# Integrating molecular typing into routine tuberculosis surveillance: An assessment of the strengths and limitations of current approaches

by

Anne Marie France

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Epidemiological Science)
in The University of Michigan
2008

Doctoral Committee:

Associate Professor Zhenhua Yang, Co-chair
Professor Betsy Foxman, Co-chair
Professor Denise Kirschner
Associate Professor Carl F. Marrs
Research Scientist Rick Riolo

For Maurice

## Acknowledgements

I am indebted to many people for their contributions to, and support of, this dissertation work. First and foremost, I thank the co-chairs of my dissertation committee, Dr. Betsy Foxman and Dr. Zhenhua Yang. I feel incredibly lucky to have had the opportunity to work with two such superb mentors, and to have benefited from their distinct working styles and areas of expertise. I have also had the tremendous fortune of having a dissertation committee that has provided guidance and advice from the start, as I ventured into areas that were a bit outside of my comfort zone. Dr. Carl Marrs has consistently provided advice and feedback that always seems to cut straight to the core of the matter. The expertise and enthusiasm of Dr. Denise Kirschner has been a constant motivation, setting a high bar to aspire to. And, I am particularly indebted to Dr. Rick Riolo. Without his endless help, otherworldly patience, and vast expertise, the modeling work included in this dissertation would not have been possible. I also thank Dr. M. Donald Cave, Dr. Joe Bates, and Dr. Leonard Musaka, at the Arkansas Department of Health, and the UAMS, for their help and support in conducting the epidemiologic investigation included in this dissertation, as well as current and former students in the Yang Lab and MAC-EPID, who have always made work much more enjoyable. Finally, I would like to acknowledge the Interdisciplinary Program in Infectious Disease and Rackham Graduate School, from which I have received funding to support this dissertation work.

## Preface

This dissertation integrates molecular typing, epidemiology, and agent-based modeling methods to characterize trends in the transmission of tuberculosis (TB) in a rural state in the southern United States, to assess the validity of molecular typing data as a marker of transmission in this population, and to gain insight into the relationship between host and microbial population factors and the information that molecular typing provides. The first chapter introduces the epidemiology, transmission, and natural history of TB infection, as well as the principals of strain typing methods, current issues in the control of TB in the US, and the use of mathematical and simulation models in the study of TB. Epidemiology is by nature a multi-disciplinary field, and this sometimes broad-ranging introductory chapter is a reflection of that.

In addition to the brief background provided in the introduction, detailed background sections on the transmission and natural history of TB infection (Appendix 1), and the molecular typing of TB (Appendix 2), can be found at the end of this dissertation. These detailed appendices provide key background to inform decisions that were made in the formulation of an Agent-Based model of TB transmission, which is presented in Chapter IV of this dissertation.

Following the introduction are reports of three studies addressing different aspects of the molecular epidemiology of TB, each with a unique research approach. Chapter II presents an investigation of TB incidence trends in Arkansas between 1996 and 2003, in which molecular typing was used as a tool to estimate the relative importance in changes in active TB transmission to overall disease trends. Chapter III reports the results of a validation study, in which extensive epidemiologic interviews, medical records, and TB control records were reviewed to establish evidence of epidemiologic linkages between patients infected with isolates that exhibited the same molecular typing pattern as the isolate from at least one other case in the study population. By comparing these pieces of information, host and microbial factors that were strongly correlated with the validity of molecular typing results in this population were identified. Chapter IV describes the development of an agent-based simulation model of TB transmission where molecular typing patterns were explicitly represented. Insights that were gained through the analysis of this model, which clarify the relationship between *M. tuberculosis* strain diversity,

typing marker stability, host demographic factors, and the validity of molecular typing data, are described. Finally, Chapter V presents a discussion of the overall findings of these three investigations, and of their contribution to the current understanding of issues related to the molecular typing of *M. tuberculosis.*

## Table of Contents

## List of Tables

## List of Figures

## List of Appendices

# Abstract

Molecular typing is increasingly integral to tuberculosis (TB) control programs, providing public health practitioners with a tool to characterize transmission patterns, track the emergence and spread of strains of particular medical and public health importance, and to identify transmission venues that contribute to the persistence of *M. tuberculosis* in populations. While molecular typing is already used extensively as a tool for TB control in many diverse populations across the globe, the sensitivity of molecular typing-based measures to characteristics of both the host and microbial populations is not well understood. To better characterize the relationship between key host and microbial factors and the validity of molecular typing measures, and to generate insights that may inform the design of a rational typing system for TB control, this dissertation work employs a multi-disciplinary research strategy which integrates molecular, epidemiologic, and computer-simulation data. In the rural, stable population of Arkansas, we found that a declining incidence of TB between 1996 and 2003 resulted primarily from a declining incidence of TB due to the reactivation of remotely acquired infection, rather than recently acquired infection. This work suggested the influence of a strong cohort effect on disease patterns in this population. A validation study of molecular typing in this same population, in which extensive epidemiologic interview data were compared to molecular typing results, identified a number of host and microbial factors associated with the validity of typing results. This study also suggested the presence of a regionally endemic strain family which was associated with false positive molecular typing results. Using an agent-based model of TB transmission, we conducted the first quantitative assessment of the importance of the diversity and stability of molecular typing markers, as well as historic and demographic characteristics of the host population, to the validity of typing results. The results of these investigations contribute to an improved understanding of the dynamics of TB transmission in rural populations of the United States, and also highlight key factors that should be considered in the interpretation of molecular typing results in all populations. Additionally, these results may inform the development of more rational approaches to the design of molecular typing systems used in TB control.

# Chapter I

# Background

''If the number of victims which a disease claims is the measure
of its significance, then all diseases, particularly the most
dreaded infectious diseases, such as bubonic plague, Asiatic
cholera, et cetera, must rank far behind tuberculosis.''
-ROBERT KOCH, 1882

## Global epidemiology of tuberculosis

Tuberculosis (TB) is a long familiar foe to human kind– references to the disease can be found in the writings of Hippocrates, and evidence of infection has been identified in prehistoric human remains, both in the old and new world [1]. Despite the development of effective chemotherapy in the 1950s, this "captain of all these men of death" remains a leading cause of mortality worldwide. In 2005, an estimated 8.8 million new cases of TB and 1.6 million deaths due to the disease occurred globally [2]. Estimates suggest that one-third of the world's population is infected with *Mycobacterium tuberculosis*, the bacterium that causes TB disease [2]. The burden of TB falls disproportionately on the developing world, with more than 95% of new cases, and 99% of deaths due to the disease, occurring in low and middle-income countries [3]. The increasing prevalence of multi-drug resistant (MDR) TB, and the recent recognition of extensively-drug resistant (XDR) TB, further magnifies the dark shadow cast by this ancient infection.

## Epidemiology of TB in the United States

As in many developed nations, the incidence of TB in the United States (US) declined dramatically over the last century, with the rate of active disease falling from 53.0 per 100,000 in 1953, when TB reporting began in the US, to 4.6 per 100,000 in 2006 (CDC 2007). Despite tremendous gains, the problem of TB in the US has not disappeared:
13,767 cases were reported in 2006 [4], a rate of 4.6 cases per 100,000 individuals.

While infrequently seen in the mainstream, TB remains firmly entrenched in the poorest and most marginalized of the US population [5]. Historically, the burden of TB has fallen most heavily on disadvantaged populations. This pattern remains true in the US, with high levels of the disease seen among homeless, drug-injecting, and prison populations [6]. HIV infected individuals, already at higher risk of developing TB disease upon infection due to depressed immune function (leading to an increased risk of developing primary or reactivation disease following infection with *M. tuberculosis*) [7], often fall into other high-risk groups, magnifying this pattern[8, 9]. Strong racial disparities are seen in the distribution of TB in the US: while 12.3 percent of the US population self-identifies as African American, more than 45 percent of cases reported among the native born population in the US occur in this group. The global TB epidemic exerts a substantial influence on the prevalence and incidence of TB in the US, and more than half of all cases reported in the US in2005 occurred in individuals born outside of the country [4].

## Etiology of TB disease

TB is caused by infection with bacterium of the *Mycobacterium tuberculosis* complex, which includes *M. tuberculosis, M. bovis, M. africanum, and M.microti.* In humans, *M. tuberculosis* is the most important cause of TB, with rare reports of disease caused by *M. africanum* [1]. *M. bovis*, which is an important cause of disease in cattle and deer populations, can cause disease in humans that is clinically, radiologically, and pathologically indistinguishable from disease caused by *M. tuberculosis*, and was once an important cause of TB disease in children. *M. bovis* is no longer considered to be an important cause of disease in humans, as a result of the introduction of pasteurization processes for milk and milk products, and the control and eradication of *M. bovis* in animal populations in the early part of the 20th century [10].

*M. tuberculosis* was first isolated in pure-culture by Robert Koch in 1882. Like other members of the genus *Mycobacterium, M. tuberculosis* is a slow-growing intracellular bacillus, characterized by a lipid-rich cell wall and a unique, lipid-rich outer membrane which acts as an effective barrier to prevent the entry of antimicrobials into the cell and helps the organism to resist phagocytosis by cells of the host immune system [11]. The mycobacteria are aerobic, non-motile, and do not form spores. The characteristic high-lipid content of their cell wall results in distinctive staining properties, allowing mycobacteria to be identified by acid-fast staining techniques. In culture, the doubling time of mycobacteria is 12-36 hours, and a detectable concentration of organisms is present after growing for an average of 15.4 days in liquid culture [11].

## Transmission and pathogenesis of *M. tuberculosis* infection

(A detailed review of the epidemiology of transmission disease outcome, which provides detailed supporting information for the architecture of the agent-based model presented in chapter IV of this dissertation, can be found in Appendix 1)

TB transmission occurs via droplet nuclei, tiny particles containing *M. tuberculosis* which are produced when individuals with respiratory TB infections sneeze, speak, sing, or induce sputum. Droplet nuclei can remain airborne for time spans ranging from minutes to hours, and result in the transmission of *M. tuberculosis* when inhaled into the lungs of a susceptible individual [1]. Although droplet nuclei may also be produced by the manipulation of tubercular lesions or processing tissue in the hospital or laboratory, such events are rare, and individuals with non-respiratory infections do not generally represent a transmission risk [12].

Following infection, an individual may develop infection within a short interval ("primary disease"), many years later ("reactivation disease"), or not at all [7]. The highest risk of developing active TB disease occurs in the first 24 months following infection, and an estimated 5-10% of newly infected individuals will develop disease within this interval (primary progressive disease). The other 90-95% will be able to contain the infection, though not clear it, through a successful cell-mediated immune response that walls off the organisms in formations called granulomas. These infected, disease-free individuals are then in a state of latent infection. While many will never go on to develop tuberculosis disease, an estimated 5-10% will, with the disease re-activating at a time of decreased immunity [12].

Although active tuberculosis infection most often affects only the lungs (85% of infections in HIV negative individuals), the disease can manifest in almost any organ system, including the central nervous system, lymphatic system, bones and joints, and the genitourinary tract [7]. Disseminated tuberculosis, in which many organs are simultaneously involved, can also occur. Extrapulmonary disease is more common in HIV infected individuals, accounting for more than 50% of cases among these individuals [13]. Among HIV negative individuals, extrapulmonary disease is particularly common among women and young children [14].

With rare exception2, only individuals with active, respiratory forms of the disease transmit infection, and individuals with a higher number of bacilli in their sputum pose a greater transmission risk. These individuals are often identified by a positive sputum smear (and hence are referred to as “smear positive” cases), when acid-fast bacilli are detected by microscopy [12]. Certain clinical characteristics are associated with higher quantities of bacilli in the sputum. Cavitary disease, which is characterized by extensive necrosis of bronchial airways, allows the discharge of infectious material into the airways. Patients with cavitary disease have a greater frequency of cough, and are more likely to have a positive sputum smear [12]. While the infection risk presented by smear-negative cases was long considered inconsequential, recent reports suggest that transmission from these cases may be responsible for a significant proportion of disease in some populations [15].

### Treatment and Disease Recurrence

Standard treatment for TB involves antibiotics taken in combination over a long course of therapy. The most standard “short course” treatment regimen consists of a 2 month course of therapy with isoniazid, rifampicin, pyrazinamide, and ethambutol, followed by a 4 month course of isoniazid and rifampicin [16]. This treatment regiment is highly effective. However, following treatment with standard short course chemotherapy, active TB recurs in 2 to 7% of cases [17]. It was long held that, following a primary infection with *M. tuberculosis*, an individual acquired relative immunity against subsequent re-infection [18]. All subsequent episodes of active disease at any point in an infected individual’s life, and at any anatomic site, were considered to be due the reactivation of dormant bacilli of the original infecting strain. While this is often the case, a number of reports in recent decades have documented the occurrence of re-infection events in individuals who had a previously resolved case of active TB [19-23].

## Strain typing of *M. tuberculosis*

(A detailed review of the strain typing methods in *M. tuberculosis*, which provides detailed supporting information for the motivation and formulation of the agent-based model presented in chapter IV of this dissertation, can be found in Appendix 2.)

### Principles of epidemiologic typing

The ability to subdivide bacterial species groups into smaller and more homogenous groups of related organisms is critically important in evolutionary, biomedical, and epidemiological research. In epidemiology, fine-grain classification of bacterial relationships allows the identification of subgroups, often identified as "strains", of medical and public health importance: for example, strains more likely to cause severe disease, or which are resistant to commonly used antibiotics. If isolates of a bacterial pathogen can be subdivided at a fine enough scale, isolates related by a transmission event may be identified, allowing investigators a powerful tool to study patterns of infection spread. The techniques and approaches of epidemiologic typing overlap considerably with those of molecular taxonomy, phylogeny, and population genetics [24], but are ultimately driven by more immediate, applied goals. Strain typing methods have proven invaluable to epidemiologic researchers and practitioners alike, providing a tool to identify factors important in determining the distribution of disease, the transmission, manifestation, and progression of infection, and to identify novel opportunities to intervene in and interrupt disease spread [24].

### History of strain typing and early *M. tuberculosis* strain typing methods

The earliest tools to subdivide bacterial species were based on phenotypic assays. The development of the first "strain typing" technique is often credited to Rebecca Lancefield, who developed a serologic assay that subdivided the hemolytic streptococci, known then as *Streptococcus haemolyticus,* based upon the expression of cell-surface antigens [25]. At the time, *Streptococcus haemolyticus* was known to cause disease in both humans and animals. Using the serotyping system she developed, Lancefied was able to demonstrate that a subgroup of this species, now known as group A streptococci, was specific to humans and human disease, including pharyngitis, scarlet fever, rheumatic fever, nephritis and impetigo.

The value of a strain-typing as an epidemiologic tool for the study of TB was recognized early, but efforts to develop a serotype-based strain typing technique were stymied, as *M. tuberculosis* was shown to comprise a single homogenous serogroup [26]. The recognition of the selective sensitivity of *M. tuberculosis* isolates to lysis by a number of distinct mycobacteriaphages, therefore, was a welcome discovery which lead to the first viable strain-typing tool for TB [27, 28]. This technique involved infecting a given *M. tuberculosis* isolate with a panel of mycobacteriaphages, and determining the sensitivity of the isolate to each, on the basis of cell-lysis. The binary "yes/no" lysis scores for each phage were concatenated into a string, which represented the "phage type". While never widely used on a population-based scale, phage typing did prove useful in the investigation of TB outbreaks and laboratory cross-contamination [29, 30]. Phage typing suffered, however, from a limited number of phage-types that could be generated using known mycobacteriaphages, greatly hampering the resolution with which this technique could differentiate *M. tuberculosis* strains. Typing techniques based on other phenotypic characteristics, such as metabolic features and susceptibility to antibiotics, were similarly limited by the limited number of strain-types that could be resolved [31].

### Genotypic typing assays for *M. tuberculosis*

The development of molecular techniques to sequence and manipulate nucleic acids presented a watershed to the development of TB typing tools, allowing variability on a genotypic, rather than phenotypic, level to be assayed. Directly assaying genotypic variability is particularly advantageous, as it also overcomes the problem that a single organism may express multiple phenotypes, depending on the environmental conditions it experiences. Genotype-based typing assays, referred to as molecular typing techniques, allow substantially higher levels of resolution than could be attained using phenotypic methods, while at the same time resulting in more robust, and hence reproducible, strain types.

Molecular typing techniques for *M. tuberculosis* exploit variable regions in an otherwise highly-conserved genome to generate DNA 'fingerprints' which are specific to a particular strain. The most commonly used technique makes use of the transposable element IS*6110*, which varies in both copy number and location throughout the genome [32]. Other commonly employed techniques include spacer oligonucleotide typing ("spoligotyping"), which assays the pattern of conserved sequences in a direct-repeat locus[33], and pTBN12 typing, which assays the number and distribution of a polymorphic GC-rich tandem repeat sequence (PGRS)[34]. pTBN12 typing and spoligotyping are often used as secondary typing methods for IS*6110* isolates with few copies of the IS*6110* element, as the variability, and hence, discriminatory power, of IS*6110* is reduced in these strains [34]. They are also used as secondary typing techniques for isolates with IS*6110* fingerprints that are nearly identical (differing by 1 band in the pattern) [35]. Recently, MIRU-VNTR, a typing technique which assays the number of tandem repeats at a number of repetitive loci throughout the TB genome [36], has begun to supplant IS*6110* typing as the most common molecular typing technique for TB, and has been adopted as the primary molecular typing method for routine TB typing in the US [37].

### Advances achieved through molecular typing

The ability to identify relationships between *M. tuberculosis* isolates at a fine scale has allowed considerable advancements in the understanding of the natural history, pathogenesis, and epidemiology of *M. tuberculosis* infection, providing a tool to investigate questions that previous research methods were unable to address. Studies utilizing molecular typing tools have shown, for example, that an individual can be infected with a new strain of the organism following a primary disease episode (exogenous re-infection), and can be simultaneously infected with multiple distinct strains, implying that complete immunity to infection does not develop following active disease [19, 38, 39]. Molecular typing studies have confirmed that *M. tuberculosis* may reactivate following decades of latent infection [40], and have revealed that transmission of *M. tuberculosis* can occur in casual settings, without the prolonged contact between infectious case and susceptible individual that was previously considered essential for transmission to occur [41].

One of the most important applications of molecular typing techniques is to the identification of TB disease due to recent transmission in the population [42-44]. By distinguishing disease due to recent infection from that due to the reactivation of latent infections, molecular typing provides a powerful tool for applied TB epidemiology and TB control.

## Control of TB in the US

While the incidence of active TB disease in the US is one to two orders of magnitude lower than that in many high-burden countries, such as the estimated 75 per 100,000 in Brazil and 639 per 100,000 in Cambodia [45], restraining the spread of TB in the US depends still on consistent, vigilant public health attention. In the 1970s and 1980s, following a period of declines in the global prevalence of TB infection, and soon after Surgeon General William H. Stewart famously noted in 1965 that “it is time to close the book on infectious diseases”, TB control programs, both in the US and abroad, were effectively dismantled [3]. Not long after, in the late 1980s and 1990s, rates of TB resurged. Between 1985 and 1992, the incidence of TB in the US increased by 20 percent [8].

In 1989, the CDC announced the goal of eliminating TB from the United States by the year 2010. After a resurgence of TB in the early 1990s, the feasibility of this goal was re-evaluated by the Institute of Medicine in 1998, which concluded that while still feasible, it would require "aggressive and decisive action beyond what is now in effect"[46]. Behind this ambitious goal is a plan that calls for the prevention of transmission of TB through timely diagnosis and treatment of those with active disease, rapid recognition of TB transmission through the use of DNA fingerprinting methods, and rapid outbreak response to quell further transmission. Additionally, the plan calls for consistent monitoring and evaluation of the progress towards the goal of tuberculosis elimination.

The patterns of TB transmission that allow TB to persist in the contemporary US setting are unlike those of a half-century ago, when the substantial bulk of transmission is thought to have occurred within the household setting [47-49]. Transmission in hospitals, prisons, nursing homes, and homeless shelters is increasingly important in sustaining transmission in the US [50-54]. In addition to these venues, where TB transmission is relatively well characterized, evidence suggests that transmission in previously unsuspected contexts contributes to sustained transmission [49, 55], and transmission in casual encounters, once considered improbable, may also play an important role [48, 56]. These non-traditional transmission events often elude TB control investigations [57], as the sites and circumstances where such transmission occurs are not well understood.

In order to most effectively focus TB control efforts towards pockets of transmission that are perpetuating the infection, the identification of transmission occurring in previously unsuspected circumstances is essential. The complex natural history of TB infection presents an obstacle to TB control programs—even if all transmission in a population has stopped, there still may be disease due to the reactivation of infections that were the result of historic transmission. In any interval, the incident cases of TB in a population will be due to both active transmission in the population (which public health efforts may be able to contain), and the re-activation of historically acquired infections (which public health efforts aimed at interrupting transmission would be unable to prevent).

Traditionally, TB control programs have relied upon contact tracing investigations to identify sources of ongoing transmission in a community. These investigations are time and labor intensive, relying on patient interviews to identify close contacts of a recently reported case of active TB. Current contact tracing evaluation protocols follow a “stone in a pond”, or concentric circle model, which was developed at a time when most contacts occurred within the home and family [47, 58, 59]. This approach is limited in its ability to detect transmission outside of conventional settings: by design, the settings and contexts in which transmission linkages are looked for are those where they are expected to be found. As changing transmission patterns shift transmission away from these expected venues, and as an increasingly marginalized population of TB patients may be less likely to trust, and therefore cooperate with, public health authorities, the effectiveness of contact tracing is increasingly limited. As few as 5-10% of cases linked by recent transmission are identified as such through routine contact investigations [42, 60-62]. Given its limitations, contact tracing alone is unlikely to be sufficient to clearly characterize TB transmission patterns that contribute to sustained transmission of TB in contemporary US settings [55].

Adequate TB control resources must be directed towards every incident TB case, in order to ensure effective treatment, limit the risk of transmission to others, and identify recently infected contacts that may benefit from prophylactic isoniazid (INH) treatment. For the purpose of TB control, however, cases due to recently acquired infection may be more important, as they often indicate transmission that is still ongoing in a community. Further investigation of such cases may reveal yet-undetected cases of active TB infection that are perpetuating transmission, as well as informing an improved understanding of transmission in the current context.

Molecular typing is increasingly integral to TB control programs, providing public health practitioners with a tool to effectively focus limited TB control resources to reduce ongoing transmission. By allowing cases related by transmission to be identified, molecular typing has contributed not only to the control of outbreaks, but has also enhanced our understanding of the locations and contexts in which TB transmission may occur. To this end, the CDC has made the universal genotyping of incident TB cases a priority, and with this support, the routine integration of molecular typing into population-based TB surveillance programs is becoming the norm in the US [63].

Population-based (or Universal) strain typing of *M. tuberculosis* isolates provides a powerful tool for TB control staff tool to characterize transmission patterns in a population. Through the application of population-based molecular TB typing, investigators have identified transmission in settings where it was previously not thought to occur, including transmission from a cadaver during autopsy [64], among unacquainted individuals who frequented the same bar, but had no clear contact with one another [65], unacquainted individuals visiting the same work site [41], and between individuals whose only contact was singing in the same church choir [66]. Such observations have deepened our understanding of TB transmission, and highlighted the deficiencies of traditional contact tracing methods to identify these transmission links [41, 57, 62]. This tool has also allowed investigators to the burden of disease due to active transmission in a population, to identify trends in active transmission over time [67-69], and clarify risk factors for recent transmission [43, 55].

## Validation of molecular typing to identify recent transmission

In the investigation of TB outbreaks, where strain typing of TB was first applied, the interpretation of genotyping data is straightforward: *M. tuberculosis* isolates from cases suspected to be linked by transmission are genotyped: identical "molecular fingerprints" are considered to confirm a transmission link, while discordant fingerprints negate it.

In the context of a population-based typing program, however, interpreting genotyping data is more complicated. In this setting, TB control staff will rarely have information on the epidemiologic relationships among cases prior to the consideration of genotyping data. In the absence of such epidemiologic data, genotyping data are relied upon to direct the attention of TB control staff towards "clusters" of related cases that warrant further investigation and TB control efforts.

Numerous reports from population-based typing programs have documented "clusters" of cases related by genotype pattern, but among which epidemiologic investigation reveals no linkage between cases [70-72]. In one such report, additional, rigorous epidemiologic investigation identified links between cases in genotype-linked clusters that had been missed by conventional contact tracing methods [57], suggesting that linkages between individuals for whom contact information is difficult to obtain, or between whom transmission occurred during casual contact, may explain such clusters. In some settings, "clustered" cases may be linked by transmission that occurred many years or even decades in the past [70]. Another explanation is that the resolution of the genotyping tool used is insufficient to discriminate between unlinked cases in a given population. In this context, cases may be "clustered" by one genotyping method but be unrelated by transmission. Using a genotyping tool with a higher level of discrimination, such false clusters may be shown to contain unrelated isolates [73].

