	1.19

35-54	134 (39)	1.53
> 55	27 (8)	1.78
TB Strain		
HP	259 (75)	1.79
Other	85 (25)	0.09
Year		
2011	58 (17)	1.84
2012	107 (31)	1.36
2013	82 (24)	1.10
2014	97 (28)	1.34

Supplemental Figure 6-1. Mean degree required to reproduce the maximum degree in the empirical network.



Supplemental Figure 6-1 Mean degree required to reproduce the maximum degree in the empirical network. Grey bars show the distribution of the number of links per case, or the degree distribution, of the empirical network (≤ 5 SNPs) from the TRAX transmission study. Each colored line shows the median degree distribution across 1000 simulated, sampled networks for the corresponding model. Line color indicates the mean degree, or the average number of transmissions per case, assumed in the complete, simulated network.

Supplemental Table 6-2. Effect of modifying network size on network models.

	Target statistic 1		**Target statistic 2**		**Target statistic 3**	
Mean degree	Average proportion[1] of isolates in sampled networks	Proportion of sampled networks with 40-60% isolates	Average maximum degree in sampled networks	Proportion of sampled networks with maximum degree > 40	Average proportion of nodes with degree > 10 in sampled networks	Proportion of sampled networks with > 10% of nodes with degree > 10
Random sampling, complete network size = 1500						
2	0.66	0.012	2.0	0	0	0
5	0.43	0.881	3.3	0	0.00001	0
8	0.33	0	4.5	0	0.00032	0
10	0.30	0	5.2	0	0.00001	0
15	0.26	0	6.8	0	0.01619	0
20	0.25	0	8.5	0	0.06679	0.01
Random sampling, complete network size = 2000 (from Table 3)						
2	0.72	0	1.74	0	0	0
5	0.48	0.99	2.82	0	0	0
8	0.35	0.02	3.77	0	0	0
10	0.30	0	4.41	0	0	0
15	0.23	0	5.44	0	0.003	0
20	0.20	0	7.03	0	0.018	0
Random sampling, complete network size = 4000						
2	0.86	0	1.2	0	0	0
5	0.70	0	1.8	0	0	0
8	0.61	0.41	2.2	0	0	0
10	0.50	0.99	2.7	0	0	0
15	0.40	0.99	3.3	0	0	0

20	0.40	0.15	4.1	0	0	0.988

[1] 1,000 networks were simulated from each model, each simulated network was sampled once.

Supplemental Table 6-3. Effect of reducing the SNP threshold (≤ 3 SNPs) on the empirical network.

	Target statistic 1		**Target statistic 2**		**Target statistic 3**	
Mean degree	Average proportion[1] of isolates in sampled networks [2]	Proportion of sampled networks with 40-60% isolates	Average maximum degree in sampled networks	Proportion of sampled networks with maximum degree > 40	Average proportion of nodes with degree > 10 in sampled networks	Proportion of sampled networks with > 10% of nodes with degree > 10
Random sampling, ≤ 5 SNP threshold in empirical network (from Table 3)						
2	0.72	0	1.74	0	0	0
5	0.48	0.99	2.82	0	0	0
8	0.35	0.02	3.77	0	0	0
10	0.30	0	4.41	0	0	0
15	0.23	0	5.44	0	0.003	0
20	0.20	0	7.03	0	0.018	0
Mean degree	Average proportion[1] of isolates in sampled networks [2]	Proportion of sampled networks with **60-80%** isolates	Average maximum degree in sampled networks	Proportion of sampled networks with maximum degree > **15**	Average proportion of nodes with degree > 10 in sampled networks	Proportion of sampled networks with **> 2%** of nodes with degree > 10
Random sampling, ≤ 3 SNP threshold in empirical network						
2	0.72	0.99	1.74	0	0	0
5	0.48	0	2.82	0	0	0
8	0.35	0	3.77	0	0	0
10	0.30	0	4.41	0	0	0
15	0.23	0	5.44	0	0.003	0
20	0.20	0	7.03	0	0.018	0.317

[1] 1,000 networks were simulated from each model, each simulated network was sampled once.

[2] Note that model results are the same at both thresholds, but the target statistics and the proportion of models meeting target statistics are different.

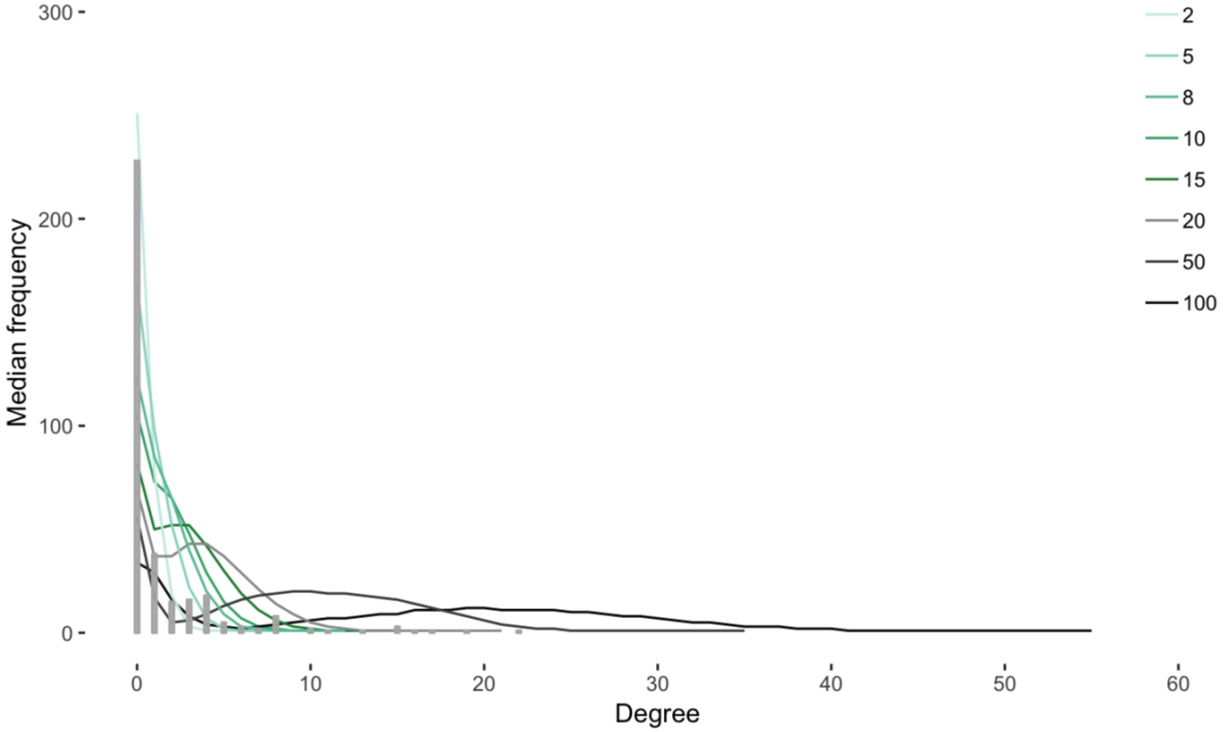
Supplemental Table 6-4. Effect of reducing the SNP threshold (≤ 3 SNPs) on the empirical network, considering an unmeasured factor.

Model [1]	**Target statistic 1**		**Target statistic 2**		**Target statistic 3**	
Mean degree	Average proportion[1] of isolates in sampled networks	Proportion of sampled networks with **60-80%** isolates	Average maximum degree in sampled networks	Proportion of sampled networks with maximum degree **> 15**	Average proportion of nodes with degree > 10 in sampled networks	Proportion of sampled networks with **> 2%** of nodes with degree > 10
Unmeasured factor (Scenario 4), ≤ 3 SNP threshold in empirical network						
10x, p = 0.10	0.33	0	4.7	0	0.001	0
10x, p = 0.20	0.35	0	4.5	0	< 0.001	0
10x, p = 0.30	0.34	0	4.8	0	0.001	0
20x, p = 0.10	0.36	0	6.4	0	0.006	0
20x, p = 0.20	0.46	0	6.0	0	0.005	0
20x, p = 0.30	0.54	0.974	7.5	0	0.039	0.988

[1] All models shown assume a mean degree in the complete network of 10.

[2] 1,000 networks were simulated from each model, each simulated network was sampled once.

Supplemental Figure 6-2. Mean degree required to reproduce the maximum degree in the ≤ 3 SNP empirical network.



Supplemental Figure 6-2. Degree distributions of empirical (≤ 3 SNPs) and simulated, sampled networks under different scenarios. Grey bars show the distribution of the number of links per case, or the degree distribution, of the empirical network (≤ 3 SNPs) from the TRAX transmission study. Each colored line shows the median degree distribution across 1000 simulated, sampled networks for the corresponding model. Line color indicates the mean degree, or the average number of transmissions per case, assumed in the complete, simulated network.

7. Public health implications and overall significance

The public health importance of this dissertation is twofold. First, describing relationships between individual-level characteristics and transmission can clarify the role of certain cases and behaviors in driving the spread of disease. The World Health Organization's End TB Strategy, which lays out guiding principles for the global effort to reduce the burden of TB, urges those charged with TB prevention and control to "know their epidemic". [261] Understanding the local forces contributing to ongoing transmission can indeed be transformative for the development of tailored and effective TB control policies. Second, identifying the limitations of using incomplete epidemiologic and sequencing data to make conclusions about population-level transmission patterns, especially in high-burden settings, is an important methodologic contribution to the field of TB transmission research. Not only does this study help us to better understand the TRAX cohort and its potential limitations for answering questions about transmission patterns, but it also provides a general framework for considering and assessing the impact of missing data in studies of partial disease transmission networks.

In Aims 1 and 2, we identified clinical characteristics, settings, and types of contact associated with transmission. In Aim 1, we found that the presence of cavitary disease on chest x-ray and reporting 2 to 3 months of cough were associated with being highly connected in the network. We also found that positive smear status was associated with poor connectivity in the network, which is contrary to the well-established notion that smear-positive cases more infectious and therefore responsible for more transmission than smear-negative cases. We hypothesized that this finding could be due to the fact that smear-negative cases are more likely to experience diagnostic delays and thus longer infectious periods, perhaps leading to more transmission than among smear-positive cases. In Aim 2, we found that time spent in urban settings was associated with being highly connected in the network, suggesting a role for urban settings in driving transmission that should be further explored. We also found that cases with extended hospital stays were less likely to be highly connected that those who did reported no or short hospital stays. Although hospitals are typically considered settings that present a high risk for transmission, this finding may reflect significantly reduced contact rates among hospitalized cases due to less time spent living, working and socializing in their communities during their infectious period.

In Aim 3, we developed and tested methods to examine the role of missing cases in the ‘partial’ transmission network that we constructed using data from the TRAX study. We found that it was unlikely that cases were missing at random from our transmission study, and that the most likely missing data scenario involved oversampling ‘low-transmitters’ or failing to account for an unmeasured factor contributing to transmission. However, our conclusions were heavily dependent on several assumptions that we made about the empirical network and key TB transmission parameters. The results from this Aim, although largely inconclusive, highlight broader uncertainty in the field about many features of TB transmission and epidemiology, including natural history parameters (e.g., the effective reproduction number of TB), metrics used to define transmission (e.g., SNP thresholds), and the true scale of drug-resistant TB epidemic in South Africa.

Our findings from Aim 3 provide important context for our findings in Aims 1 and 2. The conclusions from our network modeling studies raise uncertainty about whether cases in our study were randomly sampled from the larger population of XDR TB cases, and therefore suggest caution when interpreting associations detected in the empirical, sequencing-based network. More specifically, our results from Aim 3 suggest that our study may have oversampled cases responsible for few transmission events relative to those responsible for many transmission events. Previous research has suggested that the absence of highly connected, or central, cases in a network may have a substantial effect on network structure, which suggests that the network we constructed using *Mtb* sequencing data from TRAX cases may look materially different from the true, complete transmission network. [198] While our results from Aim 3 do not necessarily invalidate the associations we measured in Aims 1 and 2, it is unclear whether and to what extent missing cases may affect these associations. Ultimately, the results of Aim 3 suggest that the associations measured in the empirical network may not truly reflect underlying transmission patterns. These findings may also explain why we failed to see some expected associations with features known to be associated with transmission, including smear status.

Collectively, the findings from this dissertation join a growing body of evidence supporting the notion that casual contact may be driving TB transmission in high-incidence settings. In Aims 1 and 2, we found that urban contact, rather than more traditional risk factors of long hospital stays and many close contacts, were associated with transmission. These traditional risk factors are undoubtedly important and should remain targets of TB prevention and control efforts. However, when considered in the context of overall transmission, these settings may be *relatively* less important for transmission than settings in which extensive casual contact may occur. Although we did not directly measure rates of casual respiratory contact in the TRAX study, we did investigate the possibility that a factor we did not measure may contribute materially to the transmission patterns we observed. Indeed, in Aim 3, we found that an 'unmeasured' factor could potentially account for transmission heterogeneity not otherwise explained by the clinical and demographic factors included in our initial transmission network models. Rates of casual respiratory contact are in general difficult to quantify, but markers of high casual contact rates, including frequent inter-district or inter-province migration or regular use of crowded public transport, are measurable and may provide insight as to the specific behaviors that may be associated with high casual contact rates and thus transmission. If future research confirms an important role for casual contact in TB spread, it would have major implications for TB control efforts and provide clear rationale for development of interventions to reduce transmission through casual contact. These interventions could complement existing approaches to TB control that target primarily close contacts and institutional settings.

Although TB epidemiology tends to be setting-specific, there are several conclusions from these studies that may be generalizable outside of the XDR TB epidemic in South Africa. In Aim 1, we examined clinical features driving transmission. We expect that the effect of clinical features of TB disease on transmission is generally consistent across settings, since these associations are dependent upon human and pathogen physiology that is largely similar across populations. The exposures we studied in Aim 2, specifically contact with urban areas, may also be generalizable to settings outside of KwaZulu-Natal and South Africa. Circular urban-rural migration, in which individuals who live in rural areas travel frequently to and from urban areas for employment, is common in many rapidly urbanizing countries with high burdens of TB and may similarly drive transmission in other settings [108] Ultimately, however, each TB epidemic is unique and transmission patterns will be dependent upon the specific economic and social context in which disease transmission occurs. For example, the location, nature, and extent of close person-to-person contact in communities is likely very dependent upon local cultural norms, and therefore highly setting-specific. In this way, improved characterization of social milieu can inform community-based 'risk profiles' that can inform tailoring of local public health efforts. Indeed, research suggests that interventions targeting local 'catalysts' of transmission, or factors that increase contact rates and infectiousness, can be particularly effective in reducing TB incidence. [262] Understanding both universal and local drivers of TB transmission is important to furthering our understanding of TB transmission and the relative influences of biological and social factors in driving disease spread.

To our knowledge, the analytic methods used in this dissertation are a novel approach to studying patterns of endemic TB transmission. These methods fit into a broader landscape of recent research focused on the role of genomic data in improving our ability to understand and respond to infectious disease threats to public health. As pathogen genome sequencing becomes increasingly cheaper, and – perhaps most transformative for TB – methods are developed to sequence pathogens directly from patient samples, we will have access to unprecedented amounts of data that offer clues into the movement of pathogens through human populations. The availability of this data makes clear the need for new approaches to efficiently combine epidemiologic and genomic data to answer specific epidemiologic questions. Recently developed methods to resolve chains of transmission using genome sequences, including sophisticated Bayesian phylodynamic approaches, are highly computational and depend on a series of somewhat restrictive assumptions regarding the natural history parameters of TB. [263] Moreover, these methods are designed primarily for use in outbreak scenarios or low-incidence settings in which nearly all cases are sampled. [264-267] Developing approaches to understand *ongoing* transmission of endemic pathogens is critically important to further our understanding of TB transmission in high-incidence settings. To effectively capture the realities of measuring transmission in these settings, new methods must not only reconstruct transmission events but also consider the role of missing cases. Our work has aimed to fill this gap: although we use a relatively simple approach to defining transmission – pairwise SNP differences– we aimed to carefully consider the role of missing cases and the effect they may have on inference about underlying patterns of disease transmission. To the extent that new types of data integration and analytic approaches can be integrated into public health practice, they have the potential to revolutionize TB control practices.

The findings from this dissertation suggest several potential avenues for further research. First, we failed to find a complete set of factors accounting for the transmission heterogeneity suggested by our empirical transmission network. Specifically, we were unable to identify characteristics unique to cases whom were highly connected in the network. 'Superspreading' is known to be an important epidemiologic feature of other respiratory diseases and may also be critical to understanding TB transmission dynamics. [39, 41, 130, 246] Future studies explicitly aiming to identify factors related to superspreading may shed light on correlates of transmission risk that can serve as targets of intervention. Second, our studies in Aim 3 on missing cases were largely inconclusive, and our results were strongly influenced by several key TB transmission parameters on which there is a paucity of data in the literature.

This general lack of reliable data for TB transmission models has resulted in the slow application of sophisticated transmission modeling techniques in TB relative to other infectious diseases. The advent of next generation sequencing methods to understand disease dynamics may help to better define quantities, including the effective reproduction number, that will inform future quantitative modeling exercises. Additionally, initiatives which aim to better understand respiratory contact patterns in low and middle-income countries can provide critical data required to better model TB transmission in high-incidence settings. Lastly, while transmission studies in high-incidence settings struggle with a lack of data, there may be significant insight to be gained from low-incidence settings, where nearly all cases involved in transmission can be sampled. Scenarios in which the sampling fraction is high can provide an approximation of the 'complete' transmission networks that are so elusive in high-incidence settings. Moreover, highly detailed epidemiologic data is often more readily available in low-incidence settings, which can reliably identify putative transmission events and thus better define the threshold of genomic relatedness required to provide convincing evidence of transmission.