A major limitation in the application of molecular clustering data to identify active transmission is the lack of epidemiologic validations of marker systems in population-based studies, and evaluations in different populations with different disease burdens and transmission dynamics. While considerable effort has been put into the development of new typing techniques, evaluations of these tools have been limited, often relying on comparisons with other typing methods [74-76]. As no current method provides a 'gold standard' for comparison, such evaluations cannot substantiate the use of these markers to draw inferences regarding the epidemiologic relationships between isolates.

Ideally, molecular typing data would be validated using population-based samples of *M. tuberculosis* isolates, with multiple samples representing diverse populations. Genotyping results would be compared to data on the actual transmission relationships between the sampled cases, and the sensitivity, specificity, and both positive (PPV) and negative predictive value (NPV) of clustering for recently transmitted cases, would be calculated. These measures are used to assess the validity of diagnostic tests, with the PPV indicating the probability that an individual who tests positive truly has the disease or disorder in question, and the NPV indicating the probability that an individual who tests negative truly does not have the disease or disorder in question [77]. With data from a diverse range of populations, it would be possible to evaluate the sensitivity of these measures to key features of the host population, bacterial population, and molecular typing system used. Unfortunately, even in an ideal study setting, data on actual transmission linkages is unattainable. Epidemiologic investigations of transmission linkages between cases provide data that is invariably ambiguous and often incomplete, and transmission events that occur in unsuspected contexts, or by casual contact, are unlikely to be identified [57]. Additionally, these investigations are sufficiently time and resource intensive that few validations of population-based genotyping data against epidemiologic data have been undertaken [42, 70, 78]. Even with the best available epidemiologic data, only the PPV can be determined, and, as it is inevitable that true transmission links between clustered cases will be missed, even that measure cannot be stated with confidence.

## Mathematical and simulation modeling in TB research

Some of the best insight into the utility of molecular typing data to identify recent vs. remote transmission has come from mathematical models and computer simulations of TB transmission. Simulations are particularly powerful in the context of TB epidemiology as they can provide what is elusive in epidemiologic investigations in the field: an underlying distribution of cases that represents the "truth". From this distribution, data can be sampled following various strategies, and inferences made from sampled data can then be compared to the true distribution. Using this approach, investigators have shown that including less than the full proportion of cases in a sample leads to underestimates of recent transmission, as does sampling cases over a short time period [79]. In addition to providing a known distribution of cases from which to sample, mathematical and simulation models of transmission allow investigators to perform sensitivity and uncertainty analysis, varying parameters relevant to transmission and observing the impact that these factors have on both the dynamics of disease in a population and the accuracy of measurement tools and sampling strategies. Equation-based models of TB have been successfully employed to address questions at multiple levels, including the dynamics of TB transmission in large populations [80], infection outcomes on the level of individual hosts [81], and cellular interactions between *M. tuberculosis* and cells of the host immune system [82]. Molecular typing has been infrequently taken on by equation-based models. One model that did, by Vynnycky et al., assessed the impact of the age distribution in a population and the historic risk of TB infection over the life of that population on the predictive value of clustering statistics [83].

Models can provide great insight into the behavior of dynamic systems. The inferences drawn from any model, however, rely heavily upon the validity of the assumptions underlying the model. When data on the system being modeled are sparse, confident assumptions can be difficult to make. For example, it is widely believed that, within a latent granuloma, *M. tuberculosis* is in a dormant state, with little or no replication [84]. Evidence from an investigation of epidemiologically linked TB cases in the Netherlands, which found identical IS*6110* RFLP patterns in isolates separated by decades of latency, is consistent with this understanding of latent infection [40]. However, latent infection has recently been characterized as a dynamic microenvironment, with continuous activation of the immune response to restrain replication of the bacteria [85], and it has been suggested that both dormant and replicating *M. tuberculosis* might be present in latently infected individuals within different types of lesions [84].

To best address uncertainty such as that surrounding the dynamics of latent infection, sensitivity analysis, which allow an investigator to determine how robust a given model is to violations in the underlying assumptions, is essential. Sensitivity analysis has the objective of comprehensively and quantitatively evaluating the response of a particular model output to variation in selected inputs. Sensitivity analysis can identify critical parameters in a system and flaws in model design. In her EBM of molecular TB typing, Vynnycky [83, 86] allowed for the mutation of molecular fingerprint patterns, and designed her model based on the assumption that the rate of mutation of strains involved in latent infections was identical to that observed in strains involved in active infections. Whether or not molecular typing patterns evolve during latency is unclear, and this assumption is likely to exert a strong influence on the molecular typing results generated by a model. A sensitivity analysis of this assumption might show, for example, that Vynnyky's assumption of a constant rate of mutation throughout latent infection is not of importance to the conclusions she derived from the model. Because such an analysis wasn't presented, however, it is difficult to accept her results with confidence.

While EBM methods have provided considerable insight to the study of TB, the limitations of this modeling approach make them ill-suited to the study of genotyping data, where the many key components of the system-- individual hosts as well as bacterial strains – would quickly overwhelm a compartmental model. Additionally, many of the assumptions of a system demanded by EBMs – including random, instantaneous mixing of the population and an infinite population size – present considerable obstacles to the study of genotyping and "clustering", where the social mixing patterns of the human host population is of key importance.

Agent based models (ABMs), which are also known as Individual-based models (IBMs), are a relatively young modeling approach that is increasingly applied to the study of complex, adaptive systems. In contrast to EBMs, ABMs take a "bottom up", rather than "top down" approach to representing systems. ABMs represent the elements of a system as discrete entities with unique characteristics, rather than representing the distributions of individual attributes characteristic of EBMs [87]. ABMs aren't limited by assumptions of infinite population size or instantaneous total mixing made by EBMs, and the implementation allows for social mixing patterns to be more realistically specified. Additionally, because ABMs don't rely on a compartmental structure, they can handle significantly more complexity.

Two major ABMs have been developed for the study of TB. One focuses on dynamics at a cellular level, representing the process of granuloma formation in a TB-infected lung [88]. This ABM was uniquely able to represent the granuloma formation process, capturing the complex spatio-temporal interaction of bacteria, immune cells, and immune effectors. The second ABM was applied to the study of the molecular epidemiology of TB, with the goal of evaluating the hypothesis that the bacterial strains involved in large clusters are more transmissible or pathogenic than strains involved in small clusters [89]. This ABM simulated TB transmission in a hypothetical population, and tracked the identity of individual strains as they were transmitted. Data generated from this ABM suggested that cluster size varied with a range of host and population characteristics, and that a large range in cluster size would occur in a setting where all strains were of equal transmissibility and pathogenicity. While not applied to the evaluation of the validity of clustering as a measure of recent transmission, this model illustrated the utility of the ABM method to such an end.

## Rationale and Research Objectives

The majority of studies of the molecular epidemiology of TB, and particularly studies that have validated molecular typing results against epidemiologic evidence of transmission linkages have taken place in large urban populations.  Evidence from low-incidence populations suggests that the dynamics of TB transmission, and the validity of molecular typing-based measures of recent transmission, may vary considerably between large urban populations and rural populations. Large urban centers typically have very high levels of in and out migration, and serve populations from broad geographic areas, providing many opportunities for the introduction of diverse strains. By contrast, reports from a number of rural, low incidence populations suggest a high prevalence of regionally endemic strains, and low overall strain diversity [90, 91].   Endemic strain transmission, and a highly homogeneous *M. tuberculosis* strain population that may result, has also been described in high incidence settings such as Russia and South Africa [92, 93]. The low level of strain diversity resulting from endemic transmission is likely to present an obstacle to molecular typing, making it more difficult to discriminate between isolates that are not related by recent transmission events.  While molecular typing is already used extensively as a tool for TB control in rural areas of the United States, the dynamics of tuberculosis transmission in these populations, and the sensitivity of molecular typing-based measures to key characteristics of these populations, are not well understood. To clarify understanding of TB transmission and molecular typing in rural-low incidence areas, and to generate insights that may inform the design of a "rational" typing system for TB control, we have employed a multi-disciplinary research strategy which integrates many types of data, including molecular, epidemiologic, and computer-simulation data.

## Chapter II

## What’s driving the decline? A molecular epidemiologic analysis of tuberculosis trends in a rural, low-incidence population[1]

### Introduction

Following a resurgence in the late 1980s, the incidence of tuberculosis (TB) in the United States has been in steady decline, decreasing by 44 percent between 1993 and 2003 [94] and reaching a historic low of 4.8 cases per 100,000 in 2005, the lowest rate since national reporting began in 1953 [95]. TB incidence rates are not consistent across populations, however, and gaps in incidence between different race/ethnic groups and between US and foreign born persist [95]. In order to best focus resources towards the goal of the elimination of TB in the United States, it is important to understand what factors have driven the decline, and how these factors vary across sub populations.

Because of the complex natural history of TB, incident cases in a population may be due to infections that were acquired recently, and therefore represent evidence of active chains of transmission, or they may be due to the reactivation of latent infections acquired years or even decades ago [7]. Clinically, it may be difficult to distinguish between recently acquired and reactivated disease [96], but the frequency of each type in the population has important implications for infection control.

[1] This chapter was previously published as a manuscript in the *American Journal of Epidemiology*: France AM, Cave MD, Bates JH, Foxman B, Chu T, Yang Z. What's driving the decline in tuberculosis in Arkansas? A molecular epidemiologic analysis of tuberculosis trends in a rural, low-incidence population, 1997 2003. *Am J Epidemiol.* 2007 Sep 15;166(6):662-71. Epub 2007 Jul 11.

DNA genotyping of *M. tuberculosis* isolates provides a tool to draw inferences about the transmission history of a clinical isolate.  Cases infected by isolates with identical or highly similar DNA genotyping patterns, identified as clusters, reflect a common chain of transmission [90], and are considered to be caused by the same strain.  Clusters occurring within a short time period are considered to reflect active transmission followed by rapid progression to clinical disease. A number of studies have used clustering analyses to estimate the proportion of disease that results from recent transmission [42-44, 97].

In the diverse, urban population of San Francisco, investigators used molecular genotyping to analyze TB trends between 1991 and 1997.  Based on a decline in the incidence of clustered TB cases over time, the authors concluded that the reduction in the overall incidence of TB in San Francisco was driven by decreasing levels of active transmission, owing to the successful implementation of enhanced TB control programs in that population  [67].

We took a similar analytic approach to investigate TB trends in a very different population: the mostly rural, highly stable population of the state of Arkansas.  In Arkansas, the reported incidence of TB declined from 7.9 cases per 100,000 in 1997 to 4.7 cases per 100,000 in 2003.  It is uncertain whether the decline in TB incidence in Arkansas can, like that of San Francisco, be attributed to a decrease in recent transmission.  We hypothesize that, in the very distinct populations of San Francisco and Arkansas, the relative contribution of active transmission and reactivation of latent infections to the overall burden of TB may differ despite similar trends in the overall incidence and comparable control efforts.

# Methods

## Arkansas demographics

Demographic information used to characterize the population of Arkansas was obtained from the 2000 United States census data [98] and the Institute for Economic Advancement at the University of Arkansas, Little Rock. National and state TB rates were obtained from Centers for Disease Control surveillance [99-105]. Metropolitan statistical areas (MSA), defined by Arkansas' Office of Management and Budget were used to indicate urban areas: non-MSAs were considered to be rural. MSA areas included Fayetteville-Springdale-Rogers, Fort Smith, Jonesboro, Little Rock-North Little Rock, Memphis, Pine Bluff, and Texarkana. All counties within these MSA were considered urban, while all counties outside of these MSA were considered rural.

## Study Population

All patients culture-positive for *M. tuberculosis*, for which the first positive culture was collected between 1 January 1996 and 31 December 2003, were included in the study population. For each of these cases, standard demographic information was collected using the Centers for Disease Control and Prevention's "Report of a Verified Case of Tuberculosis" form. Annual case rates of TB were calculated per 100,000 using yearly population estimates from the National Center for Health Statistics. Race and ethnicity were based on self-report. Options for race were white, black, American Indian/ Alaska native, and Asian/ Pacific Islander. Options for ethnicity were Hispanic and Non Hispanic. These options were specified by the Centers for Disease Control and Prevention's "Report of a Verified Case of Tuberculosis" form. This study was approved by the Health Sciences Institutional Review Boards of the University of Michigan and the University of Arkansas for Medical Sciences.

## Genotyping

For each culture-confirmed isolate, IS*6110* restriction fragment length polymorphism (RFLP) patterns were determined using standard procedures, as previously described [106]. For all isolates with fewer than six IS*6110* bands, and for isolates with six or more IS*6110* bands that differed from another IS*6110* pattern by one band, spoligotype patterns were also generated, following standard protocol [33].

## Cluster Definition

Clusters of cases sharing identical or highly similar fingerprints may include cases for whom epidemiologic evidence of recent transmission between other cases in the cluster cannot be found [70, 107], or among whom epidemiologic evidence suggests a transmission event that occurred in the remote past [70]. In order to increase the specificity of our clustering measure for recent transmission, we used a time-restricted definition of clustering in addition to the standard definition of clustering, which we refer to here as the conventional clustering definition.

A 'conventional' cluster was defined as two or more patients whose isolates were identified as related by IS*6110* RFLP and spoligotyping. Related isolates were defined as follows: isolates with more than five IS*6110* bands with identical IS*6110* RFLP patterns or IS*6110* RFLP patterns differing by one band but with identical spoligotype patterns, or isolates with five or fewer bands with both identical IS*6110* RFLP patterns and identical spoligotype patterns.

A case was considered to be part of a 'time-restricted' cluster if it was clustered by the conventional definition with another isolate diagnosed within the one year period prior to its diagnosis date. In the literature, the cut-off to distinguish recent from reactivation disease is arbitrary, ranging from one year [67] to as many as five years in some studies [81, 108]. We chose a one year period for our definition of clustering to yield the greatest possible specificity of our definition, and also to allow our analysis to be directly comparable to previous reports regarding clustering trends over time [67].

## Statistical Analysis

The time-restricted cluster definition was used as our cluster definition for all of our analysis of trends over time. Annual TB incidence rates and chi-squared analysis of demographic variables by time-restricted cluster were generated using SAS version 9.1 [109]. Confidence intervals for yearly TB case rates were generated assuming a Poisson distribution, using the method described by Buchanan 2004 [110].

All isolates for which molecular fingerprint data were available were included in a Kaplan-Meier analysis of the distribution of the time between diagnosis dates of matching fingerprints, as suggested by Jasmer and colleagues in 1999 [67]. This survival analysis was used to estimate the probability that an isolate with a matching DNA fingerprint pattern occurred within a given time period following each culture-confirmed case. “Failure time” was the time between diagnosis dates of consecutively matching fingerprints. Isolates that did not match another isolate in the study by the end of the study period were censored.

## Results

### Demographics

In 2000, the population of Arkansas was 2,673,400 [98]. This population is largely rural, with just 29.3 percent of the population living within a MSA in 1990, compared with 77.5 percent of the US population [111]. This population is also relatively stable: in 1990, 88 percent of Arkansas residents over 5 years old had lived in the state in 1985, 79 percent had lived in the same county, and 54 percent lived within the same house, compared to 70 percent, 63 percent, and 42 percent, respectively, in the total US population [112].

### TB Cases

Arkansas reported 1402 incident cases of TB between 1996 and 2003. Of the reported cases, 1025 (73.1 percent) were confirmed by bacterial culture.

### Genotype Patterns and clustering of isolates

IS*6110* RFLP patterns were generated for 997 case isolates (71.1 percent of the culture-confirmed cases) and 264 (26.5 percent) of these had fewer than six IS*6110* bands. Spoligotype patterns were generated for each isolate with fewer than six IS*6110* bands as well as for any isolate with greater than six bands whose IS*6110* fingerprint differed from another by one band.

Using the conventional cluster definition, 551 of 997 (55.3 percent) cases were clustered with at least one other case in the population, resulting in 115 clusters. Thirty-four (30 percent) of these clusters were defined on the basis of identical IS*6110* fingerprints with fewer than six bands along with identical spoligotypes, while 81 were defined on the basis of IS*6110* patterns with six or more bands that were identical or which differed by one band but had identical spoligotype patterns. The size of the clusters ranged from two isolates, accounting for 57 (49.6 percent) clusters to, 35 isolates (one cluster). The time-span of individual clusters ranged from one year to the full eight years of the study period. A Kaplan-Meier analysis of the distribution of the time between diagnosis dates of matching isolates found that, among all isolates for which another isolate a with a matching fingerprint was identified during the remaining study period, 73.6 percent (312 of 424) were identified within one year, while 87.0 percent (369 of 424) were identified within 2 years. Of 551 isolates, 127 matched at least one other isolate in the study population, but were the last case in the cluster during the study period, and therefore were not included in the Kaplan-Meier analysis.

Of the 551 isolates clustered by conventional methods, 312 matched the isolate of a case that was diagnosed within the same one year period, and were thus considered clustered by our time-restricted cluster definition.

Epidemiological and clinical characteristics of clustered cases

Cases considered clustered by IS*6110* and spoligotype that were not considered clustered using the one year restriction were significantly less likely to have cavitary disease and more likely to be in older age categories (with a two-sided p-value < 0.05 ) than were cases that were considered clustered using the one year restriction (Table 2.1). Additionally, these cases were less likely to be homeless, heavy alcohol users, HIV positive, and sputum-smear positive than were cases that were considered clustered using the one year restriction, although none of these associations were significant at the p < 0.05 level.

Incidence trends

The incidence of culture confirmed TB in Arkansas declined by 2.7 cases per 100,000 between 1997 and 2003 (from 5.9 cases per 100,000 in 1997 to 3.2 cases per 100,000 in 2003, Figure 2.1). This overall decline resulted from a decline in both clustered cases, which declined by 1.0 case per 100,000 (from 2.0 cases per 100,000 in 1997 to 1.0 cases per 100,000 in 2003), and unique cases, which declined by 1.7 cases per 100,000 (from 3.9 to 2.2 cases per 100,000).

Stratification by age showed the largest decline in the incidence of TB in those aged 65 and older, with the overall incidence of culture-confirmed TB in this age group declining from 19.9 cases per 100,000 in 1997 to 8.5 cases per 100,000 in 2003, an absolute decline of 11.4 cases per 100,000 (Figure 2.2a). In this age group, the absolute decline of unique cases over this period was 1.8 times the absolute decline of the clustered cases (unique cases declined from 14.5 per 100,000 in 1997 to 7.1 cases in 2003, while clustered cases declined from 5.4 cases per 100,000 in 1997 to 1.3 cases per 100,000 in 2003). In the 20 to 64 age group, the overall incidence of culture-confirmed TB declined by 1.5 cases per 100,000 (from 5.0 cases per 100,000 in 1997 to

3.5 cases per 100,000 in 2003). In this age group, the absolute decline of unique and clustered cases was identical at 1.0 cases per 100,000 in each: unique cases declined from 3.0 to 2.0 cases per 100,000 and clustered cases declined from 2.4 to 1.4 cases per 100,000 (Figure 2.2b).

The incidence of culture-confirmed TB was consistently higher in blacks than in whites in Arkansas (Figure 2.3a, 2.3b). The magnitude of the decline in the overall incidence of culture-confirmed TB was greater in blacks than whites: in blacks, the incidence declined from 13.8 to 6.5 cases per 100,000 (a difference of 7.3 cases per 100,000) over the period 1997 to 2003, while the incidence in whites declined from 4.0 cases per 100,000 to 2.4 cases per 100,000 (1.6 cases per 100,000). The decline in the incidence of TB in blacks was dominated by a decline in the incidence of unique cases, with the decline in unique cases 3.6 times the decline in clustered cases (unique cases declined from 8.4 to 3.0 cases per 100,000, while clustered cases declined from 5.3 to 3.5 cases per 100,000). In contrast, in whites, the magnitude of the decline in clustered cases (from 1.3 to 0.5 cases per 100,000) was the same as the decline in unique cases (from 2.7 to 1.9 cases per 100,000).

The incidence of culture-confirmed TB was higher among the rural than the urban population for each year except 2001 (Figure 2.4 a, b). Between 1997 and 2003, the overall incidence of culture-confirmed TB declined by 3.0 cases per 100,000 in rural counties ( from 6.9 to 3.9 cases per 100,000 ) and 2.5 cases per 100,000 in urban counties ( from 5.0 to 2.5 cases per 100,000 ).

In rural counties, the decline in unique cases was 2.0 times the decline in clustered cases (a decline of 2.0 cases per 100,000 in unique cases and of 1.0 per 100,000 in clustered cases). In urban counties, the decline in unique cases was 1.4 times the decline in clustered cases (a decline of 1.4 cases per 100,000 in unique cases and 1.0 per 100,000 in clustered cases).

## Discussion

Incident TB cases may result from an infection with *M. tuberculosis* acquired recently, or from the reactivation of a latent infection acquired in the remote past. Distinguishing clinically between these types of disease is frequently not possible, although they have different implications for TB control programs. We used molecular genotyping data to estimate the relative contribution of recent and remotely acquired infection to the yearly incidence of TB in Arkansas. This analysis indicates that decline in the incidence of TB in Arkansas between 1997 and 2003 was most likely due to a decline in the incidence of reactivated latent infections in individuals over age 65. This decline was primarily seen in blacks, and was most prominent in rural areas of the state.

The steep decline in reactivation cases in the oldest age category (age 65 and older), and the minimal decline in reactivation cases in those aged 20-64, suggests that a cohort effect may be responsible for the decreased incidence of TB in Arkansas. With a steep decline in the incidence of TB in Arkansas (and the United States) in the first half of the 20th century, each successive birth cohort was exposed to a lower risk of infection with *M. tuberculosis.* As a result, the prevalence of latent TB infection is likely highest in the oldest individuals in the population, and decreases with decreasing age. The declining incidence of TB in the population may reflect earlier birth cohorts leaving the population as more recent birth cohorts enter it. This is similar to what has been reported in the Netherlands [69], a population that also experienced a steep decline in the incidence of TB early in the 20th century.

The marked decline in the incidence of TB in non-Hispanic blacks is a welcome finding. In the US, the burden of disease has fallen disproportionately on this group, with an incidence of TB eight times higher than that seen in non-Hispanic whites in 2002 [100]. However, this decline does not appear to be the result of reduced transmission in this population – the incidence of clustered cases declined minimally over the study period. Rather, a decline in the incidence of reactivated latent infections is driving this trend. While the numbers in our study were too low to stratify race/ethnic groups by age, the race-stratified yearly incidences of unique cases suggest that the cohort effect is predominately seen in the non-Hispanic black population, as the incidence of reactivated TB in non-Hispanic blacks declined steeply, while declining minimally non-Hispanic whites.

Molecular epidemiologic investigations have consistently identified non-Hispanic black race as a risk factor for molecular clustering [43, 91, 113]. It has been suggested that higher frequencies of excessive alcohol use, drug use, incarceration, and infection with HIV, all identified risk factors for clustering, may be responsible for the higher rates of TB seen among non-Hispanic blacks than non-Hispanic whites [113]. While not discounting the importance of known risk factors, our results suggest that, in Arkansas, higher rates of TB currently observed in non-Hispanic blacks are as much the result of historic trends as of contemporary risk behaviors. Historically, the incidence of TB has been higher in blacks than whites, both in Arkansas and United States. Because active TB can result from infection acquired many years or even decades in the past, the legacy of historic TB transmission may be felt in a population for generations. Similarly, reductions in transmission will continue to have effects on overall disease incidence for many years. A concerted and continuous program to treat recently infected close contacts of newly diagnosed cases, as well as other high risk individuals having latent infection, has been in place in Arkansas for more than three decades. The observed decline in the incidence of TB in non-Hispanic blacks appears to be the result of reductions in transmission and the treatment of latent infections over several decades. An additional factor working to reduce the risk for reactivation among non-Hispanic blacks could be an improvement in the overall health status of this population.

Arkansas differs from the rest of the United States, particularly from the urban areas in which molecular typing for TB has been validated as a measure for recent transmission. While the majority of the US population lives in urban areas, only a minority of the population of Arkansas does. The pattern of TB incidence also differs. In urban areas of the United States, the incidence of TB was 24 percent higher than the national rate in 2003 [99], while in the same year in Arkansas, the incidence of TB was 23 percent higher in the rural areas than the state overall. As the bulk of Arkansas' population resides in rural counties, the decline in the incidence of TB in these counties between 1997 and 2003 appears to be a key driver of the overall decline in TB in Arkansas.

Our results differ from those of a previous investigation conducted in San Francisco [67], in which the investigators concluded that the decline in the incidence of TB in that population between 1991 and 1997 was due primarily to a decline in active transmission in that city. It is not surprising that a similar analysis of two very different populations finds divergent results: the transmission dynamics between the rural, highly stable population of Arkansas are likely quite different from a diverse urban population like San Francisco. The molecular typing methods we used differed slightly from those used by Jasmer and colleagues: while we used spoligotyping as a secondary typing method, they used pTBN12 typing. Spoligotyping is somewhat less discriminatory than pTBN12 typing [114], therefore this difference could influence our cluster classifications, resulting in a higher estimate in the amount of clustering, particularly among isolates with fewer than six IS*6110* bands. However, it is unlikely that this methodological difference resulted in the distinct time-trends identified in our respective study populations.