The global drug-resistant TB epidemic is an urgent threat to public health. Interventions to reduce transmission can be effective in reducing disease burden, but preventing transmission can seem like an overwhelming task: if transmission events are indeed incredibly difficult to identify, especially in TB-endemic settings, what chance do we have of preventing them? However, there are several reasons to be hopeful. First, interventions targeted towards very specific settings that present a high and sustained risk of casual contact are feasible and potentially effective approaches to preventing transmission. For example, implementing environmental controls in locations such as schools or minibus taxis may represent an efficient use of resources. [129, 258, 268] Should casual contact be definitively implicated as driving TB transmission in high-incidence settings, future studies should investigate the efficacy and cost-effectiveness of interventions targeting transmission through casual contact. Environmental controls can also reduce transmission risk in settings that are known to promote extended close contact, particularly hospitals. The efficacy of interventions to reduce TB transmission in institutional settings, including improvement of infection control practices, installation of UV lighting, and structural improvements to increase airflow, is well-documented. Second, reducing transmission can also be achieved indirectly by early detection and treatment of TB disease. By identifying locations and types of contact associated with transmission, exposed contacts at risk of TB disease can be identified and provided with preventive therapy, and those with early-stage TB disease can be promptly diagnosed and started on treatment. Fortunately, detection and treatment of XDR TB are two areas in which there have been recent promising advances. Detection of XDR TB is largely dependent on the capacity of the healthcare system and the efficiency with which patient samples are processed and undergo microbiological testing. At present, diagnosis of XDR TB still requires culture-based susceptibility testing which can take weeks to months, but point-of-care diagnostics for XDR TB will vastly improve the turnaround time required for diagnosis and are currently in the pipeline. Once a patient is diagnosed, they can immediately be started on treatment. Treatment of XDR TB is complicated and mortality rates remain high, but this is also an area undergoing rapid improvement. The increasing availability of bedaquiline and delaminid, in South Africa and elsewhere, has dramatically improved outcomes for XDR TB patients and offers new hope for preventing XDR TB transmission by treating, and curing, XDR TB disease. Collectively, these tools can reduce transmission of XDR TB in South Africa and globally.

8. References

1. Global Tuberculosis Report 2016. Geneva, Switzerland: World Health Organization, **2016**.

2. Wilson LG. The historical decline of tuberculosis in Europe and America: its causes and significance. Journal of the history of medicine and allied sciences **1990**; 45:366-96.

3. Yuen CM, Kammerer JS, Marks K, Navin TR, France AM. Recent Transmission of Tuberculosis — United States, 2011–2014. PloS one **2016**; 11:e0153728.

4. Marais BJ, Walker TM, Cirillo DM, et al. Aiming for zero tuberculosis transmission in low-burden countries. The Lancet Respiratory medicine **2017**; 5:846-8.

5. Global Tuberculosis Report. Geneva, Switzerland: World Health Organization, **2017**.

6. Bifani PJ, Plikaytis BB, Kapur V, et al. Origin and interstate spread of a New York City multidrug-resistant Mycobacterium tuberculosis clone family. Jama **1996**; 275:452-7.

7. Shah NS, Wright A, Bai G-H, et al. Worldwide emergence of extensively drug-resistant tuberculosis. Emerging infectious diseases **2007**; 13:380-7.

8. (CDC) CfDCaP. Emergence of Mycobacterium tuberculosis with extensive resistance to second-line drugs--worldwide, 2000-2004. MMWR Morbidity and mortality weekly report **2006**; 55:301-5.

9. Baleta A. S African president criticised for lack of focus on AIDS. Lancet (London, England) **2004**; 363:541.

10. Gandhi NR, Moll A, Sturm AW, et al. Extensively drug-resistant tuberculosis as a cause of death in patients co-infected with tuberculosis and HIV in a rural area of South Africa. Lancet (London, England) **2006**; 368:1575-80.

11. Gandhi NR, Moll A, Sturm AW, et al. Extensively drug-resistant tuberculosis as a cause of death in patients co-infected with tuberculosis and HIV in a rural area of South Africa. Lancet (London, England) **2006**; 368:1575-80.

12. Shah NS, Auld SC, Brust JCM, et al. Transmission of Extensively Drug-Resistant Tuberculosis in South Africa. New England Journal of Medicine **2017**; 376:243-53.

13. Ismail NA, Mvusi L, Nanoo A, et al. Prevalence of drug-resistant tuberculosis and imputed burden in South Africa: a national and sub-national cross-sectional survey. The Lancet Infectious Diseases **2018**; 18:779-87.

14. Mvusi L. Tuberculosis Burden in South Africa: South Africa National Department of Health, **2017**.

15. Pietersen E, Ignatius E, Streicher EM, et al. Long-term outcomes of patients with extensively drug-resistant tuberculosis in South Africa: a cohort study. Lancet (London, England) **2014**; 383:1230-9.

16. Shean K, Streicher E, Pieterson E, et al. Drug-associated adverse events and their relationship with outcomes in patients receiving treatment for extensively drug-resistant tuberculosis in South Africa. PloS one **2013**; 8:e63057.

17. Ramma L, Cox H, Wilkinson L, et al. Patients' costs associated with seeking and accessing treatment for drug-resistant tuberculosis in South Africa. The international journal of tuberculosis and lung disease : the official journal of the International Union against Tuberculosis and Lung Disease **2015**; 19:1513-9.

18. Verguet S, Riumallo-Herl C, Gomez GB, et al. Catastrophic costs potentially averted by tuberculosis control in India and South Africa: a modelling study. Lancet Glob Health **2017**; 5:e1123-e32.

19. Pooran A, Pieterson E, Davids M, Theron G, Dheda K. What is the Cost of Diagnosis and Management of Drug Resistant Tuberculosis in South Africa? PloS one **2013**; 8:e54587.

20. Global Tuberculosis Control 2009: Epidemiology, Strategy, Financing. Geneva, Switzerland: World Health Organization, **2009**.

21. Andrews JR, Noubary F, Walensky RP, Cerda R, Losina E, Horsburgh CR. Risk of progression to active tuberculosis following reinfection with Mycobacterium tuberculosis. Clinical infectious diseases : an official publication of the Infectious Diseases Society of America **2012**; 54:784-91.

22. Houben RMGJ, Dodd PJ, Jaramillo E, Williams B, Raviglione M, Dye C. The Global Burden of Latent Tuberculosis Infection: A Re-estimation Using Mathematical Modelling. PLoS medicine **2016**; 13:e1002152.

23. Behr MA, Edelstein PH, Ramakrishnan L. Revisiting the timetable of tuberculosis. BMJ **2018**; 362.

24. Pai M, Behr MA, Dowdy D, et al. Tuberculosis. Nature Reviews Disease Primers **2016**; 2:16076.

25. Dowdy DW, Basu S, Andrews JR. Is passive diagnosis enough? The impact of subclinical disease on diagnostic strategies for tuberculosis. American journal of respiratory and critical care medicine **2013**; 187:543-51.

26. Barry CE, Boshoff HI, Dartois V, et al. The spectrum of latent tuberculosis: rethinking the biology and intervention strategies. Nature reviews Microbiology **2009**; 7:845-55.

27. Corbett EL, Bandason T, Cheung YB, et al. Epidemiology of Tuberculosis in a High HIV Prevalence Population Provided with Enhanced Diagnosis of Symptomatic Disease. PLoS medicine **2007**; 4:e22.

28. Wood R, Middelkoop K, Myer L, et al. Undiagnosed tuberculosis in a community with high HIV prevalence: implications for tuberculosis control. American journal of respiratory and critical care medicine **2007**; 175:87-93.

29. Riley RLM, C C; Nyka, W; Weinstock, N; Storey, P B; Sultan, L U; Riley, MC; Wells, W F. Aerial dissemination of pulmonary tuberculosis: A two-year study of contagion in a tuberculosis ward. American Journal of Epidemiology **1959**; 70:185-96.

30. Dharmadhikari AS, Mphahlele M, Venter K, et al. Rapid impact of effective treatment on transmission of multidrug-resistant tuberculosis. The international journal of tuberculosis and lung disease : the official journal of the International Union against Tuberculosis and Lung Disease **2014**; 18:1019-25.

31. Escombe AR, Moore DAJ, Gilman RH, et al. The infectiousness of tuberculosis patients coinfected with HIV. PLoS medicine **2008**; 5:e188.

32. Salpeter EE, Salpeter SR. Mathematical model for the epidemiology of tuberculosis, with estimates of the reproductive number and infection-delay function. American journal of epidemiology **1998**; 147:398-406.

33. Ma Y, Horsburgh CR, White LF, Jenkins HE. Quantifying TB transmission: a systematic review of reproduction number and serial interval estimates for tuberculosis. Epidemiology and infection **2018**:1-17.

34. Biggerstaff M, Cauchemez S, Reed C, Gambhir M, Finelli L. Estimates of the reproduction number for seasonal, pandemic, and zoonotic influenza: a systematic review of the literature. BMC infectious diseases **2014**; 14:480.

35. Guerra FM, Bolotin S, Lim G, et al. The basic reproduction number (R0) of measles: a systematic review. The Lancet Infectious diseases **2017**; 17:e420-e8.

36. VanderWaal KL, Atwill ER, Isbell LA, McCowan B. Linking social and pathogen transmission networks using microbial genetics in giraffe, *Giraffa camelopardalis*. Journal of Animal Ecology **2014**; 83:406-14.

37. L. VK, O. EV. Heterogeneity in pathogen transmission: mechanisms and methodology. Functional Ecology **2016**; 30:1606-22.

38. Wong G, Liu W, Liu Y, Zhou B, Bi Y, Gao GF. MERS, SARS, and Ebola: The Role of Super-Spreaders in Infectious Disease. Cell host & microbe **2015**; 18:398-401.

39. Shen Z, Ning F, Zhou W, et al. Superspreading SARS events, Beijing, 2003. Emerg Infect Dis **2004**; 10:256-60.

40. Faye O, Boëlle P-Y, Heleze E, et al. Chains of transmission and control of Ebola virus disease in Conakry, Guinea, in 2014: an observational study. The Lancet Infectious Diseases **2015**; 15:320-6.