The use of "clustering" of cases that exhibit identical or similar DNA fingerprint patterns as a measure of recent transmission has been used most widely in diverse, urban populations: of the limited studies that attempt to validate the approach against epidemiologic data, the great majority have been conducted in these same populations [42, 78]. However, the predictive value of clustering for recent transmission may vary across diverse populations. Evaluations of molecular clustering in rural, stable populations suggest that, in these settings, clusters of cases sharing identical or highly similar fingerprints may include cases for whom epidemiologic evidence of recent transmission between other cases in the cluster cannot be found [70, 107], or among whom epidemiologic evidence suggests a transmission event that occurred in the remote past [70]. Additionally, a contact investigation combined with molecular fingerprinting analysis in Arkansas [115] showed evidence of a single strain of *M. tuberculosis* persisting in a rural area of the state over many decades. Repetitive cycles of transmission and infection with a single strain, followed by reactivation or progression to disease, could result in a cluster of cases linked in part, but not entirely, by recent transmission. These studies suggest that, in using molecular clustering as a measure of recent transmission in Arkansas, some amount of misclassification is inevitable. By restricting our definition of clustering to cases with matching fingerprints that were diagnosed within a year of one another, we attempted to increase the specificity of this measure for recent transmission. Indeed, a comparison of known risk factors for recent transmission found that each factor assessed was more commonly found among clustered cases diagnosed within a year of one another than among cases clustered by fingerprint alone that were diagnosed more than one year apart. However, in the absence of epidemiological contact tracing data to verify this, we cannot conclude that our restricted measure is more specific for recent transmission.

Because of the uncertainty regarding molecular typing in this population, we must interpret our results with caution. However, we have no reason to believe that misclassification will differ between years in our study. Therefore, we are confident in using these tools to follow trends over time. Particularly among the oldest populations is our study, the signal is very strong, which leads us to believe that, while not a perfect measure, our current definition of molecular clustering is allowing us to detect important factors driving trends in TB in Arkansas.

Effective TB control programs are essential to reaching the goal of TB elimination in the US. Our results suggest that, while the overall incidence of TB declined in Arkansas between 1997 and 2003, the decline in the active transmission of *M. tuberculosis* infection was not as important as the decrease in the reactivation rate. Both rates declined, the former as a result of prompt and highly effective treatment of newly identified cases and the latter as a result of effective treatment of latent infection. In Arkansas and other rural, low-incidence populations, improvements in the effective identification and interruption of active pockets of transmission will be essential to reaching TB elimination goals

Tables

**Table 2.1.** Table 2.1. Known risk factors for TB transmission by clustering definition. Comparison of the frequency of previously identified risk factors for clustering among 1) all TB cases identified as clustered, 2) cases clustered within a 1 year time interval, and 3) cases clustered but with more than 1 year between clustered cases, Arkansas, 1997 to 2003.

| Risk factor | All clustered (%) | | Clustered within 1 year (%) | | Clustered outside 1 year (%) | | OR | 95 % CI |
|---|---|---|---|---|---|---|---|---|
| Sputum smear + | 219/550 | (40.0) | 130/312 | (41.7) | 89/238 | (37.4) | .836 | 0.59, 1.18 |
| **Cavitary disease** | **178/490** | **(36.3)** | **114/278** | **(41.0)** | **64/212** | **(30.2)** | **.62** | **0.42, 0.90** |
| HIV+ | 25/349 | (7.2) | 18/212 | (8.5) | 7/137 | (5.1) | .58 | 0.23, 1.42 |
| Homeless | 23/546 | (4.2) | 15/310 | (4.8) | 8/236 | (3.4) | .69 | 0.29, 1.66 |
| Excessive Alcohol Use | 104/532 | (19.6) | 64/302 | (21.2) | 40/230 | (17.4) | .78 | 0.51, 1,21 |
| Injection Drug Use | 2/363 | (0.6) | 2/0 | (1.0) | 0/159 | (0.0) | na | |
| Male | 368/551 | (66.8) | 211/312 | (67.6) | 157/239 | (65.7) | 1.09 | 0.76, 1.56 |
| **Age >65** | **200/550** | **(36.4)** | **93/311** | **(29.9)** | **107/239** | **(44.8)** | **1.9** | **1.33, 2.70** |

Odds ratios and 95% confidence intervals compare risk factor frequencies between cases clustered within one year and cases clustered but with more than 1 year between clustered cases.

# Figures

**Figure 2.1.** Annual incidence of all culture confirmed, unique (non-clustered), and clustered TB cases in Arkansas, 1997 to 2003. Clustering based on the time-restricted definition.

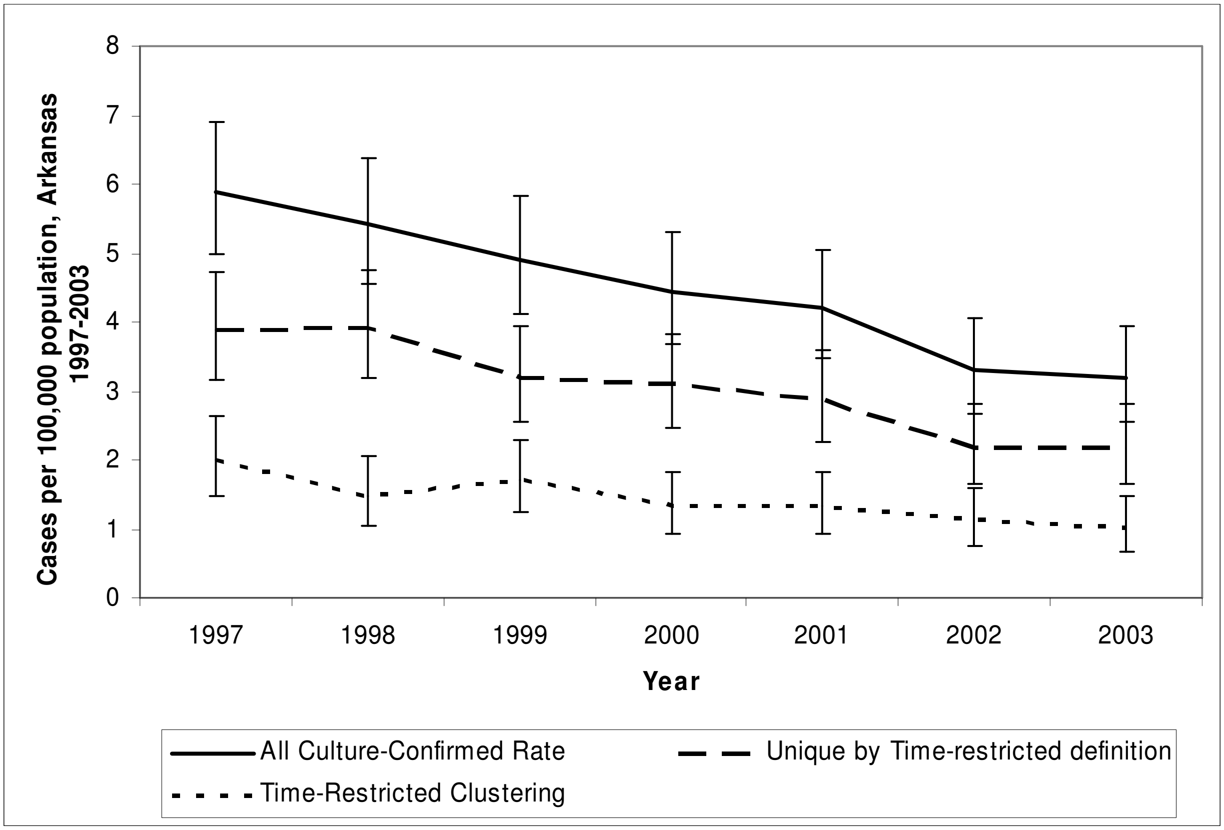


**Figure 2.2.** Annual incidence of culture confirmed, unique (non-clustered), and clustered TB cases by age group, Arkansas 1997 to 2003. Clustering based on the time-restricted definition. Panel a, Age 20-64; panel b, Age 65+

Figure 2.2, panel a.

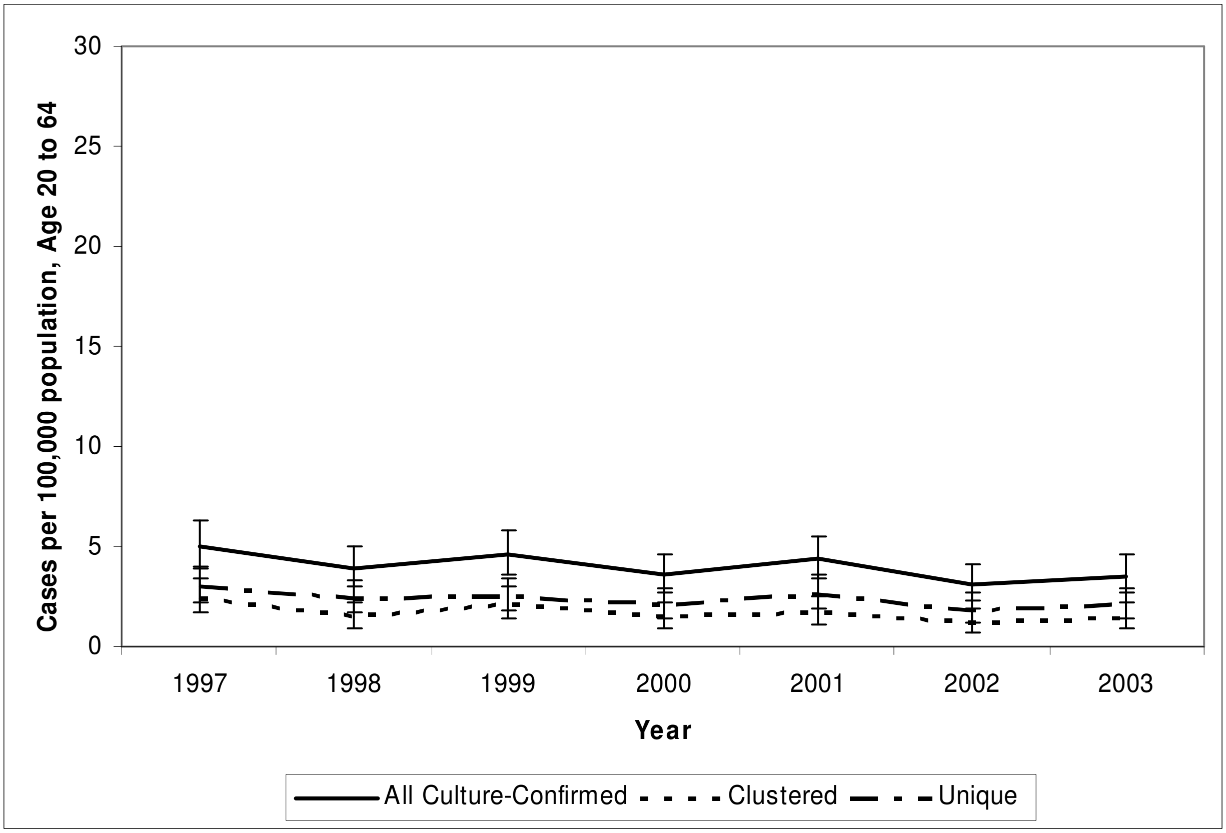


Figure 2.2, panel b.

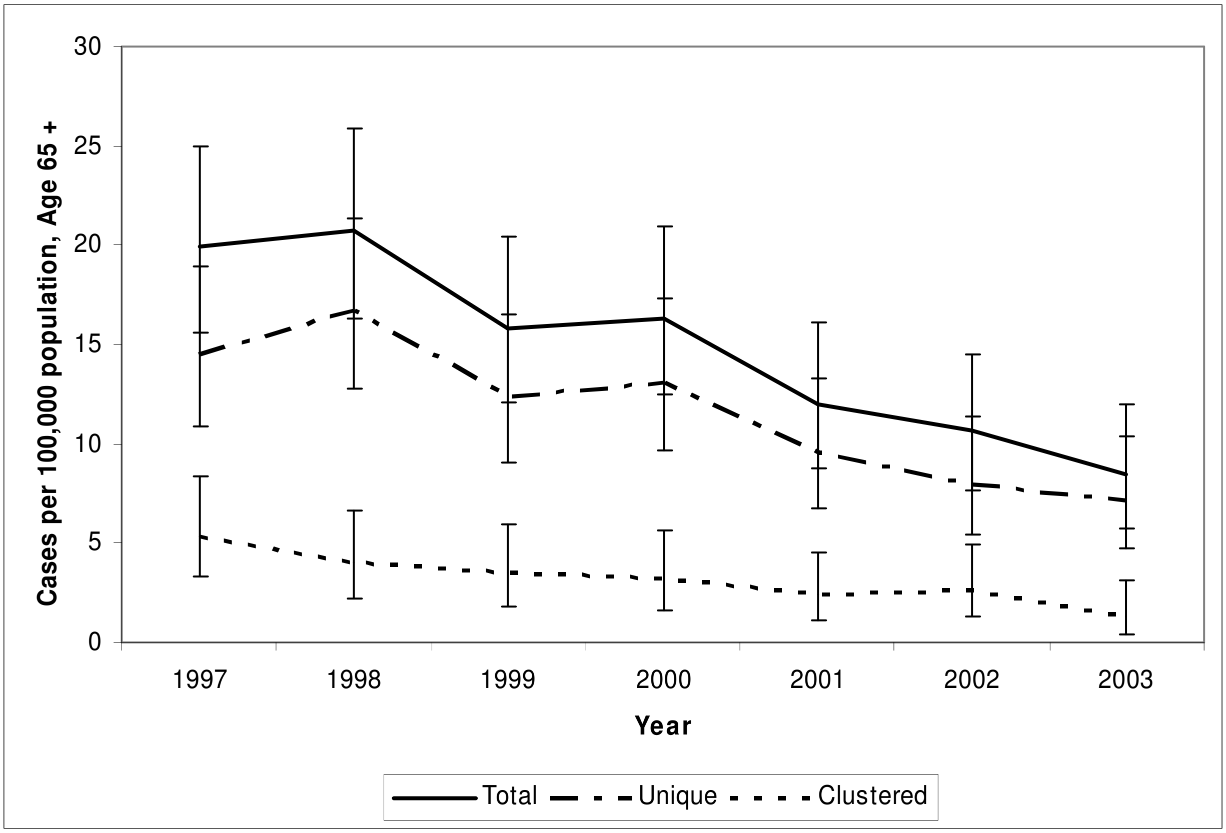


**Figure 2.3.** Annual incidence of culture confirmed, unique (non-clustered), and clustered cases of TB in non-Hispanic whites (panel a) and non-Hispanic blacks (panel b), Arkansas, 1997 to 2003. Clustering based on the time-restricted definition.

Figure 2.3, panel a.

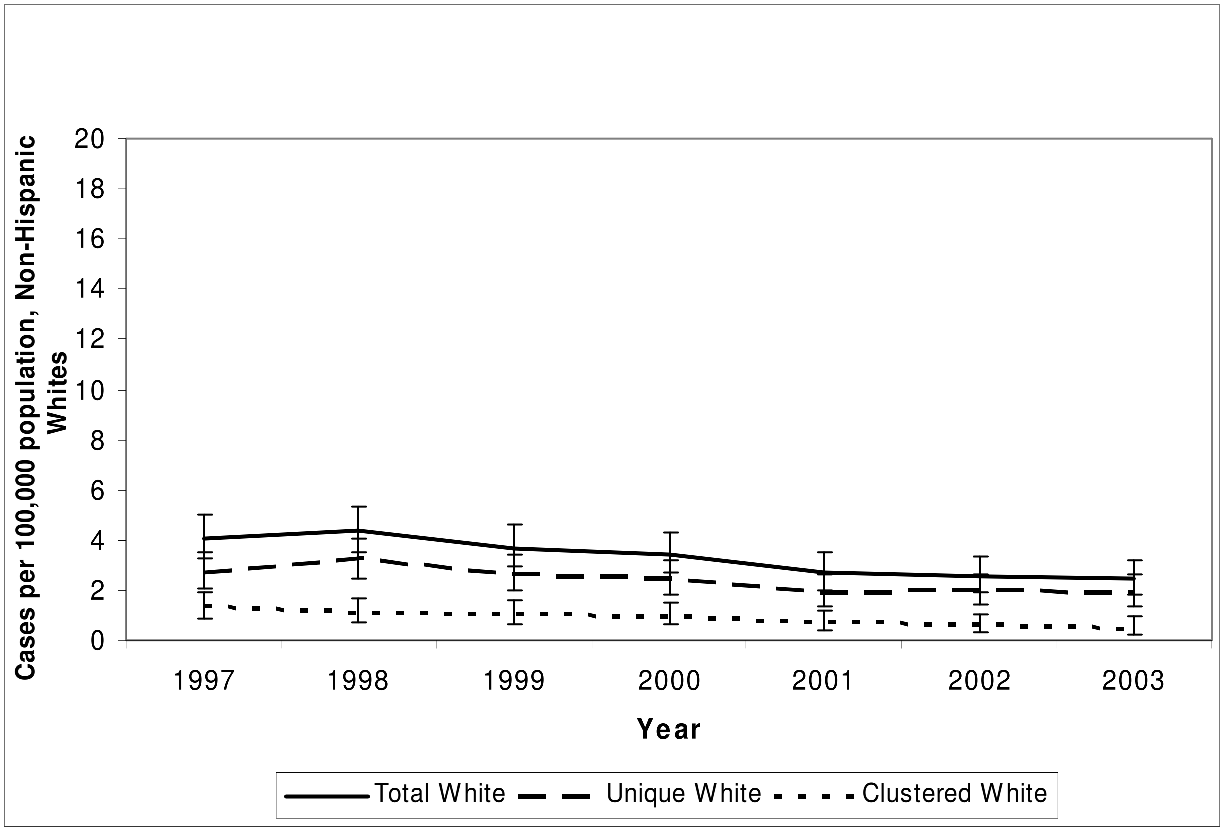


Figure 2.3, panel b.

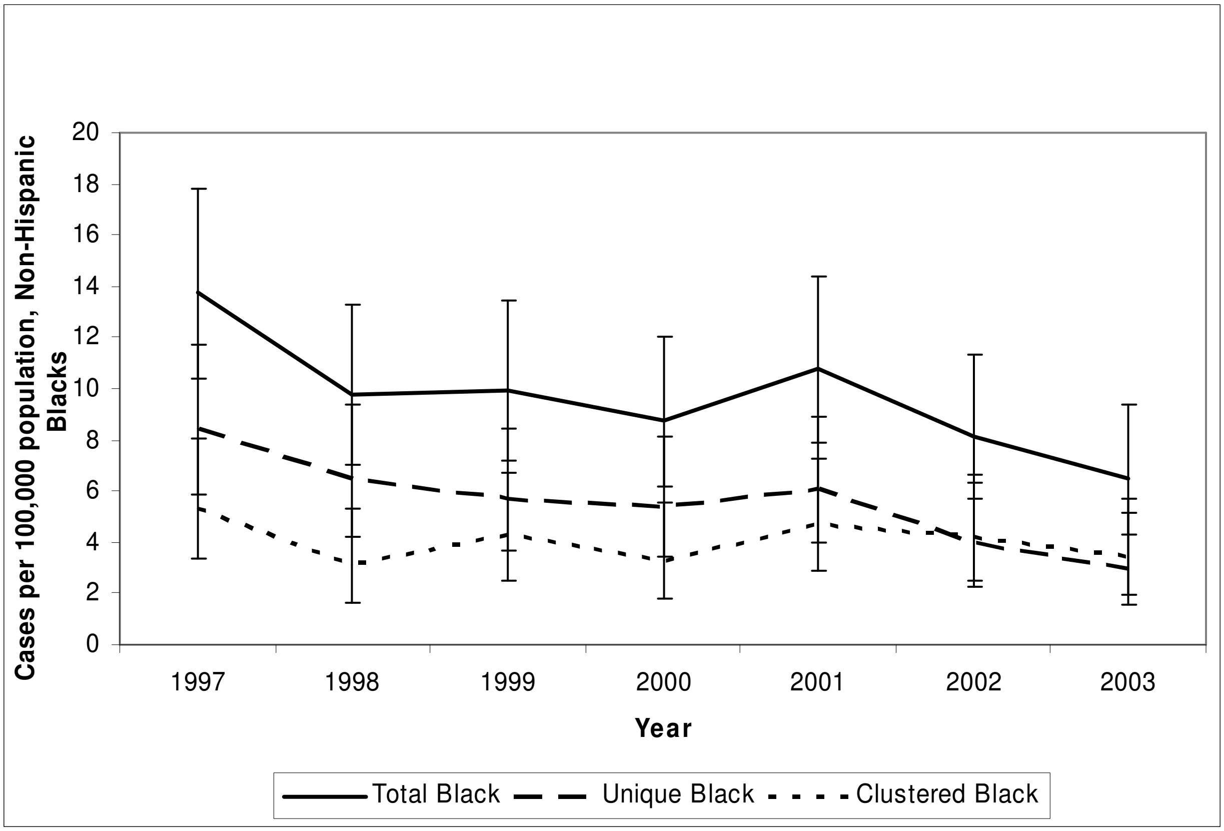


**Figure 2.4.** Annual incidence of culture confirmed, unique (non-clustered), and clustered cases of TB in rural counties (panel a) and urban counties (panel b), Arkansas, 1997 to 2003.

Figure 2.4, panel a.

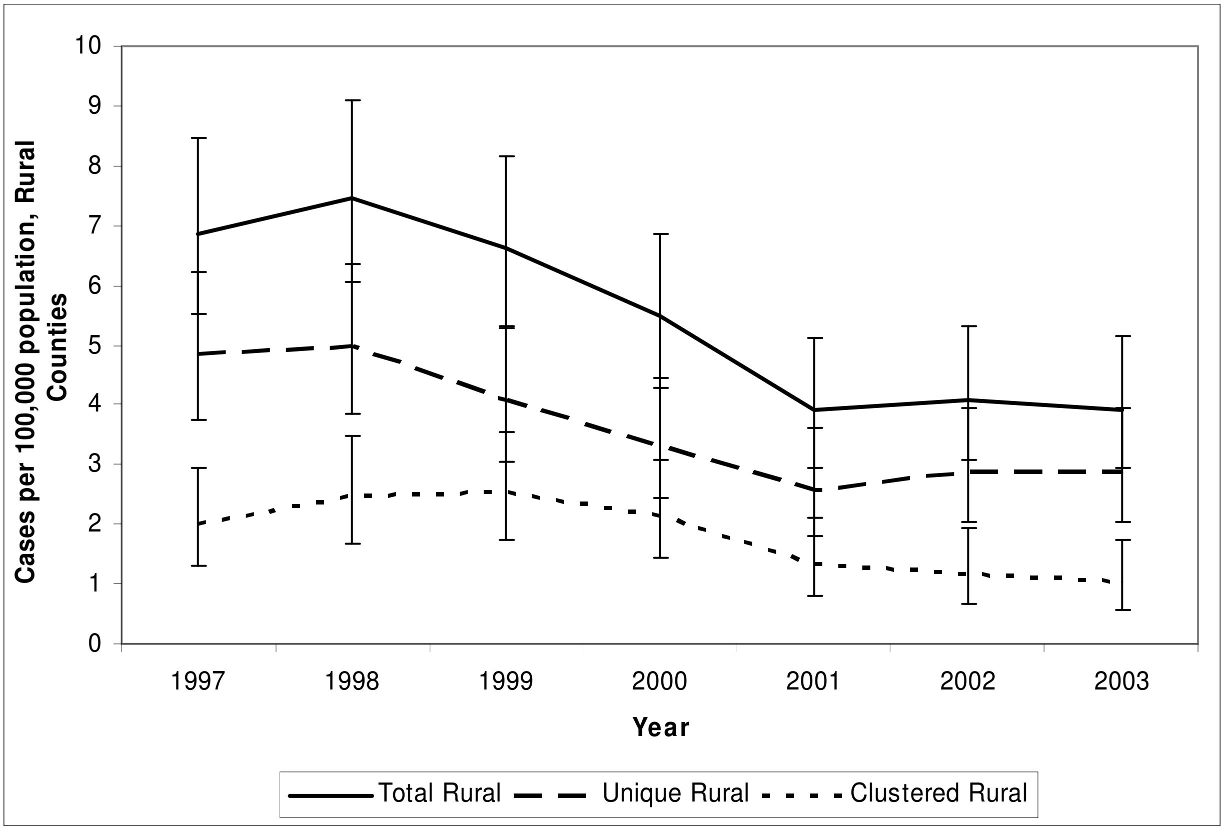
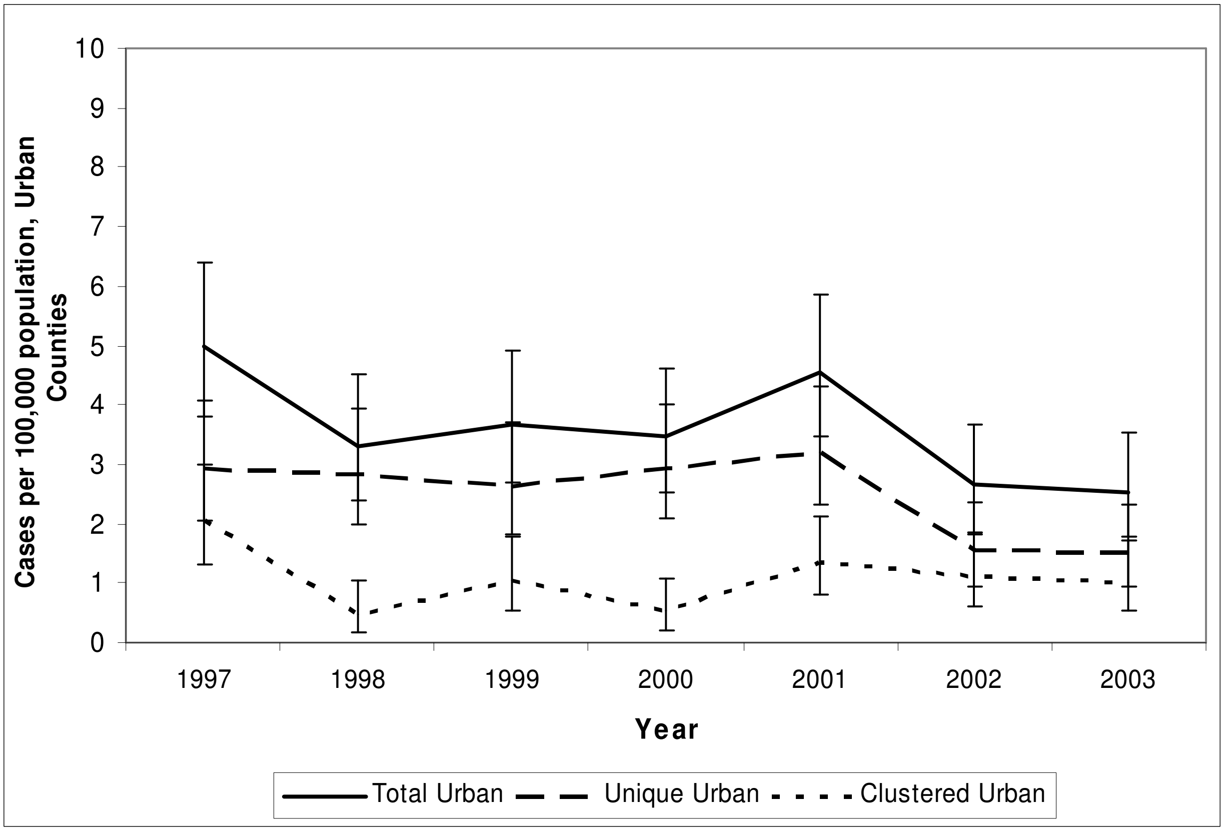


Figure 2.4, panel b.



# Chapter III

## Microbial and host predictors of tuberculosis case clustering: implications for the design and interpretation of population-based molecular typing programs

### Introduction

Molecular typing has become an essential tool in the study of *M. tuberculosis*, both in basic research and in applied public health investigations. When integrated into routine, population-based tuberculosis (TB) surveillance, molecular typing presents a powerful tool to identify patterns of infection: by allowing cases related by a recent transmission event to be identified, the spread of drug-resistant clones can be tracked [116], the proportion of disease due to recent transmission can be estimated [72, 117, 118], and, of particular importance to TB Control programs, outbreaks can be rapidly identified [118, 119]. Despite the increasing reliance on molecular typing methods to identify TB cases linked by recent transmission, however, the relationship between molecular typing data and transmission relationships remains poorly understood.

Numerous reports from population-based typing programs have documented “clusters” of cases related by molecular typing pattern that show no epidemiologic linkage between cases [70, 71]. In some cases, further investigation has shown such “unexplained clusters” to consist of truly linked cases, where case connections had been missed by conventional contact tracing methods [57]. Such clusters point to transmission occurring beneath the radar of traditional TB control programs, such as in marginalized populations where patients may mistrust public health authorities or who may not know the names of many contacts [120, 121], in a context not recognized by the traditional contact tracing protocol, or through casual contact [41, 56]. Often, however, further investigation finds that “unexplained clusters” do not reflect recent transmission events: transmission might have occurred many years or even decades in the past [70], or the particular molecular typing tool used might provide insufficient resolution to discriminate unlinked cases in a given population [73].

Distinguishing molecular case-clustering that reflects true transmission linkages (referred to here as “true clustering”) from erroneous linkages (“false clustering”) is critical to the effective use of molecular typing to direct limited TB control resources. The probability that a clustered case truly reflects recent transmission (measured by the predictive value positive (PPV)) may be influenced by a complicated array of factors, including characteristics of the host population [55, 70], the local pathogen population [90, 92, 122], the molecular typing tool and clustering definition used [35, 75], and the time period over which cases are evaluated [79]. To date, however, evaluations of molecular typing approaches have most often focused either solely on microbial factors [123-125], or solely on host population factors [55, 70], and none have addressed the interacting effects of these factors.

Evolution of the *M. tuberculosis* genome is largely clonal [126] [127], and distinct evolutionary lineages are highly associated with particular geographic regions [128-132]. In a given geographic region, the composition of the infecting *M. tuberculosis* strain population may reflect both contemporary and historically distant migration patterns of the human host population [127, 133]. In many regions TB transmission appears to be dominated by regionally endemic *M. tuberculosis* strains [90, 92, 134], with little genetic variation between *M. tuberculosis* case isolates, even those clearly unrelated by transmission. Reports from host populations that have experienced continuous immigration and emigration show relatively high levels of *M. tuberculosis* “background” diversity (the diversity amongst isolates thought to be unrelated by transmission) [122], while geographically isolated populations have been characterized by highly homogeneous *M. tuberculosis* strain populations [135].

In addition, diversity in the molecular markers that are exploited by molecular typing methods (including IS*6110* RFLP, spoligotyping, and MIRU) varies by evolutionary lineage [125, 136]. For example, using spoligotyping, which assays variation in the Direct Repeat (DR) locus, molecular markers that show high diversity among non-Beijing lineages show little diversity among isolates of the Beijing family [92, 124, 137], with the result that typing methods based on these markers cannot discriminate between unrelated Beijing-family isolates.

In the United States (US), where the human population has diverse ancestral origins, a higher diversity of *M. tuberculosis* spoligotype families is seen than that seen in geographic regions with human populations of more homogeneous ancestry [132, 138, 139]. The diversity is not uniform: both historic and contemporary migration patterns vary widely across different populations in the US, particularly between urban and rural areas. Reports from urban populations with high levels of immigration, such as San Francisco, demonstrate high *M. tuberculosis* strain diversity [127]. By contrast, reports from rural populations in both the US and Canada (demographically similar populations which both experience low rates of TB), show levels of clustering substantially higher than would be expected based on the low incidence of active TB disease [90, 91, 134]. Regionally endemic strains appear to dominate transmission in some rural populations [90, 91], and epidemiologic investigations suggest that clustering in rural populations is less likely to reflect recent transmission than it is in urban populations [70, 90]. The low positive predictive value (PPV) of clustering in rural areas may be due to low background diversity in the circulating population of *M. tuberculosis*. With current information, however, it is not possible to distinguish the effects of *M. tuberculosis* population structure from effects of the host population demographics or the molecular typing method used.

We investigated the impact of host behavioral, clinical, and demographic factors, as well as pathogen population characteristics and molecular typing method, on the predictive value of molecular typing in Arkansas by reviewing extensive epidemiologic interviews and genetic typing data collected over the five year period from 1996 to 2000. Arkansas is, a southern US state characterized by a rural, historically stable population. Arkansas has a higher than expected level of case clustering and low PPV of clustering for recent transmission, suggesting patterns of transmission very different than those described in large urban populations in the US [35, 70, 115].

## Methods

Arkansas demographic characterization

We obtained Arkansas population and demographic data from the 2000 US Census (Census 2000), and state TB and HIV/AIDS and Centers for Disease Control and Prevention Surveillance Reports (CDC HIV 2000, TB 2000). Census Bureau Metropolitan Statistical Areas (MSAs) defined by Arkansas' Office of Management and Budget were used to determine urban areas; non-MSAs were considered to be rural areas. The MSAs identified included Fayetteville-Springdale-Rogers, Fort Smith, Jonesboro, Little Rock-NorthLittle Rock, Memphis, Pine Bluff, and Texarkana. All counties within these MSAs were considered urban, while all counties outside of these MSAs were considered rural.

### Study sample description and demographic characterization

Study patients were persons with an incident case of culture-confirmed TB reported in Arkansas between 1996 and 2000. If multiple isolates were collected for any single patient, only the first isolate was included in our analysis.

Arkansas reported 976 incident cases of TB between January 1, 1996, and December 31, 2000 (CDC TB Surveillance Reports). Of these, 721 were culture confirmed, and molecular typing results were generated for 705 (97.8%). Five patients suffered a relapse of active disease within this study period: for each of these patients, only the first episode of disease was considered in our analysis. Each isolate for all relapse cases was genotyped: in each case, the episode of relapse disease produced an isolate with an identical pattern to the initial disease isolate. Therefore, restricting our analysis to only the first reported case in the study period, for a total of 971 cases and 700 genotyped isolates in the study period, had no impact on the designation of clusters.

Four hundred and eleven (42.3%) of cases lived in one of seven MSAs at the time of their diagnosis. Sixty-two percent of cases were males, and 44% were age 65 or older at diagnosis. Non-Hispanic whites made up 49% of cases, non-Hispanic blacks, 36%, Hispanics, 7.8%, Asian Pacific Islanders, 6.4%, and American Indian/Alaska Natives, 0.5%. Eleven percent of cases were foreign-born, and 6.3% were infected with a drug-resistant isolate. Thirty-three (3.4%) of patients had had a prior diagnosis of TB disease.

This study was approved by the health sciences institutional review boards of the University of Michigan and the University of Arkansas for Medical Sciences.

### Genotyping

*IS6110* restriction fragment length polymorphism (RFLP) patterns were determined using standard procedures, as previously described [32], for 700 of all 716 (97.7%) culture-confirmed cases (excluding the five relapse cases previously noted). For all 196 isolates with fewer than six *IS6110* bands (28 % of those typed) and for isolates with six or more *IS6110* bands that differed from another *IS6110* pattern by only one band, pTBN12 patterns were also generated, following a standard protocol [140]. Spoligotyping patterns were generated for 697 (99.6%), following standard protocols [33].

Compared to the 700 genotyped cases, the 17 cases that were culture confirmed but not genotyped were significantly more likely to be diagnosed with extra-pulmonary disease, (Odds Ratio 4.42, 95% CI 1.61, 12.15), and to have been born outside of the United States (Odds Ratio 3.63, 95% CI 1.15, 11.43).

### Cluster definition

We evaluated clustering using *IS6110* RFLP alone, spoligotype alone, *IS6110* RFLP with secondary typing by pTBN12 RFLP (IS*6110*-pTBN12), and *IS6110* RFLP with secondary typing by spoligotype (IS*6110*-spoligotype). For both *IS6110* RFLP alone and spoligotype alone, a cluster was defined as two or more cases with an identical typing pattern. For both methods combining *IS6110* with a secondary typing technique, separate cluster definitions were used for isolates with five or fewer *IS6110* RFLP bands (low band isolates), and isolates with 6 or more *IS6110* RFLP bands (high band isolates). For low band isolates, a cluster was defined as a cluster of two or more cases with isolates identical by both *IS6110* RFLP and the secondary typing method (pTBN12 or spoligotype). For high band isolates, a cluster was defined as two or more cases with identical *IS6110* RFLP patterns, or cases with *IS6110* RFLP patterns differing by a single band, but identical by the secondary typing method.

### Investigation of epidemiologic linkages among clustered cases.

As a sentinel surveillance site in the CDC's National Tuberculosis Genotyping and Surveillance Network, the state of Arkansas conducted a population-based sentinel study of TB which included all incident culture-positive TB patients in the state from 1996- 2000. Active case finding relied on local mycobacteriology and hospital infection control records for all facilities in the state, hospital IDC-9 discharge codes for TB, pharmacy records for prescriptions of a combination of two or more anti-TB drugs, coroners' records that showed TB as a diagnosis, and AIDS surveillance reports that indicated a diagnosis of TB[141]. For each culture positive TB patient, a Report of a Verified Case of Tuberculosis was completed, providing standard demographic information. Race/ethnicity was based on self-report.

Detailed data on patient demographics, social history, clinical characteristics, and risk factors related to TB transmission and disease were obtained by review of medical and public health records for all culture-confirmed TB cases for which an isolate was genotyped and found to be clustered with that of another case in the study period using *IS6110*- pTBN12, or with *IS6110* RFLP with spoligotype as a secondary method. In addition, notes from interviews of clustered cases were reviewed to identify epidemiological contacts between clustered cases. These interviews, conducted by ADH TB control employees, followed an extensive standardized questionnaire which included questions on patients' demographic characteristics, history of current and previous residence and employment, history of previous TB and tuberculin skin test (TST) results, dates of symptom onset and diagnosis, hospitalizations, completion of treatment, laboratory results, chest radiograph findings, and known TB risk factors (including substance abuse, alcohol abuse, and stays in long term facilities such as homeless shelters, detoxification centers, or correctional facilities, travel history, types and sites of social and leisure activities, underlying illness, and concurrent immunosuppressive conditions or treatment). In cases for which extensive interviews were not conducted, medical records and case contact tracing records were reviewed to identify epidemiological contacts.

Of 330 cases with an isolate clustered by *IS6110*- pTBN12, 232 (70.3%) had an interview available for review, 64 (19.4%) had medical and contact tracing records available, but no interview, and 34 (10.3%) had no records available. Of the 34 cases with no records available, sufficient information was provided in the records of other cases in the same cluster to establish epidemiologic linkages for 28 (82.4%). Insufficient information was available to classify 8 cases, 6 of which had no records available, and 2 of which were interviewed, but were in a cluster with missing information for all other cases in the cluster. In total, epidemiologic linkage status was classified for 322 (97.6%) of 330 clustered cases.

Cases for whom a record review was conducted in place of an interview were more likely to be foreign born (Odds Ratio = 4.6, 95%CI 1.49, 14.29), to reside in a MSA at the time of diagnosis (Odds Ratio = 1.94, 95%CI 1.11, 3.45), to receive treatment for their TB from a private provider rather than the health department (Odds Ratio= 5.26, 95% CI 1.96, 14.29), and to have a drug-susceptible isolate (Odds Ratio 3.4, 95% CI 1.31, 9.59). There were no significant demographic or clinical differences between cases for which no record was found and cases for which either an interview or a record review was available.

### Epidemiological linkage classification

Based on available information, each case that was part of a cluster was classified as having a definite epidemiologic link, a probable epidemiologic link, or having no identified link to, any other case in the same cluster.

A **definite epidemiologic link** was defined as one between patients within a cluster who lived in the same household or shared the same indoor airspace when at least one of the patients was judged to be infectious [35].
A **probable epidemiologic link** was defined as a link that did not fit the criteria for definite epidemiologic link, but for which interview suggests either a direct exposure of one infectious patient to another in the cluster, or a circumstance whereby cluster patients were in the same location at the same time [70].

Information was abstracted into a Microsoft Access database (Microsoft Corp, Redmond, WA).

### Spoligotype family assignment

Spoligotype-defined families are well described [132], and individual families often show strong phylogeographic associations. Because these family classifications can be used to characterize the population genetics of local *M. tuberculosis* isolates, we classified the spoligotype family for each isolate in our study sample. Spoligotype patterns were analyzed with 'Spotclust' [142], which implements a mixture model built on the SpolDB3 database. This model takes into account knowledge of the evolution of the DR region and assigns spoligotype patterns to families and subfamilies.

### Statistical Analysis

We compared the distribution of demographic and clinical characteristics among patients with clustered and unique isolates, and epidemiologically-linked to non-linked clustered cases, using the $\chi^2$ test or Fisher's exact test, as appropriate. Unless otherwise noted, clustering designations included in all statistical analysis were based on *IS6110*- pTBN12 typing. Predictive value positives (PPVs) were calculated to determine the probability that, given inclusion in a cluster, a definite or probable epidemiologic linkage was identified between a given case and at least one other case in that cluster.

The diversity of molecular types was estimated using the Hunter-Gaston Discrimination Index (HGI) (also referred to as Simpson's Index of diversity) [143], calculated by the following formula:

$$D = 1 - \left[ \frac{1}{N(N-1)} \sum_{j=1}^{s} \mathrm{n}_{\mathrm{j}}(n_j - 1) \right]$$

where D is the numerical index of discrimination, N is the total number of strains in the typing scheme, s is the total number of different strain types, and $n_j$ is the number of strains belonging to the jth type. D can be interpreted as the probability that any two isolates drawn at random will be of different molecular types. For the purposes of presentation, we used 1-D, which can be interpreted as the probability that any two isolates drawn at random will exhibit the same molecular type.

We used multiple logistic regression analysis to assess the importance of demographic, clinical, and strain characteristics in predicting false clustering among cases with no known relationship to other cases in the sample. For this model, our study sample was restricted to unique isolates and clustered cases for which no linkage was found. A stepwise logistic regression with forward selection was performed. Independent variables with a probability of < .10 by the Wald statistic, after adjusting for other variables in the model, were included in the model. Independent variables included in the stepwise model were then re-run in a logistic model which included age and race to adjust for confounding. Gender was not included in this model as it was not associated with either clustering or the identification of epidemiologic linkages in our study sample. Because spoligotype-defined strain families varied in the number and diversity of IS*6110* RFLP bands, and because band number is strongly correlated with the validity of IS*6110* RFLP clustering measures, we included a binary low band number/high band number variable in all models that included strain family as a predictor.

All statistical analysis was conducted using SAS version 9.1.3 (SAS Institute Inc. SAS. Version 9.1.3 Cary, NC: SAS Institute Inc, 2006)

# Results

## Arkansas Demographics

In 2000, the population of Arkansas was 2,673,400, with 77.8% of the population self-described as Non-Hispanic White, 15.7% Non-Hispanic Black, and 4.7% Hispanic, 1.1% Asian, and 0.7 % American Indian/Alaska Native. The foreign-born population in Arkansas is relatively small: in 2000, 2.8% compared to 11.1% of the US population. The Arkansas population is somewhat more stable than the overall US population: among US-Natives, 63.9% of Arkansas residents in 2000 were born in Arkansas, while 60% of the US population was born in the same state in which they currently reside. Arkansas has a largely rural population, with 49.5 percent of the population residing in a county falling within an MSA, while 50.5 percent lived in non-MSA counties (Census 2000).

The proportion of Arkansans living with HIV/AIDS is slightly lower than that in the overall US population, with 3,648 residents of Arkansas (0.14% of the population) reported to be living with HIV/AIDS in 2000 [9], compared to 450,151 individuals in the US as a whole (0.16% of the population). In 2000, the incidence of TB in Arkansas was higher than the US average, with a rate of 7.4 cases per 100,000 [144], compared to a rate of 5.8 per 100,000 in the US.

## Genotyping

An *IS6110* RFLP pattern was generated for all 700 (100%) genotyped isolates. The number of *IS6110* RFLP bands ranged from 1 to 22, with 196 isolates (28%) having less than 6 bands, 227 (32.4%) having 6-11 bands, and 277 (39.6%) with 12 or more bands. A pTBN12 RFLP pattern was generated for 181 of 196 (92.4%) of low IS*6110* band isolates, and for 132 of 504 (26.2%) of high IS*6110* band isolates. A spoligotype was generated for 697 (99.6%).

We evaluated the level of clustering using four different typing definitions: *IS6110* RFLP alone, spoligotyping alone, *IS6110*- pTBN12 RFLP, and *IS6110* –spoligotype (Figure 3.1). The proportion of isolates that were clustered ranged from a high of 80.8% clustered using spoligotype alone to a low of 47.1% using IS*6110* RFLP -pTBN12. Maximum cluster size corresponded to the proportion of clustered isolates by a given technique, with the largest *IS6110* RFLP-pTBN12 cluster having 17 isolates, and the largest spoligotype cluster having 97 isolates.

The proportion of isolates that was clustered varied according to the number of IS*6110* bands. Isolates with fewer than six bands showed consistently higher levels of clustering, despite the use of a secondary typing technique for these isolates (Figure 3.2).

### Characteristics of cases clustered by IS6110 RFLP with secondary typing by pTBN12

Clustering occurred significantly more often in cases that were aged 20-64 or residing in a correctional facility at the time of diagnosis, of non-Hispanic black race, or US-born. Additionally, cases reporting either homelessness or excessive alcohol use in the year prior to diagnosis were significantly more likely to be clustered. Clustering was significantly associated with a number of host demographic characteristics; including age group, race/ethnicity, and country of birth (Figure 3.3a). Additional host factors, including residence in a correctional facility at diagnosis, reported homelessness in the year prior to diagnosis, reported alcohol abuse in the year prior to diagnosis, and a positive sputum smear, were also significantly associated with clustering (Figure 3.3b).

### Investigation of Clustered Patients

Of 322 clustered cases for which epidemiologic linkage could be classified, a definite link was identified for 84 (26.1%), a probable link for 19 (5.9%), and no link was found for 219 (68.8%). The PPV of clustering ranged from a low of 0.17 using spoligotyping to a high of 0.32 using IS*6110* RFLP and pTBN12.

The PPV for typing by IS*6110* - pTBN12 varied by the number of IS*6110* bands, ranging from a low of 0.240 for isolates with five or fewer bands to a high of 0.405 for isolates with 12 or more IS*6110* bands (Figure 3.2). The PPV of clustering using IS*6110* alone, spoligotype alone, pTBN12 alone, and IS*6110* -spoligotype, varied considerably. The highest PPV, of 0.411, was attained by IS*6110* RFLP alone among isolates with 12 or more IS*6110* bands. IS*6110* RFLP alone also provided the highest predictive value for isolates with 6-11 IS*6110* bands. The higher PPV by IS*6110* RFLP alone than by IS*6110* RFLP with secondary pTBN12 typing is possible because this method considered only isolates with identical IS*6110* RFLP patterns as clustered, while IS*6110* RFLP with secondary pTBN12 considered isolates with 6 or more IS*6110* bands which differed by one band, but which had an identical pTBN12 pattern to another case, to be clustered.

Because IS*6110* -pTBN12 is considered the gold standard in TB typing in addition to showing the highest PPV in our study sample, we used this method to define clustering in our univariate and multivariate analysis of host and microbial factors associated with clustering and epidemiologic linkage.

Based on IS*6110* -pTBN12, the PPV of clustering was significantly associated with age, ranging from a high of 0.750 in the 0-19 age group to a low of 0.096 in the 64 to 85 age group (Figure 3.3a). The PPV was also significantly associated with both race/ethnicity and country of birth, rural vs. urban residence (Figure 3.3a), and alcohol abuse in the year prior to diagnosis (Figure 3.3b).

Epidemiologic linkages were significantly more likely to be identified for cases in younger age groups, for non-Hispanic blacks and Asian/Pacific Islanders, and for cases residing in rural areas at the time of diagnosis. Epidemiologic linkages were also significantly more likely to be found for cases reporting excessive alcohol use in the year prior to diagnosis. No clinical characteristic was significantly associated with the identification of an epidemiologic linkage.

Spoligotype families

SpotClust assigned 697 spoligotyped isolates to 29 strain families, with more than half in one of four common families. The most common family in our study sample was T1, accounting for 190 (27.3 %) isolates, followed by X1 (9.5%), X2 (9.0%), and LAM9 (8.5%). Four additional families S (7.3%), Haarlem3 (5.9%), Beijing (5.2%), and X3 (4.3%) were also common, and more than 75% of all spoligotyped isolates were classified into these 8 most common families.

Clustering was significantly associated with the spoligotype family of the infecting isolate. Among the four most common spoligotype families, cases infected with an X2 family isolate were most likely to be clustered, and cases infected with X1 family isolates were least likely to be clustered (Figure 3.4). The PPV of clustering varied substantially, though not statistically significantly (p=0.16) by spoligotype family (Figure 3.4), with the highest overall PPV, 0.61, in isolates of the S family, and the lowest PPV, 0.19, in isolates of the LAM9 family.