41. Lloyd-Smith JO, Schreiber SJ, Kopp PE, Getz WM. Superspreading and the effect of individual variation on disease emergence. Nature **2005**; 438:355.

42. Fennelly KP, Jones-López EC, Ayakaka I, et al. Variability of infectious aerosols produced during coughing by patients with pulmonary tuberculosis. American journal of respiratory and critical care medicine **2012**; 186.

43. Fennelly KP, Martyny JW, Fulton KE, Orme IM, Cave DM, Heifets LB. Cough-generated aerosols of mycobacterium tuberculosis: a new method to study infectiousness. American journal of respiratory and critical care medicine **2004**; 169.

44. Jones-López EC, Acuña-Villaorduña C, Ssebidandi M, et al. Cough Aerosols of Mycobacterium tuberculosis in the Prediction of Incident Tuberculosis Disease in Household Contacts. Clinical Infectious Diseases **2016**; 63:10-20.

45. Jones-Lopez EC, Namugga O, Mumbowa F, et al. Cough aerosols of Mycobacterium tuberculosis predict new infection: a household contact study. American journal of respiratory and critical care medicine **2013**; 187:1007-15.

46. Mathema B, Andrews JR, Cohen T, et al. Drivers of Tuberculosis Transmission. The Journal of infectious diseases **2017**; 216:S644-S53.

47. Shaw JB W-WN. Infectivity of pulmonary tuberculosis in relation to sputum status. American review of tuberculosis **1954**; 69:724-32.

48. Grzybowski S, Barnett GD, Styblo K. Contacts of cases of active pulmonary tuberculosis. Bulletin of the International Union against Tuberculosis **1975**; 50:90-106.

49. Kenyon TA, Creek T, Laserson K, et al. Risk factors for transmission of Mycobacterium tuberculosis from HIV-infected tuberculosis patients, Botswana. The international journal of tuberculosis and lung disease : the official journal of the International Union against Tuberculosis and Lung Disease **2002**; 6:843-50.

50. Beck-Sagué C, Dooley SW, Hutton MD, et al. Hospital outbreak of multidrug-resistant Mycobacterium tuberculosis infections. Factors in transmission to staff and HIV-infected patients. Jama **1992**; 268:1280-6.

51. Rodwell TC, Kapasi AJ, Barnes RFW, Moser KS. Factors associated with genotype clustering of Mycobacterium tuberculosis isolates in an ethnically diverse region of southern California, United States. Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases **2012**; 12:1917-25.

52. Scott C, Cavanaugh JS, Silk BJ, et al. Comparison of Sputum-Culture Conversion for Mycobacterium bovis and M. tuberculosis. Emerging infectious diseases **2017**; 23:456-62.

53. Gralton J, Tovey E, McLaws M-L, Rawlinson WD. The role of particle size in aerosolised pathogen transmission: a review. The Journal of infection **2011**; 62:1-13.

54. Morawska L, Johnson GR, Ristovski ZD, et al. Size distribution and sites of origin of droplets expelled from the human respiratory tract during expiratory activities. Journal of Aerosol Science **2009**; 40:256-69.

55. Turner RD, Bothamley GH. Cough and the transmission of tuberculosis. The Journal of infectious diseases **2015**; 211:1367-72.

56. Kwon S-B, Park J, Jang J, et al. Study on the initial velocity distribution of exhaled air from coughing and speaking. Chemosphere **2012**; 87:1260-4.

57. Turner RD HR, Birring SS, Bothamley Graham H. Daily Cough Frequency in Tuberculosis Is Associated with Rates of Household Infection. American Thoracic Society Conference, **2016**.

58. Behr MA, Warren SA, Salamon H, et al. Transmission of Mycobacterium tuberculosis from patients smear-negative for acid-fast bacilli. Lancet (London, England) **1999**; 353:444-9.

59. Daley CL, Small PM, Schecter GF, et al. An Outbreak of Tuberculosis with Accelerated Progression among Persons Infected with the Human Immunodeficiency Virus. New England Journal of Medicine **1992**; 326:231-5.

60. Gray J, Cohn D. Tuberculosis and HIV Coinfection. Seminars in Respiratory and Critical Care Medicine **2013**; 34:032-43.

61. Sonnenberg P, Glynn JR, Fielding K, Murray J, Godfrey-Faussett P, Shearer S. How soon after infection with HIV does the risk of tuberculosis start to increase? A retrospective cohort study in South African gold miners. The Journal of infectious diseases **2005**; 191:150-8.

62. Kwan CK, Ernst JD. HIV and tuberculosis: a deadly human syndemic. Clinical microbiology reviews **2011**; 24:351-76.

63. Schutz C, Meintjes G, Almajid F, Wilkinson RJ, Pozniak A. Clinical management of tuberculosis and HIV-1 co-infection. The European respiratory journal **2010**; 36:1460-81.

64. Sterling TR, Pham PA, Chaisson RE. HIV infection-related tuberculosis: clinical manifestations and treatment. Clinical infectious diseases : an official publication of the Infectious Diseases Society of America **2010**; 50 Suppl 3:S223-30.

65. Issarow CM, Mulder N, Wood R. Modelling the risk of airborne infectious disease using exhaled air. Journal of theoretical biology **2015**; 372:100-6.

66. Toit K, Altraja A, Acosta CD, et al. A four-year nationwide molecular epidemiological study in Estonia: risk factors for tuberculosis transmission. Public health action **2014**; 4:S34-40.

67. Guerra-Assunção JA, Crampin AC, Houben RMGJ, et al. Large-scale whole genome sequencing of M. tuberculosis provides insights into transmission in a high prevalence area. eLife **2015**; 4.

68. Munang ML, Browne C, Evans JT, et al. Programmatic utility of tuberculosis cluster investigation using a social network approach in Birmingham, United Kingdom. The international journal of tuberculosis and lung disease : the official journal of the International Union against Tuberculosis and Lung Disease **2016**; 20:1300-5.

69. Middelkoop K, Mathema B, Myer L, et al. Transmission of tuberculosis in a South African community with a high prevalence of HIV infection. The Journal of infectious diseases **2015**; 211:53-61.

70. Eldholm V, Rieux A, Monteserin J, et al. Impact of HIV co-infection on the evolution and transmission of multidrug-resistant tuberculosis. Trends in Ecology & Evolution **2016**; 5:306-13.

71. Andrews JR, Morrow C, Walensky RP, Wood R. Integrating social contact and environmental data in evaluating tuberculosis transmission in a South African township. The Journal of infectious diseases **2014**; 210:597-603.

72. Dodd PJ, Looker C, Plumb ID, et al. Age- and Sex-Specific Social Contact Patterns and Incidence of *Mycobacterium tuberculosis* Infection. American Journal of Epidemiology **2015**; 183:kwv160.

73. Johnstone-Robertson SP, Mark D, Morrow C, et al. Social mixing patterns within a South African township community: implications for respiratory disease transmission and control. American journal of epidemiology **2011**; 174:1246-55.

74. Middelkoop K, Bekker L-G, Morrow C, Lee N, Wood R. Decreasing household contribution to TB transmission with age: a retrospective geographic analysis of young people in a South African township. BMC infectious diseases **2014**; 14:221.

75. Horton KC, MacPherson P, Houben RMGJ, White RG, Corbett EL. Sex Differences in Tuberculosis Burden and Notifications in Low- and Middle-Income Countries: A Systematic Review and Meta-analysis. PLoS medicine **2016**; 13:e1002119.

76. Cruz AT, Starke JR. A current review of infection control for childhood tuberculosis. Tuberculosis **2011**; 91:S11-S5.

77. Chan ED, Kinney WH, Honda JR, et al. Tobacco exposure and susceptibility to tuberculosis: Is there a smoking gun? Tuberculosis **2014**; 94:544-50.

78. den Boon S, van Lill SWP, Borgdorff MW, et al. Association between smoking and tuberculosis infection: a population survey in a high tuberculosis incidence area. Thorax **2005**; 60:555-7.

79. Lavigne M, Rocher I, Steensma C, Brassard P. The impact of smoking on adherence to treatment for latent tuberculosis infection. BMC public health **2006**; 6:66.

80. Nicol MP, Wilkinson RJ. The clinical consequences of strain diversity in Mycobacterium tuberculosis. Transactions of the Royal Society of Tropical Medicine and Hygiene **2008**; 102:955-65.

81. Coscolla M, Gagneux S. Does M. tuberculosis genomic diversity explain disease diversity? Drug discovery today Disease mechanisms **2010**; 7:e43-e59.

82. López B, Aguilar D, Orozco H, et al. A marked difference in pathogenesis and immune response induced by different Mycobacterium tuberculosis genotypes. Clinical and experimental immunology **2003**; 133:30-7.

83. Dormans J, Burger M, Aguilar D, et al. Correlation of virulence, lung pathology, bacterial load and delayed type hypersensitivity responses after infection with different Mycobacterium tuberculosis genotypes in a BALB/c mouse model. Clinical and experimental immunology **2004**; 137:460-8.

84. Manca C, Tsenova L, Barry CEr, et al. Mycobacterium tuberculosis CDC1551 induces a more vigorous host response in vivo and in vitro, but is not more virulent than other clinical isolates. Journal of immunology (Baltimore, Md : 1950) **1999**; 162:6740-6.

85. Manabe YC, Dannenberg AMJ, Tyagi SK, et al. Different strains of Mycobacterium tuberculosis cause various spectrums of disease in the rabbit model of tuberculosis. Infection and immunity **2003**; 71:6004-11.