In order to consider the hypothesis that some proportion of false clustering might occur due to the presence of regionally endemic strain(s) and subsequent strain homogeneity, we evaluated the strain composition among clustered, US-born cases for which no evidence of a transmission linkage was identified.  Among the 211 US-born clustered cases for which no epidemiologic linkage was found, age at diagnosis was significantly associated with the spoligotype family of the infecting strain ($p < 0.0001$, Figure 3.5).  Spoligotype family X2 predominated in those aged 85 and over, while spoligotype family T1 predominated in those aged 20-44.

We assessed the "background diversity" of isolates in our study sample by examining the diversity of molecular types found among case-isolates with no identified transmission relationship with any other case-isolate in the study sample (therefore including both non-clustered isolates and clustered isolates for which no epidemiologic link was identified): among truly unrelated isolates, an ideal typing tool would identify all isolates as unique (D = 1.0). Background diversity varied substantially across spoligotype families, using either IS*6110* -pTBN12 or IS*6110* -spoligotype (Figure 3.6a), and using IS*6110* or pTBN12 alone (Figure 3.6b). The variation between strain families differed by different typing methods – for example, isolates in family X2 were most likely to be clustered at random using IS*6110* alone or IS*6110*-spoligotype, while isolates in the Beijing  were most likely to be clustered at random using pTBN12 alone or IS*6110*-pTBN12.

Multivariate analysis of falsely clustered cases

Because we were interested in identifying factors associated with false clustering among isolates not related by transmission, and because transmission-linked cases violate the assumption of independent observations required by multivariate logistic regression, we included only cases unlikely due to recent transmission (cases with a unique isolate by IS*6110*-pTBN12 and cases clustered by this method but for which no evidence of a transmission linkage was identified) in our multivariate analysis.  Compared to cases with unique isolates, apparently "falsely clustered" cases were significantly more likely to have reported excessive alcohol in the year prior to diagnosis, to be between the ages of 20 and 64, of non-Hispanic black race, and to live in the Northeast region of the state, after adjusting for all other variables in the model (Table 3.1). Additionally, clustered but unlinked cases were less likely to be infected with an isolate with more than six IS*6110* bands, and their infecting isolate was over four times more likely to be of the X2 or Beijing family than of the T1 family.

## Discussion

Strain diversity is a pre-requisite for molecular typing of *M. tuberculosis* isolates to effectively discriminate related from unrelated cases. While most geographic regions are characterized by highly homogeneous strain populations when defined by spoligotype [138], the spoligotype patterns reported from the United States are diverse, corresponding to the diverse geographic origins of the US population. The results we have presented here demonstrate that regionally endemic clones may persist at high prevalence, even within the diverse, dynamic population of the US.

Decreasing strain diversity (or a decrease in the "types" that can be resolved by a given molecular typing method) increases the probability that two unrelated isolates will exhibit the same molecular typing pattern by chance, thereby increasing the probability of false clustering. The strong link between diversity in *M. tuberculosis* and geography complicates typing not only because of the highly homogenous strain populations that often result, but also because the level of resolution provided by a particular typing tool may vary across different evolutionary lineages of the pathogen. Our observation of a disproportionate frequency of "unexplained clustering" among isolates of the X2 family is consistent with evidence that the diversity and evolution of typing markers may depend on the genetic background of the strain [125, 136]. While the influence of the genetic background of the X2 strain family has not been specifically investigated, ample evidence from isolates of the Beijing strain family demonstrates that such lineage-specific differences do occur [92, 137].

It is possible that the X2 strain family is endemic to a region which is not confined to Arkansas. The group of X2-family isolates in our study sample includes a sub group characterized by a prominent 2-band *IS6110* pattern, NATFP 00016. This *IS6110* pattern, which accounted for 5% of all isolates genotyped across the seven sites of the National Tuberculosis Genotyping Surveillance Network between 1996 and 2000 [139], was the most common RFLP pattern identified in Alabama, a southern US state with both historic and contemporary migration patterns similar to Arkansas [134]. It is speculated that this pattern identifies an endemic group of related strains that spread throughout that population during the TB epidemic of the 19th and early 20th centuries [134].

Global distribution patterns of the X strain family lend support to the speculation that a sub-family of this lineage, X2, may have been circulating in Arkansas and surrounding regions for generations. The X strain family is highly prevalent both in the British Isles as well as in former colonies, which suggest that the X family may be of British origin [132]. The Arkansas River Valley was settled in large measure by immigrants of Scotch-Irish heritage [145], so it is not surprising that a historically prevalent family may also have roots in the British Isles. The X2 family was not identified in California, despite a high incidence of TB in that state [139], suggesting that its spread in the United States may have occurred in the remote past.

Even under the best circumstances, epidemiologic investigations are unlikely to identify all transmission linkages. Individuals in high-risk populations may be reluctant to cooperate with public health officials to provide contact information[47], leading to missed linkages. Transmission may have also occurred via casual contact, which is unlikely to be identified by epidemiologic investigation [56]. Such un-identified transmission links are of critical importance to TB control efforts, as undetected transmission may allow ongoing transmission, as source cases go unrecognized and untreated. While a large amount of undetected 'true' recent transmission could plausibly explain our results, it seems unlikely that a sufficiently large proportion of all cases had undetected or misclassified transmission links. Additionally, the majority of X family isolates had fewer than 6 IS*6110* bands: as isolates with less than 6 IS*6110* bands are more likely to be falsely clustered than isolates with 6 or more IS*6110* bands [140], we considered the possibility that the association we observed between the X family and false clustering. However, this does not appear to be the case. In a multivariate model of cases assumed to be due to remote transmission (those with either a unique IS*6110*-pTBN12 type, or clustered but with no link identified), the association between false clustering and strain family was robust after adjustment for age, race, geographic region of the state, alcohol abuse, homelessness, and the number of IS*6110* bands.

This same multivariate analysis, however, does suggest that some proportion of the clustered cases for which no link was identified may be due to true recent transmission, as younger age, excessive alcohol use, and non-Hispanic Black race, all previously identified as risk factors for recent transmission in the US [43], were all significantly more likely be infected with a clustered isolate, but with no identified transmission link, than to have a unique isolate. It is likely, therefore, that some proportion of the "unexplained clustering" we observed is the result of transmission linkages that were not identified in the course of epidemiologic investigation. This underscores the need for further work to identify gaps in current epidemiologic contact tracing approaches.

Ideally, a measure of the sensitivity and specificity of clustering for recent transmission would allow us to separate the effects of differences in the prevalence of recent transmission and true differences in resolution between groups. Such a measure would allow us to better characterize the relationship between strain diversity and the validity of molecular clustering, which may have been inconsistent in our study due to confounding between strain-family and the prevalence of recent transmission. However, these measures require knowledge that is beyond the limitations of current tools in the study of TB epidemiology: the clear identification of all "true positives" and "true negatives" in the population, as a gold standard against which to compare the test results. While our current study is insufficient to provide conclusive evidence, we find a compelling argument that the X2 family is endemic in Arkansas, and that its prevalence compromises the validity of clustering detected molecular typing in this population.

The observation of homogenous strain populations in the US suggests that any widely used molecular typing technique must be able to discriminate between unrelated isolates from homogenous strain populations. A technique that has been demonstrated to discriminate between a diverse set isolates collected from across the globe may not be up to such a task [123]: validation sets, therefore, should consistently include unrelated isolates from homogenous strain populations. Differences in the diversity of molecular markers across strain families, along with differences in transmission dynamics in different population groups, suggests that the design and interpretation of molecular typing programs must be informed by an understanding of the local host and pathogen populations. Such differences present an obstacle to the development of a universal global typing technique.

# Tables

**Table 3.1.** Associations of host characteristics and SpotClust-assigned strain family with false clustering, among unique or clustered, unlinked isolates.

| Variable | Adjusted OR | 95% CI |
|---|---|---|
| Alcohol | | |
| ***Yes*** | **2.11** | **(1.04, 4.25)** |
| *No* | 1.00 | |
| Homeless | | |
| *Yes* | 3.94 | (0.86, 18.14) |
| *No* | 1 | |
| Age | | |
| *0-19* | 0.85 | (0.09, 8.53) |
| ***20-44*** | **5.13** | **(2.37, 11.07)** |
| ***45-64*** | **3.09** | **(1.52, 6.27)** |
| *65-84* | 1.68 | (0.89, 3.17) |
| *85 +* | 1.0 | |
| Race | | |
| *Non-Hispanic White* | 1.0 | |
| ***Non-Hispanic Black*** | **1.87** | **(1.09, 3.19)** |
| ***Other*** | **0.30** | **(0.12, 0.75)** |
| Spoligotype family | | |
| *T1* | 1.0 | |
| ***Beijing*** | **4.95** | **(1.84, 13.34)** |
| ***Haarlem3*** | **2.51** | **(1.01, 6.26)** |
| ***LAM9*** | **3.14** | **(1.44, 6.86)** |
| ***Other*** | **1.91** | **(1.04, 3.51)** |
| *S* | 1.62 | (0.62, 4.25) |
| *X1* | 0.65 | (0.25, 1.67) |
| ***X2*** | **4.75** | **(1.75, 12.87)** |
| Region | | |
| ***Northeast*** | **1.97** | **(1.07, 3.63)** |
| *Northwest* | 0.94 | (0.47, 1.84)) |
| *Southeast* | 1.06 | (0.53, 2.14) |
| *Southwest* | 0.73 | (0.37, 1.42) |
| IS*6110* RFLP Bands | | |
| *Less than 6* | 1.0 | |
| ***6 or more*** | **0.524** | **(0.28, 0.99)** |

# Figures

**Figure 3.1.** Impact of typing method on the PPV of clustering. The proportion of TB cases diagnosed in Arkansas between 1996 and 2000 that were involved in a cluster (light bars), the proportion of cases for which a transmission link was identified (dark bars), and the PPV (end of bar), varied by the typing method used.

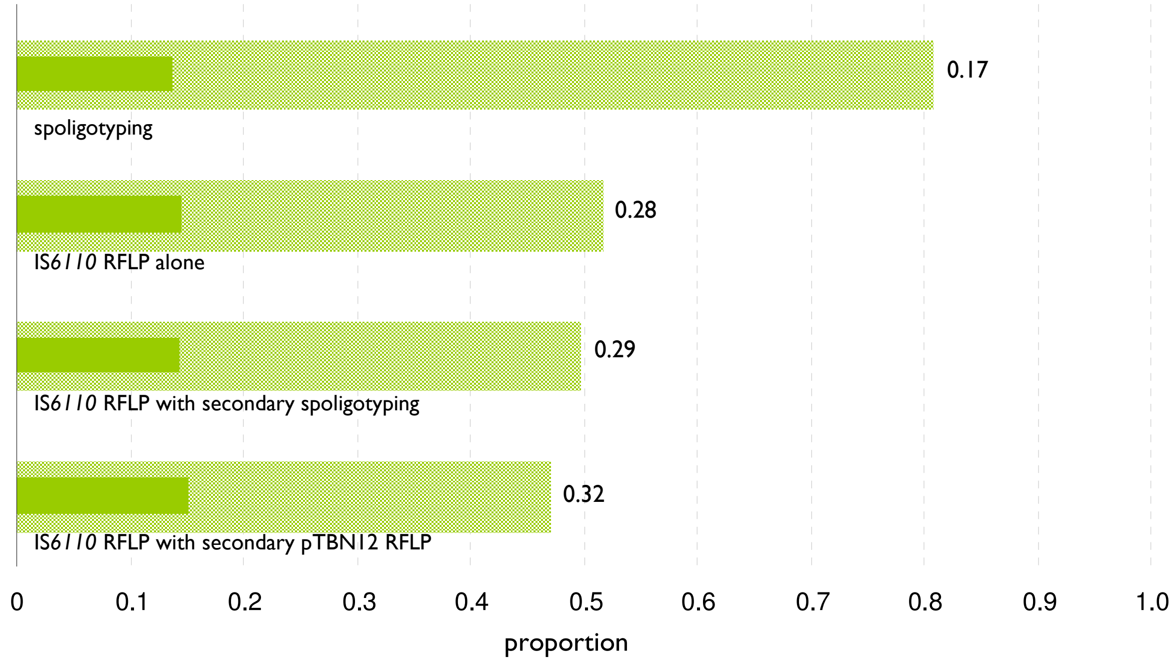


**Figure 3.2.** Impact of IS*6110* copy number on clustering. The proportion of TB cases diagnosed in Arkansas between 1996 and 2000 that were involved in a cluster (light bars), the proportion of cases for which a transmission link was identified (dark bars), and the PPV (end of bar), varied by the number of IS6110 elements and the secondary typing method used.

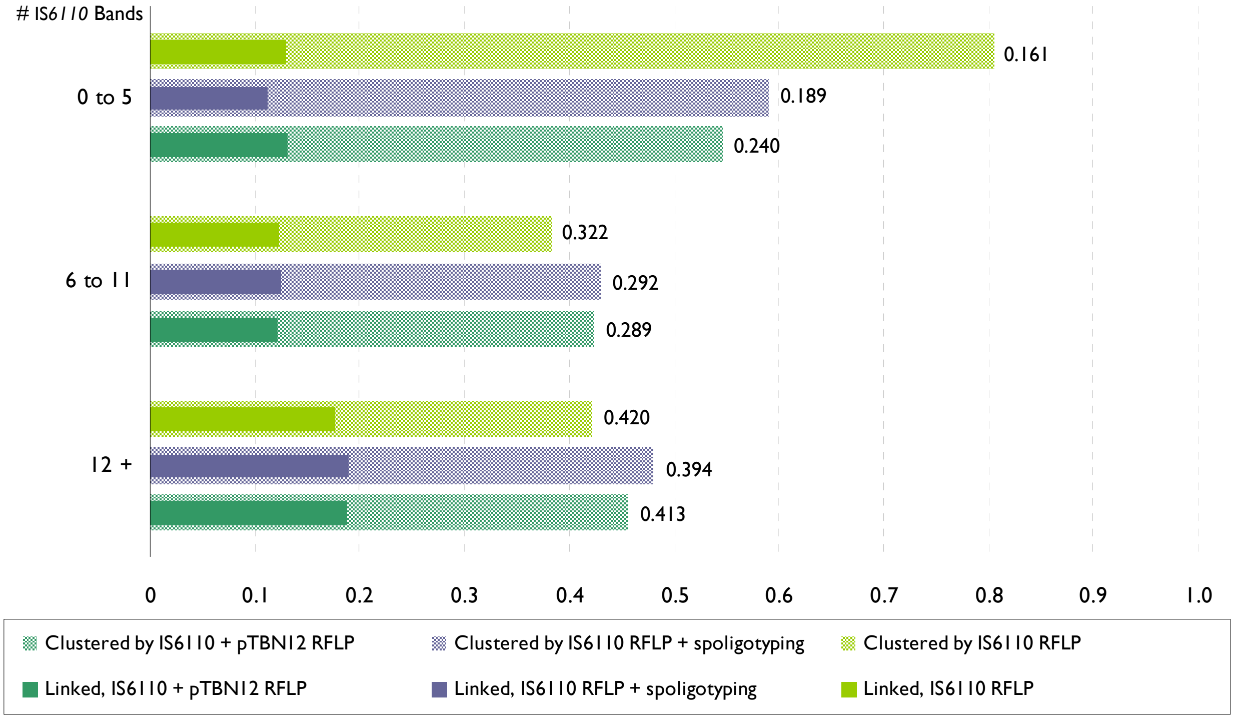


**Figure 3.3a.** Host demographic characteristics associated clustering or the identification of epidemiologic linkages for clustered cases. The proportion of TB cases diagnosed in Arkansas between 1996 and 2000 that were involved in a cluster (light bars), the proportion of cases for which a transmission link was identified (dark bars), and the PPV (end of bar), varied by host demographic group. Only characteristics that were significantly associated with either clustering or the PPV of clustering are represented in the figure.

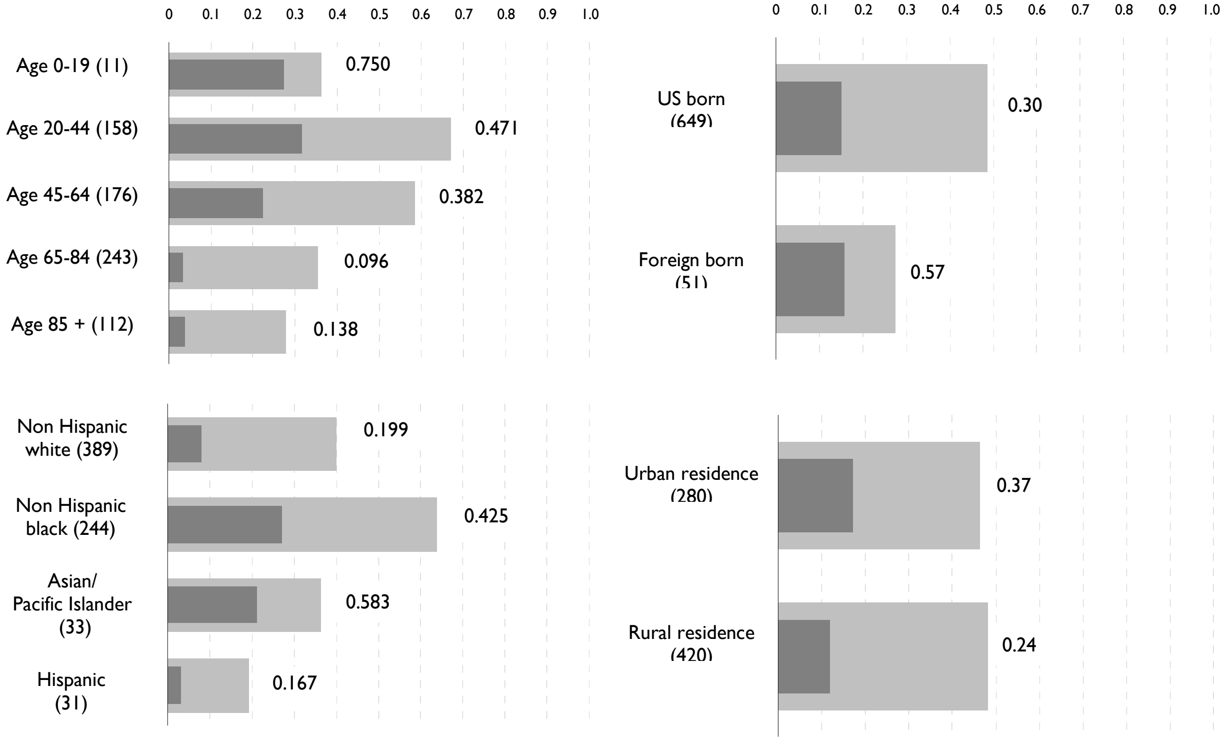


**Figure 3.3b.** Host clinical characteristics and risk behaviors associated with clustering or the identification of epidemiologic linkages for clustered cases. The proportion of TB cases diagnosed in Arkansas between 1996 and 2000 that were involved in a cluster (light bars), the proportion of cases for which a transmission link was identified (dark bars), and the PPV (end of bar), varied by host demographic group. Only characteristics that were significantly associated with either clustering or the PPV of clustering are represented in the figure.

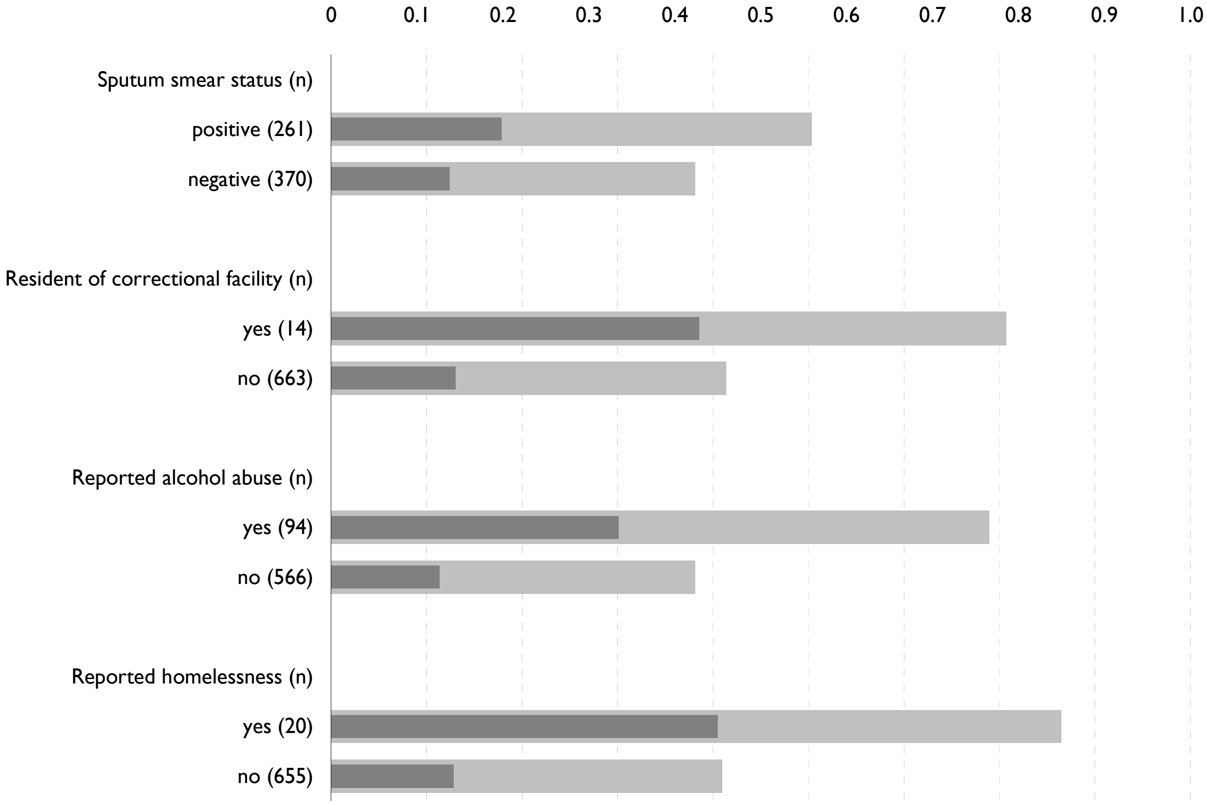


**Figure 3.4.** Spoligotype-defined strain family and predictive value of clustering**.** The proportion of TB cases diagnosed in Arkansas between 1996 and 2000 that were involved in a cluster (light bars), the proportion of cases for which a transmission link was identified (dark bars), and the PPV (end of bar), by spoligotype-defined strain family.

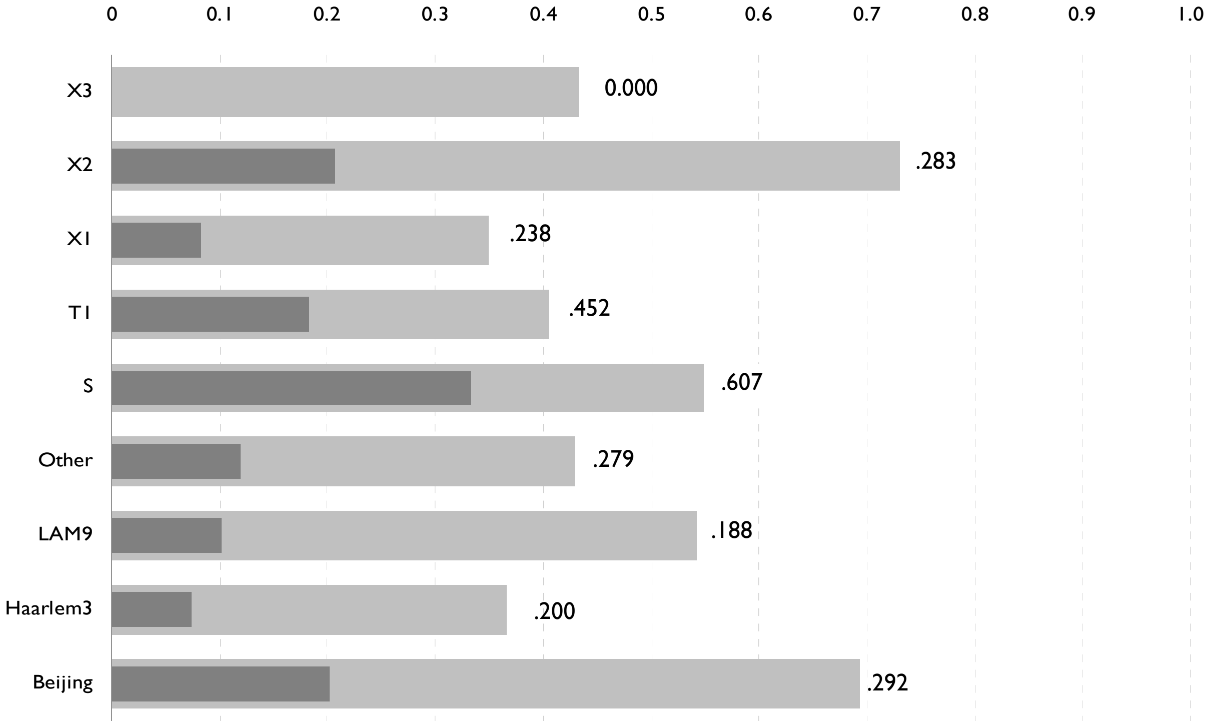


**Figure 3.5.** Falsely clustered cases by age and spoligotype family. Association between age group and spoligotype family among clustered, US-born TB cases for which no transmission link was identified, Arkansas 1996 - 2000.

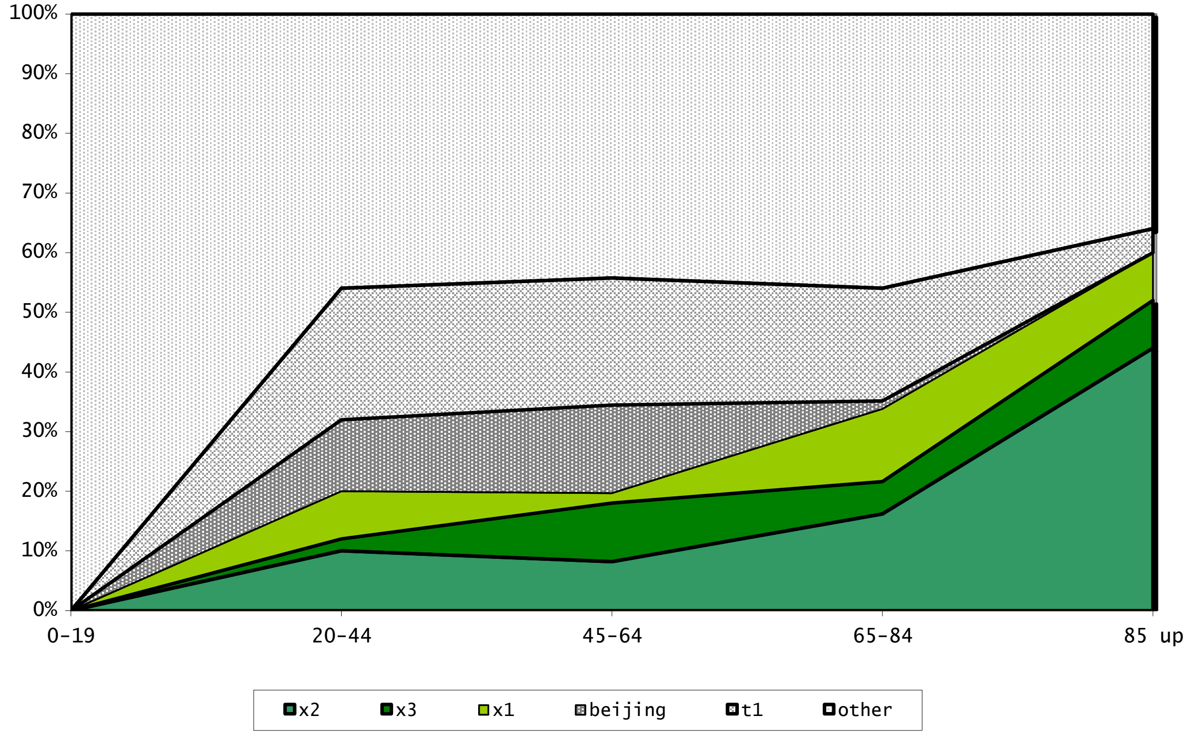
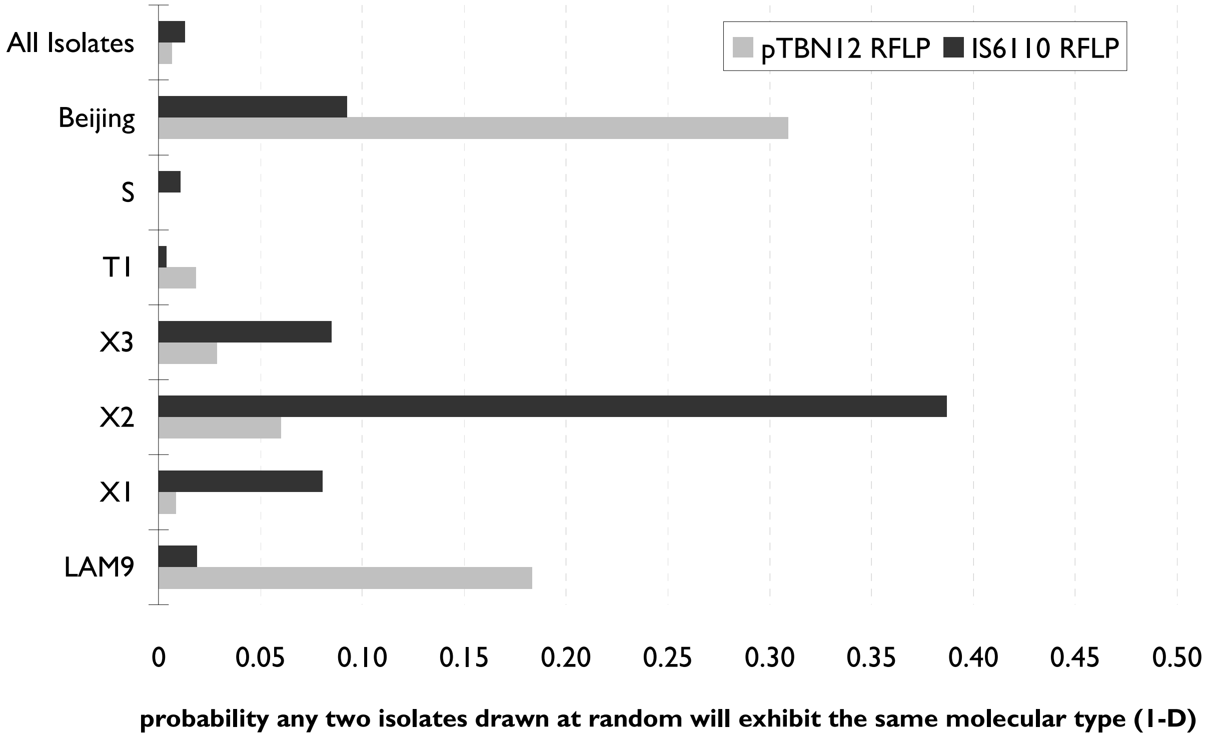
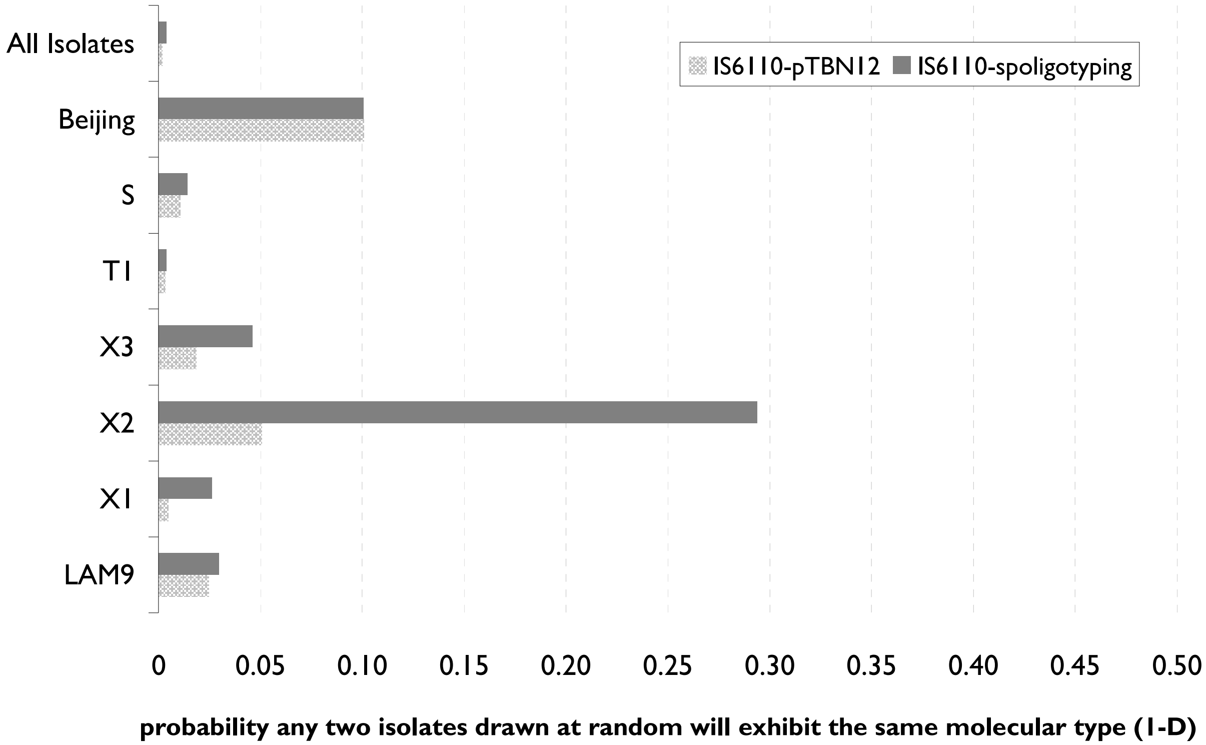


**Figure 3.6.** Background diversity by spoligotype family**.** Probability that two non-transmission related isolates, drawn at random, will exhibit the same molecular type by a) pTBN12 RFLP or IS*6110* RFLP alone, b) IS*6110* RFLP with secondary pTBN12 RFLP or IS*6110* RFLP with secondary spoligotyping. Calculated using the reciprocal of the Hunter-Gaston Index of Diversity (1-D) on the basis of incident case-isolates collected in Arkansas between 1996 and 2000 for which no evidence of a transmission linkage was identified.

**a)**



**b)**


All Isolates
Beijing
S
T1
X3
X2
X1
LAM9
▨ IS6110-pTBN12
■ IS6110-spoligotyping
0
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
0.45
0.50
probability any two isolates drawn at random will exhibit the same molecular type (1-D)

# Chapter IV

## Mutation, migration, and genetic variation in *M. tuberculosis* typing: separating the wheat from the chaff

### Introduction

Tuberculosis (TB) control programs in the United States and other developed countries increasingly rely upon the molecular typing of incident TB cases to direct contact investigation efforts. Used in conjunction with traditional contact tracing approaches, molecular typing data provides a mechanism to rapidly identify sources of infection in a population, allowing for more timely intervention. Typing data can bring attention to transmission events that remain undetected by traditional contact tracing methods, revealing previously unsuspected transmission venues that may be important in sustaining transmission in a population [57, 146]. Investigations aided by molecular typing have revealed the occurrence of transmission in bars [51, 57, 118], including transmission between unacquainted individuals who frequent the same bar [65], as well as transmission between unacquainted individuals visiting the same worksite [41], members of a church choir [66], residents of single-room occupancy hotels [57], and crack-cocaine abusers frequenting the same crack-houses [57].

While molecular typing provides an invaluable tool to the study of TB epidemiology, our understanding of the information these techniques can provide remains limited. Commonly, “clusters” of cases whose isolates generate the same molecular typing patterns using a particular molecular typing tool, are considered to be related by a recent transmission event, while cases whose isolates generate unique molecular typing patterns are considered to have disease caused by the reactivation of a remotely acquired latent infection. However, population-based investigations have identified cases that are clustered by molecular typing for which no evidence of a transmission linkage can be identified [70, 90], and cases with clear transmission linkages that exhibit unique molecular typing patterns [35, 147, 148]. Evidence suggests that the likelihood of “false clustering” (clustering of cases that are unrelated by a recent transmission) is influenced by demographic characteristics of the host population[70] [90] [44, 135, 149], the relative burden of TB in the host population [83, 150-152], characteristics of the local *M.*

*tuberculosis* population [153] [114] [136], and the particular molecular typing method used [37, 73, 75].

Mycobacteria of the *M. tuberculosis* complex are characterized by an unusually high degree of conservation in their housekeeping genes [126], and a clonal population structure. One of the greatest limitations to the development of effective molecular typing techniques for *M. tuberculosis* has been the identification of molecular markers with sufficient variability to distinguish between *M. tuberculosis* strains. An ideal typing marker would be sufficiently variable to distinguish unrelated cases, and evolve at a pace fast enough that isolates from unrelated cases are distinct, but not so fast as to obscure the relationship between case isolates that are truly related by transmission. What level of diversity and evolution might satisfy these criteria is unclear, and, despite the focus of considerable resources towards the identification and characterization of novel molecular typing markers [36, 123, 154, 155], this fundamental question has never been addressed. Perhaps more fundamentally, whether any single rate of marker evolution could be simultaneously both sufficiently slow to minimize false negatives and sufficiently rapid to minimize false negatives has not been established.

In recent years, increasingly sophisticated tools and *M. tuberculosis* genome sequence data have allowed the identification of a growing number of diverse genetic polymorphisms with potential as genetic typing markers. Many of these polymorphisms are tandem repeats of between 40 and 100 base pairs (bp), variously called *V*ariable *N*umber *T*andem *R*epeats (VNTR), *M*ycobacterial *I*nterspersed *R*epetitive *U*nits (MIRU), and *E*xact *T*andem *R*epeats (ETR). These polymorphic loci are thought to be the most variable structures in the *M. tuberculosis* genome[156], and an increasingly favored genetic typing approach, most commonly referred to as MIRU or MIRU-VNTR, relies on the PCR amplification of a “typing panel” of between 5 and 29 of these loci. The number of repeats at each locus is determined, and these data are concatenated into a digital string that can easily be communicated and compared across laboratories. This approach is highly analogous to microsatellite typing in higher eukaryotes, and the high-throughput methods that were originally developed for typing of these organisms has been adapted to use with *M. tuberculosis* [157]. MIRU-VNTR is rapid, economical, and reports suggest that results are highly reproducible [155]. At the same time, this technique presents potential to provide a highly flexible level of discrimination, as the number of loci included in, or interpreted from, the typing set could be varied.

An optimized MIRU-VNTR typing set may achieve very high levels of discrimination, and has the potential to allow the user to ‘tune’ discrimination to suit the needs of a given application. With current knowledge, however, we are unable to fully harness the power of this tool. While we may have a tool powerful enough to achieve a specified level of discrimination, we are not any closer to understanding what level of discrimination would best suit the needs of epidemiologic typing. The level of diversity used to determine the potential of each of these alleles as a typing marker remains an arbitrary one, set at a time when the few typing tools available, mainly based on phenotypic rather than genotypic assays, achieved limited levels of discrimination[143].

To inform the design and interpretation of an optimized MIRU-VNTR typing panel, it is essential that we clearly understand the aim we are tying to achieve. What would an ideal typing system look like? Epidemiology provides clear guideposts for the evaluation of tests: the sensitivity, specificity, predictive value positive (PPV), and predictive value negative (NPV) are key measures by which to compare alternative testing approaches. Calculating these measures requires a gold standard against which to compare test results, however, and no currently available test or evaluation can accurately identify the “truth” against which to compare molecular typing results: cases of active TB truly linked by recent transmission. The handful of validation studies that have been conducted have been limited to investigations of transmission linkages among cases known to be linked by molecular clustering, allowing estimations of the PPV of clustering as a measure of recent transmission. The sensitivity, specificity, and NPV of clustering, however, cannot be evaluated in empirical investigations, given the inherent ambiguity of even the highest quality epidemiologic data. These limitations of empirical validation studies preclude the rigorous comparison of various molecular typing methods and strategies, and present an obstacle to the development of more accurate molecular typing techniques.

Model representations of real-world systems provide a powerful opportunity to gain insight into questions that cannot be effectively addressed through investigation in real populations. Agent-based stochastic computer simulation models in particular allow investigators to effectively conduct systematic experiments in a system in which the “truth” is known. Unlike equation-based models (EBMs), a formalism more widely applied to the modeling of epidemiological and biological systems, agent-based models (ABMs) represent elements of the system as discrete elements which interact with other elements of the model according to a specified set of rules. Interactions occurring at the individual level generate larger, system-wide behavior of ABMs, and this “bottom up” structure is uniquely suited to represent the complex interaction of biological and social systems that drives epidemiologic processes. ABMs (also known as individual based models, or IBMs), have been widely applied to the study of biological systems, including complex interactions between the host and pathogen governing granuloma formation in lung tissue following infection with *M. tuberculosis* [88]. To date, one ABM has been applied to the study of the molecular epidemiology of TB, to evaluate the hypothesis that *M. tuberculosis* strains involved in large clusters are more transmissible or pathogenic than those involved in small clusters [89]. This ABM simulated TB transmission in a hypothetical population, and tracked the identity of individual strains as they were transmitted. While this ABM was not applied to evaluate the validity of clustering as a marker of recent transmission, and was not formulated in a way that would allow such an application, it illustrated the potential utility of the ABM method towards this end.

To better characterize the impact of typing marker diversity and stability on the validity of “clustering” as a marker of recent transmission, and to gain insight into the influence of host demographic factors and pathogen population structure on the occurrence and validity of clustering, we developed an ABM of tuberculosis transmission which tracks the identity of individual strains as they are transmitted through a simulated population. This model was informed by the theoretical framework of a previously described ABM of TB transmission [89], but extends and modifies key features of this model in fundamental ways. Our model explicitly represents molecular typing markers, and allows markers to mutate over the course of infection. The typing system represented is based on MIRU-VNTR, and allows for various aspects of the typing panel, including the number of loci, the average allelic diversity at each loci, and the mutation rate for a given loci, to be modified and evaluated. Using this model, we consider the diversity of molecular ‘types’ achieved by a range of different typing panels, and evaluate the relationship between this diversity and the validity of clustering. We investigate these relationships in three distinct, population-specific transmission scenarios, allowing us to explore the sensitivity of clustering to population-specific factors. Lastly, we consider the conflicting goals of minimizing false clustering while maximizing sensitivity to recent transmission, asking whether a single typing panel can simultaneously achieve these aims.

## Model description and methods

### Simulation Model

An explicit description of the model rules in outline form can be found in the appendix at the end of this chapter.

A model was developed to simulate the dynamic transmission of discrete "strains" of *M. tuberculosis* through a population of human hosts. The model specifies a population of discrete individuals, each characterized by variables indicating age, TB infection status and infection state (if infected), household, and neighborhood. If infected with TB, an individual is further characterized by a variable linking it to its infecting bacterial "isolate", which itself carries an identifier corresponding to the parent strain from which it descended. A typing system to resolve infecting isolates, characterized by variables indicating the number of loci assessed, the initial allelic diversity and level of instability observed at each loci, and the maximum possible number of alleles for each loci, is also specified. The model is run in time-steps of one week, at which time processes governing the transmission of *M. tuberculosis,* progression to disease following infection, and vital dynamics of the population occur according to a specified set of parameters. Strains are tracked as they are transmitted through the population, and incident infections and cases of active disease are monitored and recorded. "Clustering" of cases, based on the specified typing system and study period, is determined, allowing the validity of this measure to be assessed across a range of simulated scenarios.

### Initialization of the model

At the initialization of each model run, a simulated population of individual "agents" is created, with each individual assigned an age, household, and neighborhood according to specified distributions (for example, the age distribution for the Arkansas-based transmission scenario follows the most recent age distribution reported by the US Census for that state). This formulation of host contact structure is analogous to that presented in an earlier ABM of TB transmission [89]. HIV infection status is assigned according to the specified prevalence. Each individual is also assigned a group of friends from the same neighborhood and age group (ages 0-9, 10-19, 20-44, 45-64, and 65+), following a normal distribution around a specified average number of friends. This average, and the distribution around it, was based on named contacts identified in the course of TB contact tracing investigations, reviewed in the course of a previous investigation (described in Chapter III of this dissertation). Neighborhood members, household members, and friends remain essentially fixed throughout each model run, modified only by birth, death, and migration.

While cases of active TB are reported more often in men than in women, the reasons for this difference are unclear [158]. Because the relationship between gender and the parameters governing the transmission and progression of *M. tuberculosis* infection is unknown, gender was not specified in the model. Neighborhood size, corresponding to definitions of neighborhood used previously in studies of neighborhood and health, is 500 agents [159]. Household size is specified based on census estimates in the specific population represented in the model. In order to ensure that no household is composed entirely of children (and to ensure the between-age group mixing that occurs within the typical household setting), the age distribution within households is specified such that any given household with one more agents under 18 years old will also include at least one agent over age 25.

The annual risk of infection (ARI) is the risk that, in a given year, a previously uninfected individual will become infected with *M. tuberculosis*. The ARI has historically provided a key measure of tuberculosis transmission, and is commonly estimated either from consecutive tuberculin survey data in a single cohort, or from tuberculin survey data in multiple cohorts for a single time point (with the assumption of a constant risk of infection over the lifetime of the cohorts) [108]. Based on the specified initial ARI, and trend in ARI over time, agents created at the initialization of each model run are "seeded" with a prevalent latent infection. This is implemented in an age-specific manner, such that each agent experiences the cumulative infection risk corresponding to the calendar years over which that agent was alive. The identity of each isolate is specified, and the time since infection, which determines the risk that that infection will progress to active disease, is recorded. Because latently infected individuals who were infected through the same chain of transmission will have closely related isolates, a proportion of initially infected individuals are infected with isolates of the same strain (and thus identical at all typing markers). The proportion of individuals in the model initially infected with such a "clustered" isolated is specified, based on estimates of the proportion cases in a population that are truly related by transmission. In accordance with historic transmission patterns, initial latent cases within the same household, or which have a similar time since infection (e.g., were infected close in calendar time) [108], are more likely to be clustered at the initiation of the model run.

The number of loci included in the typing panel is specified, with a value of 12 loci chosen to correspond to the size of the standard MIRU typing panel [37]. The diversity of alleles, or values, present at each loci is specified based on a best-case estimate reported in a group of diverse isolates collected from locations across the globe [123]. Based on this allele diversity, each initially infecting isolate is assigned a value for each of the loci included in the typing panel. The first value for each allele is randomly drawn from a uniform range between 1 and a specified maximum number of alleles. As dictated by the specified level of allele diversity at each locus, each isolate is either randomly assigned an already existing allele value for that locus, or a new value. New allele values for a given locus draw the lowest or highest current allele value, and add or subtract one repeat, as possible within the given range and resulting in a new value. This stepwise mutation pattern is consistent with observations of the distributions of variability at MIRU loci [125]. As they correspond to the physical number of repeats present at a given locus, allele values cannot be below 1, and cannot be higher than the maximum number of alleles specified.

Each locus is assigned a mutation rate drawn from a normal distribution around the specified mutation rate. Mutation rate governs the probability an allele will change in any one week interval over the course of active infection, and is specified based the rate of MIRU allele change observed in serial isolates collected from patients with active TB disease [160]. Bacterial replication is not explicitly represented in the model.

The model run.

The model is run in time-steps of one week. At each time step, general population processes occur: individuals enter the model through birth and immigration, leave the model through death and emigration, and age. At the same time, processes relevant to TB transmission occur: individuals contact other individuals, transmit infection, develop active disease, and recover, and infecting bacteria mutate to create variant allele profiles, with specified probabilities. Upon infection, an individual acquires a variable that indicates the identity of the infecting strain. Parameter definitions and default values are defined in Table 1.

For a schematic illustration of model flows, see Figure 4.1. Please also refer to Appendix 1, in addition to parameter estimates and variable descriptions included in this section, for a more comprehensive background on the processes represented in the model.

### Transmission of infection

The probability of contact between a susceptible and infected individual is a function of the social contact structure of the population, as well as the prevalence of active, infectious disease in the population. Little information on any actual social contact structure (as relevant to TB transmission) is available. Historically, most TB transmission is thought to have occurred within the home and family [58, 59], however, recent reports have documented the occurrence of transmission through casual contact [41, 56]. We defined a contact in our model as a relationship with the potential to facilitate the transmission of *M. tuberculosis.* The social contact structure represented by the model is consistent with evidence that the majority of transmission occurs within the household, but that transmission also occurs between close non-household contacts, and may occur between casual contacts. Each individual belongs to a household and neighborhood: an individual is in contact with every member of that household, as well as a fixed group of age-group specific friends in the same neighborhood, at every time-step. In addition, an individual may come into "casual" contact with another individual in the same neighborhood, and, less frequently, with an individual in another neighborhood. Consistent with transmission rates observed in contact tracing studies [161-163], the risk of transmission is highest between household contacts, with transmission between friends or casual contacts occurring at a specified fraction of that between household contacts.

Among individuals with respiratory forms of active TB disease, the detection of acid-fast bacilli (AFB) in the sputum provides a relative indicator of the infectiousness of the case [12]. Sputum-smear positive cases (in which AFB is detected in the sputum) are more likely to transmit disease than are sputum-smear negative cases [15, 164]. Both smear positive and smear negative cases are represented in the model: transmission from smear negative cases is defined relative to transmission from smear-positive cases.