86. Valway SE, Sanchez MPC, Shinnick TF, et al. An Outbreak Involving Extensive Transmission of a Virulent Strain of Mycobacterium tuberculosis. New England Journal of Medicine **1998**; 338:633-9.

87. Wood R, Morrow C, Barry CE, III, et al. Real-Time Investigation of Tuberculosis Transmission: Developing the Respiratory Aerosol Sampling Chamber (RASC). PloS one **2016**; 11:e0146658.

88. Anderson J, Jarlsberg LG, Grindsdale J, et al. Sublineages of lineage 4 (Euro-American) Mycobacterium tuberculosis differ in genotypic clustering. The international journal of tuberculosis and lung disease : the official journal of the International Union against Tuberculosis and Lung Disease **2013**; 17:885-91.

89. Click ES, Winston CA, Oeltmann JE, Moonan PK, Mac Kenzie WR. Association between Mycobacterium tuberculosis lineage and time to sputum culture conversion. The international journal of tuberculosis and lung disease : the official journal of the International Union against Tuberculosis and Lung Disease **2013**; 17:878-84.

90. Nahid P, Bliven EE, Kim EY, et al. Influence of M. tuberculosis lineage variability within a clinical trial for pulmonary tuberculosis. PloS one **2010**; 5:e10753.

91. Parwati I, Alisjahbana B, Apriani L, et al. *Mycobacterium tuberculosis* Beijing Genotype Is an Independent Risk Factor for Tuberculosis Treatment Failure in Indonesia. The Journal of infectious diseases **2010**; 201:553-7.

92. Parwati I, van Crevel R, van Soolingen D. Possible underlying mechanisms for successful emergence of the Mycobacterium tuberculosis Beijing genotype strains. The Lancet Infectious diseases **2010**; 10:103-11.

93. Luciani F, Sisson SA, Jiang H, Francis AR, Tanaka MM. The epidemiological fitness cost of drug resistance in *Mycobacterium tuberculosis*. Proceedings of the National Academy of Sciences **2009**; 106:14711-5.

94. Gagneux S, Long CD, Small PM, Van T, Schoolnik GK, Bohannan BJ. The competitive cost of antibiotic resistance in Mycobacterium tuberculosis. Science (New York, NY) **2006**; 312:1944-6.

95. Grandjean L, Gilman RH, Martin L, et al. Transmission of Multidrug-Resistant and Drug-Susceptible Tuberculosis within Households: A Prospective Cohort Study. PLoS medicine **2015**; 12:e1001843; discussion e.

96. Cohen T, Murray M. Modeling epidemics of multidrug-resistant M. tuberculosis of heterogeneous fitness. Nature medicine **2004**; 10:1117-21.

97. MIDDLEBROOK G. Isoniazid-resistance and catalase activity of tubercle bacilli; a preliminary report. American review of tuberculosis **1954**; 69:471-2.

98. Pym AS, Saint-Joanis B, Cole ST. Effect of katG mutations on the virulence of Mycobacterium tuberculosis and the implication for transmission in humans. Infection and immunity **2002**; 70:4955-60.

99. Gagneux S, Burgos MV, DeRiemer K, et al. Impact of bacterial genetics on the transmission of isoniazid-resistant Mycobacterium tuberculosis. PLoS pathogens **2006**; 2:e61.

100. Burgos M, DeRiemer K, Small PM, Hopewell PC, Daley CL. Effect of drug resistance on the generation of secondary cases of tuberculosis. The Journal of infectious diseases **2003**; 188:1878-84.

101. Mathema B, Kurepina N, Fallows D, Kreiswirth BN. Lessons from molecular epidemiology and comparative genomics. Seminars in respiratory and critical care medicine **2008**; 29:467-80.

102. Anderson LF, Tamne S, Brown T, et al. Transmission of multidrug-resistant tuberculosis in the UK: a cross-sectional molecular and epidemiological study of clustering and contact tracing. The Lancet Infectious Diseases **2017**; 14:406-15.

103. Gandhi NR, Weissman D, Moodley P, et al. Nosocomial transmission of extensively drug-resistant tuberculosis in a rural hospital in South Africa. The Journal of infectious diseases **2013**; 207:9-17.

104. Yang C, Luo T, Shen X, et al. Transmission of multidrug-resistant Mycobacterium tuberculosis in Shanghai, China: a retrospective observational study using whole-genome sequencing and epidemiological investigation. The Lancet Infectious Diseases **2017**; 17:275-84.

105. Jean Dubos RJD. The White Plague: Tuberculosis,Man, and Society. Boston, Little, Brown, **1952**.

106. Ali M. Treating tuberculosis as a social disease. The Lancet **2014**; 383:2195.

107. Dye C, Lönnroth K, Jaramillo E, Williams BG, Raviglione M. Trends in tuberculosis incidence and their determinants in 134 countries. Bulletin of the World Health Organization **2009**; 87:683-91.

108. Hidden Cities: Unmasking and Overcoming Health Inequities in Urban Settings. Geneva, Switzerland: The World Health Organization, **2010**.

109. Taylor JG, Yates TA, Mthethwa M, Tanser F, Abubakar I, Altamirano H. Measuring ventilation and modelling M. tuberculosis transmission in indoor congregate settings, rural KwaZulu-Natal. The international journal of tuberculosis and lung disease : the official journal of the International Union against Tuberculosis and Lung Disease **2016**; 20:1155-61.

110. Chamie G, Wandera B, Luetkemeyer A, et al. Household ventilation and tuberculosis transmission in Kampala, Uganda. The international journal of tuberculosis and lung disease : the official journal of the International Union against Tuberculosis and Lung Disease **2013**; 17:764-70.

111. Escombe AR, Oeser CC, Gilman RH, et al. Natural ventilation for the prevention of airborne contagion. PLoS medicine **2007**; 4.

112. Chamie G, Wandera B, Luetkemeyer A, et al. Household ventilation and tuberculosis transmission in Kampala, Uganda. The International Journal of Tuberculosis and Lung Disease **2013**; 17:764-70.

113. Odone A, Crampin AC, Mwinuka V, et al. Association between socioeconomic position and tuberculosis in a large population-based study in rural Malawi. PloS one **2013**; 8:e77740.

114. Escombe AR, Moore DAJ, Gilman RH, et al. Upper-Room Ultraviolet Light and Negative Air Ionization to Prevent Tuberculosis Transmission. PLoS medicine **2009**; 6:e1000043.

115. Vella V, Racalbuto V, Guerra R, et al. Household contact investigation of multidrug-resistant and extensively drug-resistant tuberculosis in a high HIV prevalence setting. The international journal of tuberculosis and lung disease : the official journal of the International Union against Tuberculosis and Lung Disease **2011**; 15:1170-5, i.

116. Shah NS, Yuen CM, Heo M, Tolman AW, Becerra MC. Yield of contact investigations in households of patients with drug-resistant tuberculosis: systematic review and meta-analysis. Clinical infectious diseases : an official publication of the Infectious Diseases Society of America **2014**; 58:381-91.

117. Marquez L, Feske ML, Teeter LD, Musser JM, Graviss EA. Pediatric Tuberculosis. The Pediatric Infectious Disease Journal **2012**; 31:1144-7.

118. Lambert LA, Armstrong LR, Lobato MN, Ho C, France AM, Haddad MB. Tuberculosis in Jails and Prisons: United States, 2002-2013. Am J Public Health **2016**; 106:2231-7.

119. Mindra G, Wortham JM, Haddad MB, Powell KM. Tuberculosis Outbreaks in the United States, 2009-2015. Public health reports (Washington, DC : 1974) **2017**; 132:157-63.

120. Basu S, Stuckler D, McKee M. Addressing institutional amplifiers in the dynamics and control of tuberculosis epidemics. The American journal of tropical medicine and hygiene **2011**; 84:30-7.

121. Shilova MV, Dye C. The resurgence of tuberculosis in Russia. Philosophical Transactions of the Royal Society B: Biological Sciences **2001**; 356:1069-75.

122. Godfrey-Faussett P, Sonnenberg P, Shearer S, et al. Tuberculosis control and molecular epidemiology in a South African gold-mining community. The Lancet **2000**; 356:1066-71.

123. Lobacheva T, Sazhin V, Vdovichenko E, Giesecke J. Pulmonary tuberculosis in two remand prisons (SIZOs) in St Petersburg, Russia. Euro surveillance : bulletin Europeen sur les maladies transmissibles = European communicable disease bulletin **2005**; 10:93-6.

124. Warren JL, Grandjean L, Moore DAJ, et al. Investigating spillover of multidrug-resistant tuberculosis from a prison: a spatial and molecular epidemiological analysis. BMC medicine **2018**; 16:122.

125. Verver S, Warren RM, Munch Z, et al. Proportion of tuberculosis transmission that takes place in households in a high-incidence area. Lancet (London, England) **2004**; 363:212-4.

126. Crampin AC, Glynn JR, Traore H, et al. Tuberculosis transmission attributable to close contacts and HIV status, Malawi. Emerging infectious diseases **2006**; 12:729-35.

127. Glynn JR, Guerra-Assuncao JA, Houben RMGJ, et al. Whole Genome Sequencing Shows a Low Proportion of Tuberculosis Disease Is Attributable to Known Close Contacts in Rural Malawi. PloS one **2015**; 10:e0132840.