### Progression from infection to active disease

In order to simplify the model, only respiratory forms of active disease are explicitly represented, as only these forms of disease may transmit infection. The risk of progressing to extrapulmonary disease (briefly reviewed in Appendix 1) is essentially ignored: the risks of progression that are used specify only the risk of progressing to respiratory forms of disease. Risks of progression are specified according to age group and the number of years since infection. Consistent with models presented by Sutherland[108] and Vynnycky and Fine [81], the risk of progressing to active disease is highest the first year following infection and declines each subsequent year to the fifth year following infection (corresponding to the risk of rapid or primary progression). Although over a slightly longer interval following infection than is often described (5 years as compared to 2), the declining risk by year means that the risk of primary progression is highest in the first two years after infection, corresponding to the conventional model of disease progression. From the 6th year following infection on, the risk of infection remains constant, at a level corresponding to the risk of developing reactivation disease. Each new active case of disease is designated as either sputum smear positive or sputum smear negative. This designation depends only on age.

### Re-infection

Consistent with available evidence, primary infection is not assumed to provide any protection against subsequent re-infection[17, 19-23, 81], with the exception that individuals in the first five years since any infection event (primary or otherwise) cannot be re-infected (consistent with previous TB transmission models [108]][81]). In any other state, any individual in the model is equally susceptible to infection or re-infection, given contact with an infectious individual. While not providing protection against re-infection, a primary infection does provide protection against developing active disease subsequent to a re-infection event. This protection is reflected in the estimates of the age-specific risks of progressing to disease following a re-infection event, which are those estimated by Vynnycky and Fine [81]. Our formulation of re-infection is consistent with that used by Vynnycky and Fine, as well as Sutherland [165].

Once infected with a given strain, individuals are assumed to be infected with that strain for the duration of their life. Because of re-infection, therefore, individuals may be infected with more than one strain. A multiply infected individual, however, risks disease only from the most recently infecting strain.

Consistent with the theoretic models that informed our formulation of re-infection, disease following the reactivation of a latent infection, or "endogenous disease", is disease with onset 5 or more years after infection, or the most recent re-infection. Therefore, after the 5th year post re-infection, the probability of developing reactivation disease is identical to the rate of reactivation disease. At 5 years post re-infection, reactivation due to the original infecting strain can no longer occur [81, 108].

### Duration of infectious period

We have calculated an estimated average duration of the infectious period from data on the average delay to diagnosis (including both patient and health care delays) and the average duration from initiation of treatment until the resolution of infectiousness (measured by culture conversion). Using 75 days average median delay to diagnosis [166], and 35 days median time between initiation of chemotherapy and culture conversion [167] gives a median infectious period of 110 days, or approximately 16 weeks.

A single parameter, the recovery probability, controls the duration of the infectious period. According to a preliminary diagnostic model experiment, a recovery probability of ~0.06 gives an average infectious period of approximately 16 weeks [166].

### Relapse

Following recovery, an individual moves to a "recovered" state. In this state, an individual is susceptible to re-infection, as well as to reactivation disease caused by the initial infecting organism.

### Case fatality

In the United States in 1989, the case fatality for TB  was 8.38 deaths per 100 TB cases; this was an increase from the lowest TB case fatality recorded in the United States, which was 7.08 per 100 TB cases in 1981 [168]. Case fatality is related to age, with a higher risk of death due to TB in older individuals. In order to approximate these case fatality rates, the weekly risk of death due to TB was calculated relative to the average duration of disease. For the purposes of the model, this is identical to the average duration of infectiousness. Weekly age-specific risks of death due to TB were calculated based on a disease duration of 16 weeks.

### Vital dynamics

Individuals agents can migrate into or out of the model, die, and give birth to new agents. Birth rates are specified according to reported population-specific estimates of total fertility rate (TFR), and only agents of reproductive age can give birth to new agents. Rates of immigration and emigration are based on population-specific census estimates, where available.

## Implementation of the model

The ABM was implemented in JAVA, using the Recursive Porous Agent Simulation Toolkit (Repast). The model is updated asynchronously, on a time-scale of one week. The program runs for a simulated time period of 20 years. Table 4.1 summarizes values for all model parameters used to simulate the specific transmission scenario of Arkansas. Table 4.2 summarizes parameter differences between three population-specific transmission scenarios; Arkansas, Malawi, and Afghanistan.

## Model measurements

Over the course of each simulation, the model generates calculations including measurements of the incidence and prevalence of disease, the proportion of cases that are clustered by molecular typing, the validity of clustering, and the diversity of molecular typing patterns among both incident cases of active TB disease and among all prevalent TB infections. All calculations are based on the model-specified study duration. The study duration defines the time period over which cases are evaluated, thus defining the study population in which both clustering and transmission linkages are identified.

Four key measurements of the validity of clustering are calculated: PPV, NPV, sensitivity, and specificity. These measurements are calculated as follows:

| | Linked by transmission to another case diagnosed during the study period | Not linked by transmission to another case diagnosed during the study period | |
|---|---|---|---|
| Clustered (Identical typing pattern to that of at least one other case diagnosed within the study period) | A | B | A + B |
| Non-clustered (Unique typing pattern within the study period) | C | D | C + D |
| | A + C | B + D | |

$$\text{PPV} = \frac{\text{A}}{\text{A} + \text{B}}$$

$$\text{NPV} = \frac{\text{D}}{\text{C} + \text{D}}$$

$$\text{Sensitivity} = \frac{\text{A}}{\text{A} + \text{C}}$$

$$\text{Specificity} = \frac{\text{D}}{\text{B} + \text{D}}$$

For a more straightforward interpretation, we also present the false-positive rate (FPR = 1 – specificity) and the true positive rate (TPR = sensitivity), rather than sensitivity and specificity, in some sections of our results.

The diversity of molecular typing patterns was estimated using the Hunter-Gaston Discrimination Index (HGI) (also referred to as Simpson's Index of diversity) [143], calculated by the following formula:

$$D = 1 - \left[ \frac{1}{N(N-1)} \sum_{j=1}^{s} \mathrm{n}_{\mathrm{j}} (n_j - 1) \right]$$

where D is the numerical index of discrimination, N is the total number of strains in the typing scheme, s is the total number of different strain types, and $\mathrm{n}_\mathrm{j}$ is the number of strains belonging to the jth type. D can be interpreted as the probability that any two isolates drawn at random will be of different molecular types.

All calculations for each run were calculated during the last time-step of the model, with a study population defined according to the model-specified study duration.

Experiments

In all experiments, including those conducted as part of model validation and sensitivity analysis, simulations for each set of experimental conditions (defined by a specific parameter values) were replicated with 5 model runs. All model measurements were averaged over these 5 runs, and the standard deviation (SD) was calculated.

Evaluating the validity of the model

With the purpose of testing the basic assumptions of our model across a range of simulated scenarios, we generated population-specific transmission scenarios representing three diverse settings: Malawi, characterized by both a high burden of TB and a high burden of HIV infection, Afghanistan, characterized by a high burden of TB and a negligible burden of HIV, and Arkansas, a southern US state characterized by a low burden of TB and a low burden of HIV. Population-specific parameters used to generate each of these scenarios are presented in Table 2.

Sensitivity and Uncertainty analysis

Many of the parameters in our model could be confidently estimated from demographic and surveillance data. Parameters governing certain key aspects of the transmission and natural history of TB infection, however, are less well understood, and the values used in the model for these parameters are less certain. We explored the sensitivity of our key model measures: including the sensitivity, specificity, PPV, and NPV of clustering as a marker of recent transmission, to key parameters. These parameters were identified on the basis of two criteria: first, uncertainty in the parameter estimate, and second, a biologically plausible relationship with the outcome. For these parameters, we evaluated the correlation between the given parameter and each outcome measure listed above. The parameter range evaluated was chosen based on biologically plausible values each individual parameter might take. For parameters where uncertainty was high, we explored the entire range of possible values, while for more confidently estimated parameters, a more limited range of plausible values was explored. With the exception of HIV, which was evaluated against the background of the transmission scenario based on Malawi, all parameters were evaluated against the background of the transmission scenario based on Arkansas.

The parameters evaluated can be classified into two groups: parameters governing host factors, and parameters governing microbial factors (Table 3).

# Results

## Validation of the model to country-specific settings

Transmission-related statistics generated from the model together with available empiric measurements of these same statistics for each setting is presented in Table 4.3. In each case, model-generated estimates compared well to empiric measurements without varying other model parameter values. Estimates of the prevalence of latent infection showed the highest level of agreement with measured values, while estimates of ARI% showing the highest discrepancy.

## Sensitivity Analysis

### HIV prevalence

Increasing HIV prevalence was positively correlated with incidence, prevalence, and ARI%. Interestingly, increasing HIV prevalence was actually associated with a slight decline in the proportion of disease due to recent infection, while at the same time corresponding to a slight increase in the proportion of clustered cases. Increasing HIV prevalence was positively correlated with sensitivity, and a slight negatively correlated with specificity. However, no association was seen between the prevalence of HIV and the PPV or NPV.

### Transmission probability

Similarly to HIV, increasing the transmission probability was positively correlated with incidence, prevalence, and ARI%. It was also associated with a higher proportion of disease due to recent transmission, and a higher proportion of clustering. Additionally, it showed a strong correlation to sensitivity and PPV, but no association with specificity or NPV.

### ARI% and ARI% trend

A higher initial ARI% was associated with a higher incidence and prevalence of disease – this trend was magnified with increasing levels of ARI% trend (since ARI% trend reflects the historic yearly decline in the ARI%, a higher value for this parameter corresponds to a higher prevalence of infection among increasingly older birth cohorts, with the birth cohort born the year prior to the model initiation showing a prevalence that corresponds to experiencing 1 year at the initial ARI%). The impact of ARI and ARI trend was highly dependent on the level of true transmission-clustering at initiation. If all initial infected cases (active and prevalent) were linked by transmission to another case in the population, increasing ARI% showed a strong positive association with clustering. By contrast, when no initial infected cases were linked by transmission to another case in the population , increasing ARI% showed a strong negative association with clustering. The opposite pattern was observed for the relationship between ARI%, initial transmission-linkage and specificity: an increasing ARI% corresponded to increasing specificity of clustering for recent transmission when few initially infected cases were linked by transmission, was low, while it corresponded to a decreasing specificity when many initially infected cases were linked by transmission. Both NPV and PPV increased with increasing initial ARI%, although unstable estimates at low values of initial ARI% meant that this trend was only clearly observed when the yearly trend in ARI% was greater than 0.1.

### Recovery probability

An increasing weekly probability of recovery was associated with declining incidence, prevalence, and ARI%, as well as a declining proportion of disease resulting from recent transmission. No clear association was seen, however, between recovery probability and clustering, or with the sensitivity, specificity, PPV, or NPV of clustering as a measure of recent transmission.

### Maximum number of alleles

No association was seen between the maximum number of alleles and clustering, or with the sensitivity, specificity, PPV, or NPV of clustering as a measure of recent transmission.

### Mutation rate during latency

No association was seen between the maximum mutation rate during latency and clustering, or with the sensitivity, specificity, PPV, or NPV of clustering as a measure of recent transmission.

### Migration

Increasing migration was associated with increasing overall diversity when the proportion of initial infected cases (active and prevalent) that were linked by transmission to another case in the population was high, but was associated with decreasing overall diversity when the proportion of initially transmission-linked cases was low. No clear relationship was observed between the rate of migration and the proportion of clustered cases, nor with the sensitivity, specificity, PPV, or NPV of clustering as a measure of recent transmission.

### Historic transmission patterns

The proportion of initial infected cases (active and prevalent) that were linked by transmission to another case in the population was strongly associated with the validity of clustering when more than 80% of initially infected cases were linked by transmission. When the proportion of initially infected cases were linked by transmission increased beyond this level, the overall diversity of molecular types began to fall rapidly, the NPV declined sharply, and the false positive rate increased sharply. The PPV and sensitivity, however, were not affected.

This initial proportion of clustered cases also appeared to modify the impact of other model parameters, including ARI% and the rate of migration, on diversity and clustering.

## Key Characteristics of Interest

### Population-specific transmission setting and study period

For each population-specific setting, we assessed clustering results across a range of simulated study periods. In these model settings, proportion of case clustering was sensitive to both the population-specific transmission scenario and the study duration over which molecular typing results and transmission linkages were considered (Figure 4.3). The variation in the levels of clustering and recent transmission was greatest when the duration of the study was long. In the transmission settings corresponding to both Afghanistan and Malawi, the PPV of clustering for recent transmission increased with increasing duration of study period (Figures 4.4). The FPR of clustering showed a slight positive correlation with increasing study period in Afghanistan, but no clear association with study period in Malawi (Figure 4.4). Measures for Arkansas were unstable, and are not included in Figure 4.4 as no clear trend was observed between study period the validity of clustering in this setting.

## Polymorphism at individual MIRU loci

The overall diversity of molecular types present in the population, as measured by the HGI, was positively correlated with increasing diversity at individual alleles in all three population-specific transmission scenarios considered (Figure 4.6). In each scenario, the overall diversity increased consistently sharply between as the average diversity at individual alleles increased from 0 and 0.1, and rose consistently thereafter, until the upper limit of pattern diversity was reached after an average individual allele diversity of approximately 0.8. The relationship between individual allele diversity and the total pattern diversity differed when only the molecular types that were observed in the population (those associated with an incident case of respiratory disease) were considered. In this case, the observed diversity of types increases following an approximately sigmoid curve, with the upper limit of pattern diversity reached as the average individual allele diversity rose above 0.75.

The false positive rate of clustering remained steady, at approximately 1.0, when individual allele diversity values were between 0 and 0.6 (Figure 4.5). As average individual allele diversity rose above 0.6, however, the false positive rate declined in each population-specific transmission scenario. The decline was greatest in the Arkansas-specific transmission scenario, and least in the Malawi-specific transmission scenario. In parallel with the false-positive rate, the PPV of clustering remained steady in each transmission scenario until an individual allele diversity of 0.7, beyond which the PPV increased dramatically in the Arkansas scenario, slightly in the Afghanistan scenario, and not at all in the Malawi scenario. In each scenario, the true positive rate was consistently close to 1.0 regardless of the average individual allele diversity.

## Marker instability and the number of markers in a typing panel

In order to gain insight into the relationship between marker stability and the number of markers included in a typing panel on the validity of clustering as a measure of recent transmission, we examined combinations of these parameters in a setting where all initially infecting cases were clustered, and where no novel strains were introduced by immigration. At levels of marker instability above 0.00005, the observed level of diversity in molecular types increased sharply (Figure 4.8). Increasingly higher allele instability corresponded with a sharp decline in the false positive rate, but also with a sharp decline in the true positive rate. PPV remained steady with increasing allele instability, while NPV declined slightly (Figure 4.7).

## Discussion

For the purposes of epidemiologic typing, molecular typing markers “should be sufficiently polymorphic to distinguish unrelated strains yet be stable enough to identify isolates of the same strains”[160]. Using an agent-based model of TB transmission which explicitly represents molecular typing based on MIRU-VNTR, we have generated the first quantitative insights into just what “sufficient” polymorphism and “enough” stability may be, describing the relationship between the diversity and stability of individual typing markers and the validity of molecular “clustering” as a marker of transmission relationships. Our simulation data suggest that underlying, “base” allele diversity, and diversity resulting from highly unstable typing markers, impact the validity of clustering measures in distinct ways, and that the mechanism by which marker diversity is generated in a population should be carefully considered when identifying potential typing markers, and when interpreting the results of molecular typing investigations. Additionally, we describe the importance of population-specific factors on the validity of typing measures, and suggest that a single universal typing panel is unlikely to provide consistent and valid results across diverse global populations.

In a single geographically defined population, diversity in the alleles identified at a given typing marker may result from a long history of evolution and divergence within that population, from rapid marker evolution occurring over short time scales, or from the introduction of novel strains from distant populations. The strong phylogeographic associations observed in the global population structure of *M. tuberculosis* [132] suggest that, on the small geographic scale most relevant to TB transmission, the latter two process most likely contribute to the diversity seen. Of the simulated transmission scenarios we evaluated, high levels of “existing” marker diversity, such as that achieved through the introduction of novel strains from diverse populations, provided optimal typing results compared to diversity achieved through the rapid accumulation of changes in highly unstable markers.

Diversity achieved through increasing allele stability presents a trade off: while increasing instability decreases the probability that bacterial isolates from unrelated cases will exhibit identical molecular typing patterns, it also increases the probability that bacterial cases from truly related cases will be unique. This presents a concern for TB control programs in particular, which increasingly rely on molecular typing tools to direct limited TB control resources. Failing to identify pockets of recent transmission, as will occur at any true positive rate below 1.0, may blind investigators to important transmission venues in a population, result in undiagnosed, infectious cases of active TB remaining under the radar, or both. In the hypothetical scenario for which we evaluated this question, a TPR of 1 corresponded to, at best, a FPR of more than 0.9. Misclassifying 90% of unrelated cases as due to recent transmission in the population, as would occur at this level, would effectively overburden any TB control program.

By contrast, high levels of "existing" marker diversity allowed for a reduced FPR without a parallel reduction in the TPR. Importantly, we observed a critical threshold value in the average individual allele diversity, below which overall diversity is low, and the specificity of clustering as a measure of transmission is negligible. Diversity increases dramatically once average allele diversity exceeds this value, however, and specificity increases in parallel. For the 12-loci typing panel assessed in our model, this threshold value appears to be an HGI of approximately 0.7. Population factors modified this relationship, however, and the lowest FPR achieved, using a 12-loci typing panel with typing markers that are perfectly diverse in unrelated isolates (HGI = 1.0), ranged from less than 0.10 in the low-TB, low-HIV setting of Arkansas, to above 0.7 in the high HIV, high TB setting of Malawi.

This finding has significant implications for the interpretation of results generated with current MIRU typing panels, and for the selection of MIRU loci for new typing panels.
Multiple reports have described low levels of MIRU allele diversity among the Beijing family of isolates, which is highly prevalent in Asia and the countries of the former Soviet Union [169]. While the average diversity of alleles at individual MIRU loci is as high as HGI = 0.74 in some isolate populations [124], the diversity of alleles at many individual MIRU loci is below HGI = 0.15 among isolates of the Beijing family [170]. The Beijing family has been associated with a number of outbreaks of multi-drug resistant (MDR) TB in recent years [169], and studies in animal models have suggested that it may be associated with increased virulence [171]. A recently proposed global standard MIRU typing panel, however, did not include MIRU loci with high diversity in Beijing family isolates, although a number of such loci have been identified [123, 137, 154]. In order to be valuable as a public health tool, any global standard typing panel will have to effectively discriminate isolates of the Beijing family.

While the problem of low MIRU allele diversity is best characterized in the Beijing family, allele diversity may vary substantially across TB strain families [125] . The diversity of proposed typing panels is routinely assessed using highly selected collections of TB isolates from diverse geographic locations [123, 157]. As isolates in any single geographically localized population will be considerably less diverse than such a collection, and will often have a clonal, highly homogeneous population structure, it may be more reasonable to assess the diversity of proposed typing panels in population-based collections drawn from a single geographic location, which better reflects the context in which TB typing tools are most often employed.

In observational studies, it is not possible to accurately identify and unambiguously classify the transmission relationships between cases. The major strength of this investigation is it's reliance on an ABM of TB transmission, which simulates a "world" in which the true status of each case is known, and transmission relationships between cases can be characterized unambiguously. This ABM was specified to best represent the transmission and natural history of TB, as well as social and vital dynamic processes, of three real populations. Such model representations of the "real world" however, are approximations of reality, and are limited by the assumptions that they make about the systems they represent.

Wherever possible, we based our model formulation on available data and well-vetted theoretical models of TB transmission and pathogenesis, and measurements of host population processes. For some components of the system we modeled, however, such as the probability of transmission given contact between an infectious and susceptible individual, little data is available, and the model was informed using a best-estimate from the available literature. Given such uncertainty, a critical component of our investigation is a sensitivity and uncertainty analysis to determine how sensitive our key outcome measures are to variation in the parameters that govern key model processes. The analysis presented here represents only an initial sensitivity and uncertainty analysis, and a more rigorous evaluation in the future is essential to interpreting our results with confidence.

Our sensitivity analysis revealed that parameters reflecting the historic patterns of disease transmission in a population were strongly associated with the validity of clustering as a marker of cases linked by recent transmission. These parameters: the annual risk of infection, yearly trends in the annual risk of infection, and the amount of "historic clustering" (the proportion of latently infected individuals in a population who are linked by transmission to another individual in that population), influence the proportion of latently-infected individuals in a population, the age-distribution of those cases, and the distribution of distinct *M. tuberculosis* strains among latently infected individuals. The annual risk of infection, while not perfect, is consistently estimated in many populations, using well-tested methods [108]. By contrast, the level of "historic clustering" is uncertain, and estimates based on current levels of clustering can be made only in settings where molecular typing investigations have been previously undertaken.

Our initial sensitivity analysis did not identify any association between either the mutation rate during latency or the weekly recovery probability and clustering. However, both of these parameters were evaluated in the context of a transmission scenario based on the population of Arkansas. In this low-incidence setting, the population of infecting isolates will be small, and low numbers of cases mean that measures are often unstable. It is likely that, in a setting with a higher burden of TB, both of these parameters may influence both the level of clustering and the validity of molecular clustering estimates.

A major assumption of our model is that all incident cases of respiratory TB will be reported and that all cases will yield a viable culture from which a molecular type can be generated. Perfect case ascertainment, however, is unrealistic: even if every incident case were identified, some proportion will not yield a viable culture. Incomplete sampling will result in the misclassification of some clustered cases as unique [79]. Our model estimates, therefore, will be higher than empiric measurements from population-based genotyping studies.

While this report presents analysis only of the independent effects of these individual parameters, it will be important to investigate associations between these and other key parameters and the impact of these individual and joint relationships on the key results presented here. Additionally it will be essential to further characterize our key results in the three population-specific transmission scenarios we considered. For example, we considered the impact of the number of loci included in a typing panel, and the stability of individual typing loci, in a hypothetical transmission scenario, which did not provide insight into the influence of diverse population factors on this relationship. Further, population-specific simulation experiments will be necessary to better understand how important this trade-off may be in real transmission-scenarios, and whether it is more relevant in some transmission contexts than others. At the same time, a number of the parameters we considered only in a sensitivity analysis warrant further investigation in their own right, such as the population-prevalence of HIV and the average duration of infectivity.  These parameters may also have important interactions with the key relationships with other model parameters which were not identified in the univariate analysis presented here.

In conclusion, this is the first investigation directly assessing the relationship between diversity and the validity of molecular typing as a marker of recent transmission, and the first model-based evaluation of the sensitivity, specificity, PPV, and NPV of molecular typing in population-based epidemiologic studies.  We have demonstrated the relationship between diversity in typing markers and key measures of the validity of molecular typing data. While far from conclusive, these results provide key insights that may inform the interpretation of results from population-based molecular typing studies, and contribute to the development of a typing system that makes most effective use of current knowledge.