128. Murray M, Alland D. Methodological problems in the molecular epidemiology of tuberculosis. American journal of epidemiology **2002**; 155:565-71.

129. Andrews JR, Morrow C, Wood R. Modeling the Role of Public Transportation in Sustaining Tuberculosis Transmission in South Africa. American Journal of Epidemiology **2013**; 177:556-61.

130. McCreesh N, White RG. An explanation for the low proportion of tuberculosis that results from transmission between household and known social contacts. Scientific Reports **2018**; 8:5382.

131. Taylor Z. Guidelines for the Investigation of Contacts of Persons with Infectious Tuberculosis. In: Morbidity and Mortality Weekly Report.1-37.

132. Fox GJ, Barry SE, Britton WJ, Marks GB. Contact investigation for tuberculosis: a systematic review and meta-analysis. The European respiratory journal **2013**; 41:140-56.

133. Read JM, Edmunds WJ, Riley S, Lessler J, Cummings DAT. Close encounters of the infectious kind: methods to measure social mixing behaviour. Epidemiology and infection **2012**; 140:2117-30.

134. Smieszek TBEUSRSRW. Collecting close-contact social mixing data with contact diaries: reporting errors and biases. Epidemiology and infection **2012**; 140:744-52.

135. Mathema B, Kurepina NE, Bifani PJ, Kreiswirth BN. Molecular Epidemiology of Tuberculosis: Current Insights. Clinical Microbiology Reviews **2006**; 19:658-85.

136. Murray M. Sampling bias in the molecular epidemiology of tuberculosis. Emerging infectious diseases **2002**; 8:363-9.

137. Gardy JL, Johnston JC, Ho Sui SJ, et al. Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. The New England journal of medicine **2011**; 364:730-9.

138. Kato-Maeda M, Ho C, Passarelli B, et al. Use of whole genome sequencing to determine the microevolution of Mycobacterium tuberculosis during an outbreak. PloS one **2013**; 8:e58235.

139. Stucki D, Ballif M, Bodmer T, et al. Tracking a tuberculosis outbreak over 21 years: strain-specific single-nucleotide polymorphism typing combined with targeted whole-genome sequencing. The Journal of infectious diseases **2015**; 211:1306-16.

140. Casali N, Broda A, Harris SR, et al. Whole Genome Sequence Analysis of a Large Isoniazid-Resistant Tuberculosis Outbreak in London: A Retrospective Observational Study. PLoS medicine **2016**; 13:e1002137.

141. Nikolayevskyy V, Kranzer K, Niemann S, Drobniewski F. Whole genome sequencing of Mycobacterium tuberculosis for detection of recent transmission and tracing outbreaks: A systematic review. Tuberculosis (Edinburgh, Scotland) **2016**; 98:77-85.

142. Liu X, Gutacker MM, Musser JM, Fu Y-X. Evidence for recombination in Mycobacterium tuberculosis. Journal of bacteriology **2006**; 188:8169-77.

143. Ford CB, Shah RR, Maeda MK, et al. Mycobacterium tuberculosis mutation rate estimates from different lineages predict substantial differences in the emergence of drug-resistant tuberculosis. Nature genetics **2013**; 45:784-90.

144. Achtman M. Evolution, population structure, and phylogeography of genetically monomorphic bacterial pathogens. Annu Rev Microbiol **2008**; 62:53-70.

145. Achtman M. Insights from genomic comparisons of genetically monomorphic bacterial pathogens. Philosophical Transactions of the Royal Society of London B: Biological Sciences **2012**; 367:860-7.

146. Comas I, Homolka S, Niemann S, Gagneux S. Genotyping of genetically monomorphic bacteria: DNA sequencing in Mycobacterium tuberculosis highlights the limitations of current methodologies. PloS one **2009**; 4:e7815.

147. Brites D, Gagneux S. Co-evolution of Mycobacterium tuberculosis and Homo sapiens. Immunological reviews **2015**; 264:6-24.

148. Brown TS, Narechania A, Walker JR, et al. Genomic epidemiology of Lineage 4 Mycobacterium tuberculosis subpopulations in New York city and New Jersey, 1999–2009. BMC Genomics **2016**; 17:947.

149. Ioerger TR, Feng Y, Chen X, et al. The non-clonality of drug resistance in Beijing-genotype isolates of Mycobacterium tuberculosis from the Western Cape of South Africa. BMC genomics **2010**; 11:670.

150. Cohen KA, Abeel T, Manson McGuire A, et al. Evolution of Extensively Drug-Resistant Tuberculosis over Four Decades: Whole Genome Sequencing and Dating Analysis of Mycobacterium tuberculosis Isolates from KwaZulu-Natal. PLoS medicine **2015**; 12:e1001880.

151. Hatherell H-A, Colijn C, Stagg HR, Jackson C, Winter JR, Abubakar I. Interpreting whole genome sequencing for investigating tuberculosis transmission: a systematic review. BMC medicine **2016**; 14:21.

152. Ford CB, Lin PL, Chase MR, et al. Use of whole genome sequencing to estimate the mutation rate of Mycobacterium tuberculosis during latent infection. Nature genetics **2011**; 43:482-6.

153. Folkvardsen DB, Norman A, Andersen AB, Michael Rasmussen E, Jelsbak L, Lillebaek T. Genomic Epidemiology of a Major Mycobacterium tuberculosis Outbreak: Retrospective Cohort Study in a Low-Incidence Setting Using Sparse Time-Series Sampling. The Journal of infectious diseases **2017**; 216:366-74.

154. Comas I, Coscolla M, Luo T, et al. Out-of-Africa migration and Neolithic coexpansion of Mycobacterium tuberculosis with modern humans. Nature genetics **2013**; 45:1176-82.

155. Wlodarska M, Johnston JC, Gardy JL, Tang P. A Microbiological Revolution Meets an Ancient Disease: Improving the Management of Tuberculosis with Genomics. Clinical Microbiology Reviews **2015**; 28:523-39.

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

157. Branton D, Deamer DW, Marziali A, et al. The potential and challenges of nanopore sequencing. Nat Biotech **2008**; 26:1146-53.

158. Votintseva AA, Bradley P, Pankhurst L, et al. Same-Day Diagnostic and Surveillance Data for Tuberculosis via Whole-Genome Sequencing of Direct Respiratory Samples. Journal of clinical microbiology **2017**; 55:1285-98.

159. U. Köser C, M. Bryant J, Becq J, et al. Whole-Genome Sequencing for Rapid Susceptibility Testing of M. tuberculosis. The New England journal of medicine **2013**; 369:10.1056/NEJMc1215305.

160. Campbell PJ, Morlock GP, Sikes RD, et al. Molecular Detection of Mutations Associated with First- and Second-Line Drug Resistance Compared with Conventional Drug Susceptibility Testing of Mycobacterium tuberculosis. Antimicrobial Agents and Chemotherapy **2011**; 55:2032-41.

161. Editors TPM. It's the Network, Stupid: Why Everything in Medicine Is Connected. PLoS medicine **2008**; 5:e71.

162. Koopman J. Modeling infection transmission. Annual review of public health **2004**; 25:303-26.

163. Danon L, Ford AP, House T, et al. Networks and the epidemiology of infectious disease. Interdisciplinary perspectives on infectious diseases **2011**; 2011:284909.

164. Keeling MJ, Eames KTD. Networks and epidemic models. Journal of the Royal Society, Interface **2005**; 2:295-307.

165. Ray B, Ghedin E, Chunara R. Network inference from multimodal data: A review of approaches from infectious disease transmission. Journal of biomedical informatics **2016**; 64:44-54.

166. Laumann EO, Youm Y. Racial/ethnic group differences in the prevalence of sexually transmitted diseases in the United States: a network explanation. Sexually transmitted diseases **1999**; 26:250-61.

167. Rodriguez-Hart C, Liu H, Nowak RG, et al. Serosorting and Sexual Risk for HIV Infection at the Ego-Alter Dyadic Level: An Egocentric Sexual Network Study Among MSM in Nigeria. AIDS and behavior **2016**; 20:2762-71.

168. Klovdahl AS, Graviss EA, Yaganehdoost A, et al. Networks and tuberculosis: an undetected community outbreak involving public places. Social science & medicine (1982) **2001**; 52:681-94.

169. Broadhead RS, Heckathorn DD, Weakliem DL, et al. Harnessing peer networks as an instrument for AIDS prevention: results from a peer-driven intervention. Public health reports (Washington, DC : 1974) **1998**:42-57.

170. Rothenberg R, Narramore J. The relevance of social network concepts to sexually transmitted disease control. Sexually transmitted diseases **1996**; 23:24-9.

171. Latkin CA, Sherman S, Knowlton A. HIV prevention among drug users: outcome of a network-oriented peer outreach intervention. Health psychology : official journal of the Division of Health Psychology, American Psychological Association **2003**; 22:332-9.

172. Luke DA, Harris JK. Network Analysis in Public Health: History, Methods, and Applications. Annual Review of Public Health **2007**; 28:69-93.

173. Potterat JJ, Muth SQ, Rothenberg RB, et al. Sexual network structure as an indicator of epidemic phase. Sexually transmitted infections **2002**; 78 Suppl 1:i152-8.

174. Potterat JJ, Phillips-Plummer L, Muth SQ, et al. Risk network structure in the early epidemic phase of HIV transmission in Colorado Springs. Sexually transmitted infections **2002**; 78 Suppl 1:i159-63.

175. Stoner BP, Whittington WL, Hughes JP, Aral SO, Holmes KK. Comparative epidemiology of heterosexual gonococcal and chlamydial networks: implications for transmission patterns. Sexually transmitted diseases **2000**; 27:215-23.

176. Wylie JL, Jolly A. Patterns of chlamydia and gonorrhea infection in sexual networks in Manitoba, Canada. Sexually transmitted diseases **2001**; 28:14-24.

177. Meyers LA, Pourbohloul B, Newman MEJ, Skowronski DM, Brunham RC. Network theory and SARS: predicting outbreak diversity. Journal of theoretical biology **2005**; 232:71-81.

178. Jenness SM, Goodreau SM, Morris M, Cassels S. Effectiveness of combination packages for HIV-1 prevention in sub-Saharan Africa depends on partnership network structure: a mathematical modelling study. Sexually Transmitted Infections **2016**; 92:619-24.

179. Riley S, Fraser C, Donnelly CA, et al. Transmission dynamics of the etiological agent of SARS in Hong Kong: impact of public health interventions. Science (New York, NY) **2003**; 300:1961-6.

180. Baraff AJ, McCormick TH, Raftery AE. Estimating uncertainty in respondent-driven sampling using a tree bootstrap method. Proceedings of the National Academy of Sciences **2016**; 113:14668-73.

181. Krivitsky PN, Morris M. Inference for social network models from egocentrically sampled data, with application to understanding persistent racial disparities in HIV prevalence in the US. Ann Appl Stat **2017**; 11:427-55.

182. Goel S, Salganik MJ. Assessing respondent-driven sampling. Proceedings of the National Academy of Sciences **2010**; 107:6743-7.

183. Gile KJ, Handcock MS. 7. Respondent-Driven Sampling: An Assessment of Current Methodology. Sociological Methodology **2010**; 40:285-327.

184. Logan JJ, Jolly AM, Blanford JI, et al. The Sociospatial Network: Risk and the Role of Place in the Transmission of Infectious Diseases. PloS one **2016**; 11:e0146915.

185. Read J M; Edmunds WJR, S; Lessler, J; Cummings D A T. Close encounters of the infectious kind: methods to measure social mixing behaviour. Epidemiology and infection **2012**; 140:2117-30.

186. Morris M, Handcock MS, Hunter DR. Specification of Exponential-Family Random Graph Models: Terms and Computational Aspects. Journal of statistical software **2008**; 24:1548-7660.

187. Holland PW, Leinhardt S. An Exponential Family of Probability Distributions for Directed Graphs. Journal of the American Statistical Association **1981**; 76:33-50.

188. Frank O, Strauss D. Markov graphs. Journal of the american Statistical association **1986**; 81:832-42.

189. Strauss D, Ikeda M. Pseudolikelihood Estimation for Social Networks. Journal of the American Statistical Association **1990**; 85:204.

190. Snijders TAB. Markov Chain Monte Carlo Estimation of Exponential Random Graph Models. **2002**.

191. Robins G, Pattison P, Kalish Y, Lusher D. An introduction to exponential random graph (p*) models for social networks. Social Networks **2007**; 29:173-91.

192. Laumann EO, Marsden PV, Prensky D. The boundary specification problem in network analysis. Research methods in social network analysis **1989**; 61:87.

193. Bliss CA, Danforth CM, Dodds PS, Vespignani A, Slooten E. Estimation of Global Network Statistics from Incomplete Data. PloS one **2014**; 9:e108471.

194. Costenbader E, Valente TW. The stability of centrality measures when networks are sampled. Social networks **2003**; 25:283-307.

195. Wang DJ, Shi X, McFarland DA, Leskovec J. Measurement error in network data: A re-classification. Social Networks **2012**; 34:396-409.

196. Borgatti SP, Borgatti SP, Carley KM, Krackhardt D. On the robustness of centrality measures under conditions of imperfect data. SOCIAL NETWORKS **2004**:2006.

197. Kossinets G. Effects of missing data in social networks *. **2003**.

198. Smith JA, Moody J, Morgan JH. Network sampling coverage II: The effect of non-random missing data on network measurement. Social Networks **2017**; 48:78-99.

199. Walker TM, Lalor MK, Broda A, et al. Assessment of Mycobacterium tuberculosis transmission in Oxfordshire, UK, 2007-12, with whole pathogen genome sequences: an observational study. The Lancet Respiratory medicine **2014**; 2:285-92.

200. Walker TM, Ip CL, Harrell RH, et al. Whole-genome sequencing to delineate Mycobacterium tuberculosis outbreaks: a retrospective observational study. The Lancet Infectious Diseases **2013**; 13:137-46.

201. Hu Y, Mathema B, Jiang W, Kreiswirth B, Wang W, Xu B. Transmission pattern of drug-resistant tuberculosis and its implication for tuberculosis control in eastern rural China. PloS one **2011**; 6:e19548.

202. Dheda K, Limberis JD, Pietersen E, et al. Outcomes, infectiousness, and transmission dynamics of patients with extensively drug-resistant tuberculosis and home-discharged patients with programmatically incurable tuberculosis: a prospective cohort study. The Lancet Respiratory Medicine **2017**; 5:269-81.

203. Cook VJ, Sun SJ, Tapia J, et al. Transmission Network Analysis in Tuberculosis Contact Investigations. The Journal of infectious diseases **2007**; 196:1517-27.

204. Andre M, Ijaz K, Tillinghast JD, et al. Transmission Network Analysis to Complement Routine Tuberculosis Contact Investigations. American Journal of Public Health **2007**; 97:470-7.

205. Leigh Brown AJ, Lycett SJ, Weinert L, et al. Transmission network parameters estimated from HIV sequences for a nationwide epidemic. The Journal of infectious diseases **2011**; 204:1463-9.

206. Kouyos Roger D, von Wyl V, Yerly S, et al. Molecular Epidemiology Reveals Long-Term Changes in HIV Type 1 Subtype B Transmission in Switzerland. The Journal of infectious diseases **2010**; 201:1488-97.

207. Chan PA, Hogan JW, Huang A, et al. Phylogenetic Investigation of a Statewide HIV-1 Epidemic Reveals Ongoing and Active Transmission Networks Among Men Who Have Sex With Men. Journal of acquired immune deficiency syndromes (1999) **2015**; 70:428-35.

208. Lewis F, Hughes GJ, Rambaut A, Pozniak A, Leigh Brown AJ. Episodic sexual transmission of HIV revealed by molecular phylodynamics. PLoS medicine **2008**; 5:e50.

209. Brenner BG, Roger M, Stephens D, et al. Transmission Clustering Drives the Onward Spread of the HIV Epidemic Among Men Who Have Sex With Men in Quebec. Journal of Infectious Diseases **2011**; 204:1115-9.

210. Wertheim JO, Oster AM, Johnson JA, et al. Transmission fitness of drug-resistant HIV revealed in a surveillance system transmission network. Virus evolution **2017**; 3:vex008.

211. Zarrabi N, Prosperi M, Belleman RG, Colafigli M, De Luca A, Sloot PMA. Combining epidemiological and genetic networks signifies the importance of early treatment in HIV-1 transmission. PloS one **2012**; 7:e46156.

212. Brenner BG, Wainberg MA. Future of Phylogeny in HIV Prevention. JAIDS Journal of Acquired Immune Deficiency Syndromes **2013**; 63:S248-S54.

213. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics **2009**; 25:1754-60.

214. Li H, Handsaker B, Wysoker A, et al. The sequence alignment/map format and SAMtools. Bioinformatics **2009**; 25:2078-9.

215. Shisana O RT, Simbayi L, Zuma KK, Jooste S, Zungu NP, Labadarios D, Onoya D, Wabiri N. South African National HIV Prevalence, Incidence, and Behaviour Survey, 2012.

216. Nel An, Mabude Z, Smit J, et al. HIV Incidence Remains High in KwaZulu-Natal, South Africa: Evidence from Three Districts. PloS one **2012**; 7:e35278.

217. O'Donnell MR, Padayatchi N, Kvasnovsky C, Werner L, Master I, Horsburgh CR, Jr. Treatment outcomes for extensively drug-resistant tuberculosis and HIV co-infection. Emerg Infect Dis **2013**; 19:416-24.

218. Martinez L, Shen Y, Mupere E, Kizza A, Hill PC, Whalen CC. Transmission of Mycobacterium Tuberculosis in Households and the Community: A Systematic Review and Meta-Analysis. Am J Epidemiol **2017**; 185:1327-39.

219. Bryant JM, Harris SR, Parkhill J, et al. Whole-genome sequencing to establish relapse or re-infection with Mycobacterium tuberculosis: a retrospective observational study. The Lancet Respiratory medicine **2013**; 1:786-92.

220. Pérez-Lago L, Comas I, Navarro Y, et al. Whole genome sequencing analysis of intrapatient microevolution in Mycobacterium tuberculosis: potential impact on the inference of tuberculosis transmission. The Journal of infectious diseases **2014**; 209:98-108.

221. Lessler J, Salje H, Grabowski MK, Cummings DAT. Measuring Spatial Dependence for Infectious Disease Epidemiology. PloS one **2016**; 11:e0155249.

222. Riley S. Large-Scale Spatial-Transmission Models of Infectious Disease. Science (New York, NY) **2007**; 316.

223. Ndjeka N. Strategic Overview of MDR-TB Care in South Africa.

224. Lim JR, Gandhi NR, Mthiyane T, et al. Incidence and Geographic Distribution of Extensively Drug-Resistant Tuberculosis in KwaZulu-Natal Province, South Africa. PloS one **2015**; 10:e0132076.

225. Census 2011: Census in Brief. Pretoria, South Africa: Statistics South Africa, **2011**.

226. Camlin CS, Hosegood V, Newell M-L, McGrath N, Bärnighausen T, Snow RC. Gender, Migration and HIV in Rural KwaZulu-Natal, South Africa. PloS one **2010**; 5:e11539.

227. Eldholm V, Monteserin J, Rieux A, et al. Four decades of transmission of a multidrug-resistant Mycobacterium tuberculosis outbreak strain. **2015**; 6:7119.

228. Pebesma EJ BR. Classes and methods for spatial data in R. R News **2005**; 5.

229. Hijmans R. geosphere: Spherical Trigonometry. R package version 1.5-5 ed, **2016**.

230. Lurie M, Harrison A, Wilkinson D, Karim SA. Circular migration and sexual networking in rural KwaZulu/Natal: implications for the spread of HIV and other sexually transmitted diseases. Health Transition Review **1997**; 7:17-27.

231. Huang CC, Tchetgen ET, Becerra MC, et al. The effect of HIV-related immunosuppression on the risk of tuberculosis transmission to household contacts. Clinical infectious diseases : an official publication of the Infectious Diseases Society of America **2014**; 58:765-74.

232. Espinal MA, Peréz EN, Baéz J, et al. Infectiousness of Mycobacterium tuberculosis in HIV-1-infected patients with tuberculosis: a prospective study. Lancet (London, England) **2000**; 355:275-80.

233. Sullivan BJ, Esmaili BE, Cunningham CK. Barriers to initiating tuberculosis treatment in sub-Saharan Africa: a systematic review focused on children and youth. Global Health Action **2017**; 10:1290317.

234. South African Demographic and Health Survey 2003. Pretoria: Department of Health, Medical Research Council, **2007**.

235. Ribeiro FK, Pan W, Bertolde A, et al. Genotypic and Spatial Analysis of Mycobacterium tuberculosis Transmission in a High-Incidence Urban Setting. Clinical infectious diseases : an official publication of the Infectious Diseases Society of America **2015**; 61:758-66.

236. Zelner JL, Murray MB, Becerra MC, et al. Identifying Hotspots of Multidrug-Resistant Tuberculosis Transmission Using Spatial and Molecular Genetic Data. The Journal of infectious diseases **2016**; 213:287-94.

237. Bryant JM, Schürch AC, van Deutekom H, et al. Inferring patient to patient transmission of Mycobacterium tuberculosis from whole genome sequencing data. BMC infectious diseases **2013**; 13:110.

238. Gandhi NR, Brust JCM, Moodley P, et al. Minimal diversity of drug-resistant Mycobacterium tuberculosis strains, South Africa. Emerging infectious diseases **2014**; 20:426-33.

239. Chihota VN, Müller B, Mlambo CK, et al. Population Structure of Multi- and Extensively Drug-Resistant Mycobacterium tuberculosis Strains in South Africa. Journal of Clinical Microbiology **2012**; 50:995-1002.

240. Alene KA, Viney K, McBryde ES, et al. Spatial patterns of multidrug resistant tuberculosis and relationships to socio-economic, demographic and household factors in northwest Ethiopia. PloS one **2017**; 12:e0171800.

241. Targeted tuberculin testing and treatment of latent tuberculosis infection. This official statement of the American Thoracic Society was adopted by the ATS Board of Directors, July 1999. This is a Joint Statement of the American Thoracic Society (ATS) and the Centers for Disease Control and Prevention (CDC). This statement was endorsed by the Council of the Infectious Diseases Society of America. (IDSA), September 1999, and the sections of this statement. American journal of respiratory and critical care medicine **2000**; 161:S221-47.

242. Kapwata T, Morris N, Gandhi N, et al. Spatial distribution of extensively drug-resistant tuberculosis (XDR-TB) patients in KwaZulu-Natal, South Africa. bioRxiv **2017**.

243. Cohen T, van Helden PD, Wilson D, et al. Mixed-Strain Mycobacterium tuberculosis Infections and the Implications for Tuberculosis Treatment and Control. Clinical Microbiology Reviews **2012**; 25:708-19.

244. Becerra MC, Appleton SC, Franke MF, et al. Tuberculosis burden in households of patients with multidrug-resistant and extensively drug-resistant tuberculosis: a retrospective cohort study. Lancet (London, England) **2011**; 377:147-52.

245. Devaux I, Kremer K, Heersma H, Van Soolingen D. Clusters of Multidrug-Resistant Mycobacterium tuberculosis Cases, Europe. Emerging Infectious Diseases **2009**; 15:1052-60.

246. Paull SH, Song S, McClure KM, Sackett LC, Kilpatrick AM, Johnson PTJ. From superspreaders to disease hotspots: linking transmission across hosts and space. Frontiers in ecology and the environment **2012**; 10:75-82.

247. Li LM, Grassly NC, Fraser C. Quantifying Transmission Heterogeneity Using Both Pathogen Phylogenies and Incidence Time Series. Molecular Biology and Evolution **2017**; 34:2982-95.

248. Dowdy DW, Golub JE, Chaisson RE, Saraceni V. Heterogeneity in tuberculosis transmission and the role of geographic hotspots in propagating epidemics. Proc Natl Acad Sci USA **2012**; 109.

249. Oster AM, Wertheim JO, Hernandez AL, Ocfemia MC, Saduvala N, Hall HI. Using Molecular HIV Surveillance Data to Understand Transmission Between Subpopulations in the United States. J Acquir Immune Defic Syndr **2015**; 70:444-51.

250. Morgan E, Oster AM, Townsell S, Peace D, Benbow N, Schneider JA. HIV-1 Infection and Transmission Networks of Younger People in Chicago, Illinois, 2005-2011. Public Health Reports **2017**; 132:48-55.

251. Oster AM, France A, Mermin J. Molecular epidemiology and the transformation of hiv prevention. Jama **2018**; 319:1657-8.

252. Blower SM, Chou T. Modeling the emergence of the 'hot zones': tuberculosis and the amplification dynamics of drug resistance. Nature medicine **2004**; 10:1111-6.

253. Nelson KN, Shah NS, Mathema B, et al. Spatial Patterns of Extensively Drug-Resistant Tuberculosis Transmission in KwaZulu-Natal, South Africa. The Journal of infectious diseases **2018**:jiy394-jiy.

254. Bartlett SR, Wertheim JO, Bull RA, et al. A molecular transmission network of recent hepatitis C infection in people with and without HIV: Implications for targeted treatment strategies. Journal of viral hepatitis **2017**; 24:404-11.

255. Lurie MN, Williams BG. Migration and health in Southern Africa: 100 years and still circulating. Health Psychology and Behavioral Medicine **2014**; 2:34-40.

256. Lurie MN, Williams BG, Zuma K, et al. The impact of migration on HIV-1 transmission in South Africa: a study of migrant and nonmigrant men and their partners. Sexually transmitted diseases **2003**; 30:149-56.

257. Stuckler D, Basu S, McKee M, Lurie M. Mining and risk of tuberculosis in sub-Saharan Africa. American journal of public health **2011**; 101:524-30.

258. Feske ML, Teeter LD, Musser JM, Graviss EA. Giving TB wheels: Public transportation as a risk factor for tuberculosis transmission. Tuberculosis (Edinburgh, Scotland) **2011**; 91 Suppl 1:S16-23.

259. Stimson J, Gardy JL, Mathema B, Crudu V, Cohen T, Colijn C. Beyond the SNP threshold: identifying outbreak clusters using inferred transmissions. bioRxiv **2018**.

260. Naidoo CC, Pillay M. Increased in vitro fitness of multi- and extensively drug-resistant F15/LAM4/KZN strains of Mycobacterium tuberculosis. Clinical microbiology and infection : the official publication of the European Society of Clinical Microbiology and Infectious Diseases **2014**; 20:O361-9.

261. The End TB Strategy. Geneva, Switzerland: World Health Organization, **2015**.

262. Dowdy DW, Azman AS, Kendall EA, Mathema B. Transforming the Fight Against Tuberculosis: Targeting Catalysts of Transmission. Clinical Infectious Diseases **2014**; 59:1123-9.

263. Didelot X, Fraser C, Gardy J, Colijn C. Genomic infectious disease epidemiology in partially sampled and ongoing outbreaks. Molecular Biology and Evolution **2017**; 34:msw075.

264. Dudas G, Carvalho LM, Bedford T, et al. Virus genomes reveal factors that spread and sustained the Ebola epidemic. Nature **2017**; 544:309-15.

265. Faria NR, Quick J, Claro IM, et al. Establishment and cryptic transmission of Zika virus in Brazil and the Americas. Nature **2017**; 546:406-10.

266. Metsky HC, Matranga CB, Wohl S, et al. Zika virus evolution and spread in the Americas. Nature **2017**; 546:411-5.

267. Worobey M. Epidemiology: Molecular mapping of Zika spread. Nature **2017**; advance on.

268. Zayas G, Chiang MC, Wong E, et al. Effectiveness of cough etiquette maneuvers in disrupting the chain of transmission of infectious respiratory diseases. BMC public health **2013**; 13:811.