# Tables

**Table 4.1.** Parameter definitions for the model

| **Parameter** | **Value** | **Unit** | **Reference** |
|---|---|---|---|
| *I. Parameters governing population size and social contact structure* | | | |
| Initial Host Population Size | 50,000 | agents | |
| Average Neighborhood Size | 500 | agents | Local neighborhood size as defined in some studies of local neighborhood and health [159] |
| Average Household Size | 2.8 | agents | Based on 2000 US Census |
| Average Number of Friends | 10 | agents | Based on the number of non-household contacts identified by tuberculosis cases in during contact-tracing investigations in Arkansas between 1996 and 2000. (Average of 10, range of 0 to 112) |
| HIV prevalence | 0.0014 | Prevalence | Based on 2000 CDC HIV/AIDS Surveillance estimate for Arkansas[9] |
| Probability of within-neighborhood casual contact | 0.02 | | Estimated based on assumptions that casual transmission, while possible, occurs relatively infrequently. |
| Probability of outside of neighborhood casual contact | 0.01 | | |
| *II. Parameters governing the initial specification of prevalent infections* | | | |
| Annual Risk of Infection Percent (ARI%) | 0.0015 | | Based on unpublished data previously cited [89] |
| Annual trend in ARI% prior to initiation of model (ARI%trend) | 0.015 | | Based on unpublished data previously cited [89] |
| Initial prevalence of active infections | 1.5 per 100,000 | | Based on the annual incidence of active respiratory TB disease and the average duration of disease |
| Proportion of initially specified infections (active and latent) that are linked by | 0.3 | | Based on data from Arkansas reviewed in chapter III of this |

| | | | | |
|---|---|---|---|---|
| transmission to another infected case in the population. | | | | dissertation. |
| *III. Parameters governing host population vital dynamics* | | | | |
| Birth Rate | | 0.0331/year for agents aged 15-44 | | Based on the general fertility rate in the US for the year 2003. Modified (divided by half) as gender is not specified in model |
| Death Rate | | | | Modified from US Vital Statistics: Death rates by age and sex in the US, 1995. Averaged across gender. |
| | *Group* | | | |
| | Age 0 to 9 | 0.0007 | /year | |
| | Age 10 to 19 | 0.0004 | /year | |
| | Age 20 to 44 | 0.003 | /year | |
| | Age 45 to 64 | 0.02 | /year | |
| | Age 65 and up | 0.1 | /year | |
| Immigration rate | | 0.15 | /year | Based on 2000 US Census figures for Arkansas: 1.5% of the total AR population is foreign-born and entered between 1990 and 2000, giving a rate of 0.15% per year. |
| Emigration rate | | 0.15 | /year | Set to balance immigration |
| *IV. Parameters governing the risk of transmission given contact* | | | | |
| Probability of transmission given contact | | | | [163] |
| | Household contacts | 0.06 | /week | |
| | Non-household contacts | 0.001 | /week | |
| Risk of transmission from smear negative case, relative to smear positive cases. | | 0.2 | | [164], [15] |
| *V. Risk of progression to active disease following infection* | | | | |
| Following Primary Infection | | | | [81, 165] |
| Within 1st year of infection | | | | |

|  |  |  |
|---|---|---|
| Age 0 – 10 | 2.48 | /year |
| Age 11-19 | 5.57 | /year |
| Age 20 up | 8.66 | /year |
| Within 2nd year of infection |  |  |
| Age 0 – 10 | 1.02 | /year |
| Age 11-19 | 2.28 | /year |
| Age 20 up | 3.55 | /year |
| Within 3rd year of infection |  |  |
| Age 0 – 10 | 0.32 | /year |
| Age 11-19 | 0.72 | /year |
| Age 20 up | 1.13 | /year |
| Within 4th year of infection |  |  |
| Age 0 – 10 | 0.21 | /year |
| Age 11-19 | 0.48 | /year |
| Age 20 up | 0.74 | /year |
| Within 5th year of infection |  |  |
| Age 0 – 10 | 0.07 | /year |
| Age 11-19 | 0.16 | /year |
| Age 20 up | 0.24 | /year |
| Following re-infection |  |  |
| Within 1st year of infection |  |  |
| Age 0 – 10 | 4.25 | /year |
| Age 11-19 | 4.68 | /year |
| Age 20 up | 5.11 | /year |
| Within 2nd year of infection |  |  |
| Age 0 – 10 | 1.74 | /year |
| Age 11-19 | 1.92 | /year |
| Age 20 up | 2.10 | /year |
| Within 3rd year of infection |  |  |
| Age 0 – 10 | 0.55 | /year |
| Age 11-19 | 0.61 | /year |
| Age 20 up | 0.66 | /year |
| Within 4th year of infection |  |  |

| | | | |
|---|---|---|---|
| Age 0 – 10 | 0.37 | /year | |
| Age 11-19 | 0.40 | /year | |
| Age 20 up | 0.44 | /year | |
| Within 5th year of infection | | | |
| Age 0 – 10 | 0.12 | /year | |
| Age 11-19 | 0.13 | /year | |
| Age 20 up | 0.14 | /year | |
| Beyond 6 years after infection or re-infection (latent reactivation) | | | |
| Age 0 – 10 | 9.8 x 10$^{-8}$ | /year | |
| Age 11-19 | 0.0150 | /year | |
| Age 20 up | 0.030 | /year | |
| Probability that an incident case is sputum-smear positive | | | Based on data from Norway, 1951-69, as previously described [81] |
| Age 0 – 10 | 0.1 | /case | |
| Age 11-19 | 0.4 | /case | |
| Age 20 up | 0.8 | /case | |
| Relapse probability following recovery from active disease | 0.0359 | /year | [17] |
| Probability of death from TB during active disease | | | Based on age specific case fatality rates, US, 1989. |
| Age 0 -14 | 0.0005 | | |
| Age 15-19 | 0.00063 | | |
| Age 20-24 | 0.0013 | | |
| Age 25-44 | 0.0032 | | |
| Age 45-54 | 0.0045 | | |
| Age 55-64 | 0.0066 | | |
| Age 65+ | 0.0123 | | |
| *V. Molecular typing technique and bacterial alleles at typing loci.* | | | |
| Number of Loci in typing panel | 12 | | Based on current standard MIRU typing panel [36] |
| Weekly rate of allele change observed in | 5.249E- | | Estimated from [160] |

| | | |
|---|---|---|
| serial samples | 05 | |
| Relative rate of allele change during latent infection | 0 | Based on limited available evidence (see appendix II), plan to evaluate. |
| Average individual allele diversity, based on HGI | 0.74 | Highest average identified in non-Beijing isolates [124] |
| Maximum number of alleles | 14 | Estimate from available literature [124, 125, 137, 170] |

**Table 4.2**. Parameters to generate population-specific transmission scenarios

| | Arkansas | Malawi | Afghanistan |
|---|---|---|---|
| ++Age distribution | | | |
| 0-14 | 9 | 49.7 | 44.6 |
| 15-64 | 77.1 | 47.7 | 53.0 |
| 65+ | 13.9 | 2.6 | 2.4 |
| Birth Rate<br>(per 1000 individuals of reproductive age) | 33 | 970 | 1034 |
| †ARI% | 0.0015 | 0.015 | 0.03 |
| †ARI trend | 0.015 | 0.01 | 0 |
| HIV prevalence | 0.0014 | †0.1492 | †0.0001 |
| *Initial clustering | 0.3 | 0.7 | 0.5 |
| ‡Average household size | 2.8 | 4.3 | 8.0 |
| †Weekly probability of recovery | 0.060 | 0.022 | 0.020 |

† Estimate used in prior ABM of TB[89]

* Estimates for initial clustering in Arkansas based on molecular epidemiologic data included in this dissertation. Estimates for Malawi were based on data reported from a molecular epidemiologic investigation of TB in that country [152]. Estimate for Afghanistan was based on a best guess, as no data on recent transmission or molecular clustering are available.

++ Age distribution and birth rate for Arkansas is based on 2000 US Census.
(DP-1. Profile of General Demographic Characteristics: 2000
Data Set: Census 2000 Summary File 1 (SF 1) 100-Percent Data)
Estimates for Malawi and Afghanistan based on data compiled by UNICEF
(http://www.unicef.org/infobycountry/index.html)

**Table 4.3.** Key parameters identified for initial sensitivity analysis.

| Host-related parameters | | | |
|---|---|---|---|
| Parameter | Description | Model value | Range evaluated |
| tP | Weekly probability of transmission from a smear-positive infectious host to a susceptible household member | 0.06 | 0.01 to 1.0 |
| ARI% | Annual risk of infection experienced by the simulated population each year prior to the initiation of the model | 0.0015 (for Arkansas) | 0 to 0.004 |
| ARItrend | Annual decline in the ARI% experienced by the simulated population prior to initiation of the model | 0.01 | 0 to 0.04 |
| HIVp | Prevalence of HIV infection in the population | 0.1492 (for Malawi) | 0.0001 to 0.1492 |
| reProb | Weekly probability that an actively infected individual recovers to a latent state | 0.06 (for Arkansas) | 0.01 to 0.1 |
| Microbe-related parameters | | | |
| mNA | Maximum number of alleles allowed at any single typing locus | 14 | 5 to 50 |
| mRl | Rate of allele change at an individual loci during latent infection, relative to rate during active infection | 0 | 0.1, 1 |

**Table 4.4** Comparison of model-generated and empirically measured TB transmission statistics for three population-specific transmission settings.

| | Incidence (per 100,000) | Prevalence | ARI% | % clustered | PPV |
|---|---|---|---|---|---|
| **Arkansas** | | | | | |
| Measured | 5.1 | 0.05 | 0.005 | 47.1* | 0.320* |
| Model | 4.6 | 0.043 | 0.007 | 54.7 | 0.139 |
| **Malawi** | | | | | |
| Measured | 401 | 0.299 | 0.015 | 72.0* | -- |
| Model | 569 | 0.276 | 0.029 | 87.9 | 0.284 |
| **Afghanistan** | | | | | |
| Measured | 333 | 0.410 | 0.030 | -- | -- |
| Model | 160 | 0.394 | 0.015 | 50.3 | 0.402 |

* Measured clustering and PPV statistics are based on typing with IS*6110* RFLP, which is considered more discriminatory than the 12-loci MIRU panel the model results presented here were based on.

# Figures

**Figure 4.1.** Variability in the number of repetitive units at molecular typing loci based on MIRU-VNTR. In ABM, the specific allele value for each locus corresponds to the number of repetitive elements present. Figure modified from Frothingham, 1998 [172].

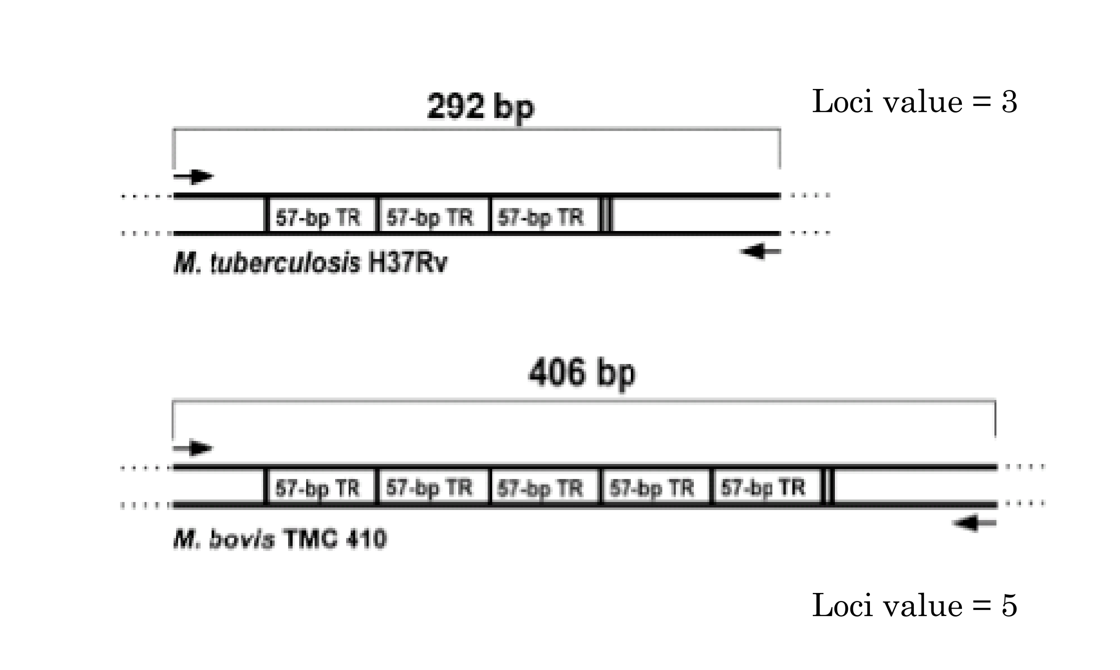


**Figure 4.2.** Schematic illustration of model flows. Infection states, represented by variables in the model, are represented here by compartments, similarly to a stock and flow diagram representing an equation-based model.

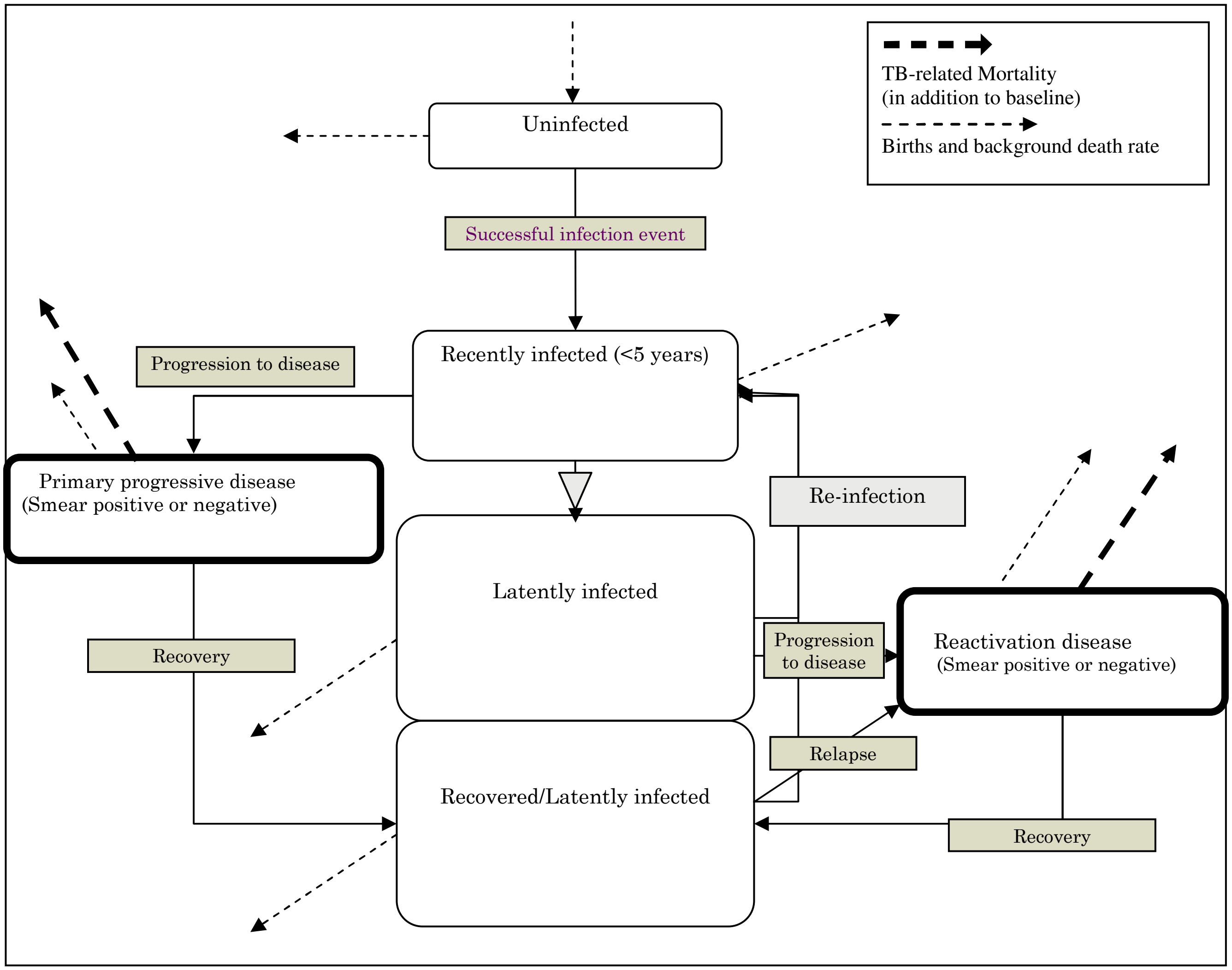
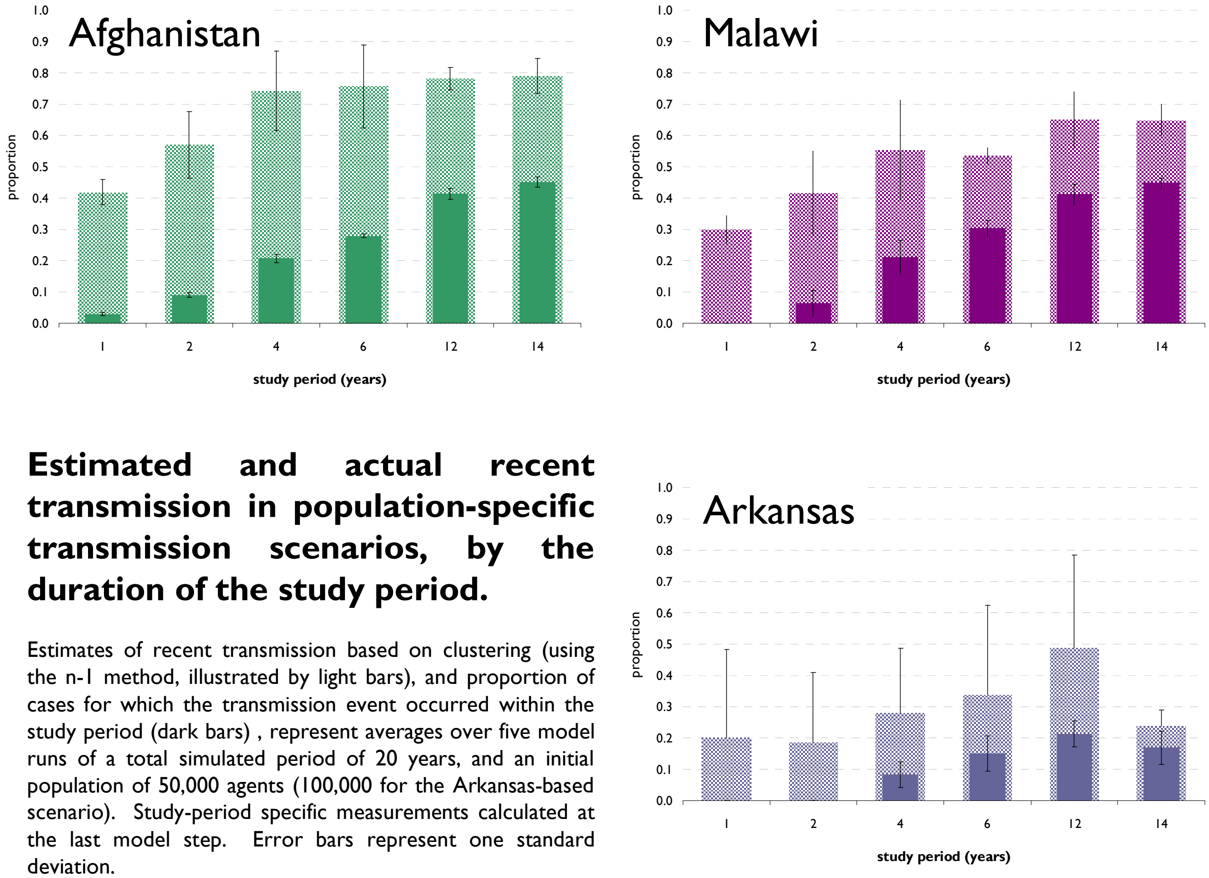

TB-related Mortality
(in addition to baseline)
Births and background death rate
Uninfected
Successful infection event
Recently infected (<5 years)
Progression to disease
Primary progressive disease
(Smear positive or negative)
Re-infection
Latently infected
Progression
to disease
Reactivation disease
(Smear positive or negative)
Recovery
Relapse
Recovered/Latently infected
Recovery

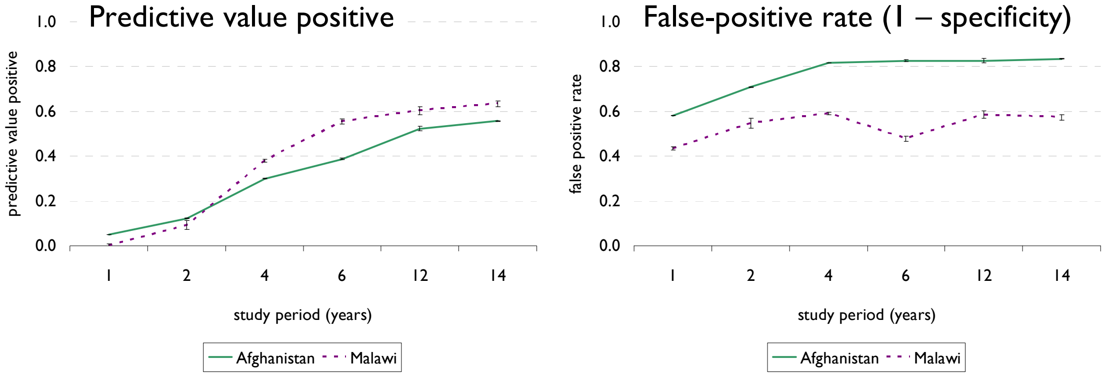
**Figure 4.3.** Estimated and actual recent transmission by population and study duration. Comparison of the proportion of cases truly related by recent transmission to clustering-based estimates of recent transmission in three population-specific transmission scenarios represented in an agent-based computer simulation.



## Estimated and actual recent transmission in population-specific transmission scenarios, by the duration of the study period.

Estimates of recent transmission based on clustering (using the n-1 method, illustrated by light bars), and proportion of cases for which the transmission event occurred within the study period (dark bars) , represent averages over five model runs of a total simulated period of 20 years, and an initial population of 50,000 agents (100,000 for the Arkansas-based scenario). Study-period specific measurements calculated at the last model step. Error bars represent one standard deviation.

**Figure 4.4.** Validity of clustering by population and study duration. Study duration and the validity of clustering as a measure of transmission in two population-specific transmission scenarios represented in an agent-based computer simulation.



### Impact of study duration on the validity of clustering as a measure of transmission, in transmission scenarios specific to Malawi and Afghanistan.

Calculations of the predictive value positive and false positive rate (1-specificity) represent averages over five model runs of a total simulated period of 20 years, and an initial population of 50,000 agents . Study-duration specific measurements were calculated at the end of each 20 year simulation. Error bars represent one standard deviation.

**Figure 4.5.** Rate of allele change and the validity of clustering. Allele stability and validity of clustering as a measure of transmission in a hypothetical transmission scenario represented in an agent-based computer simulation.

## Rate of allele change and the validity of clustering as a measure of transmission, using a 12 loci typing pattern.

Rate of allele change corresponds to the weekly probability of marker change in at individual typing loci over the course of active infection. Measures presented represent averages over five model runs of a total simulated period of 20 years, and an initial population of 50,000 agents. All measurements calculated at the last model step, over a study period equivalent to 4 years. Error bars represent one standard deviation.

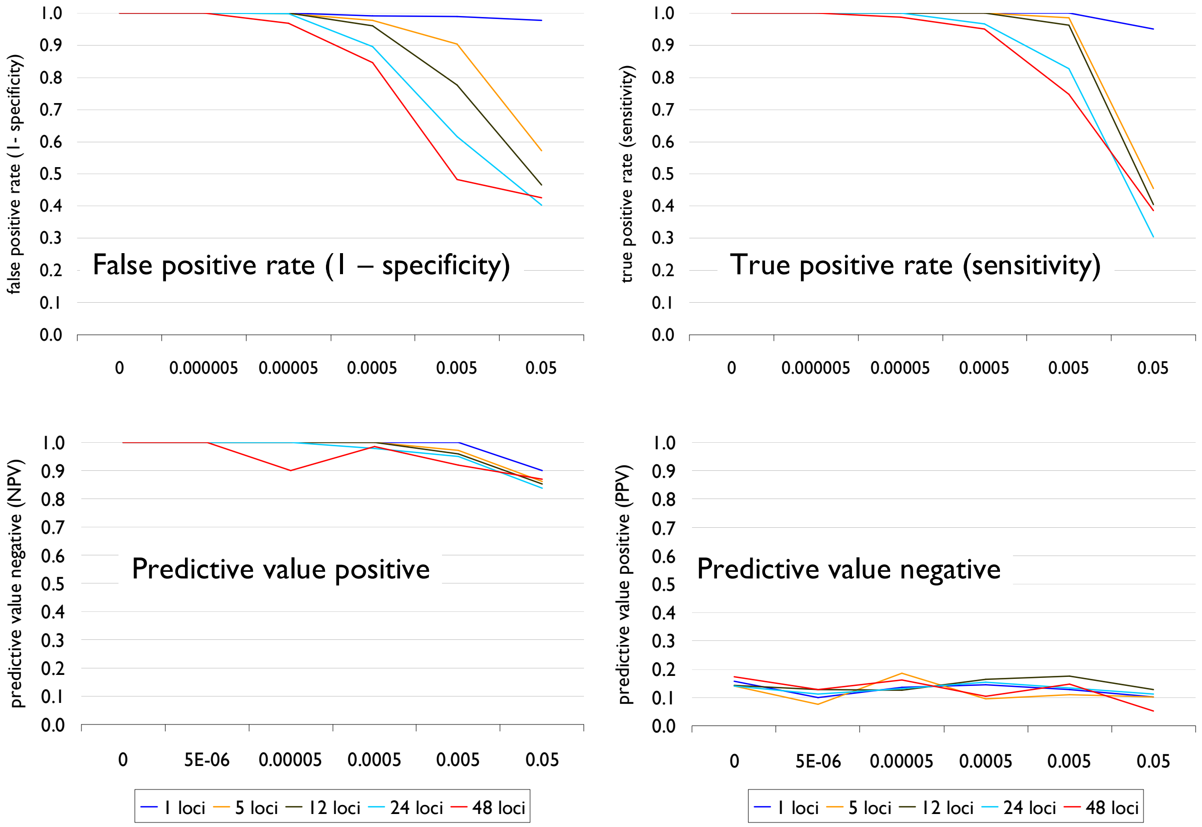
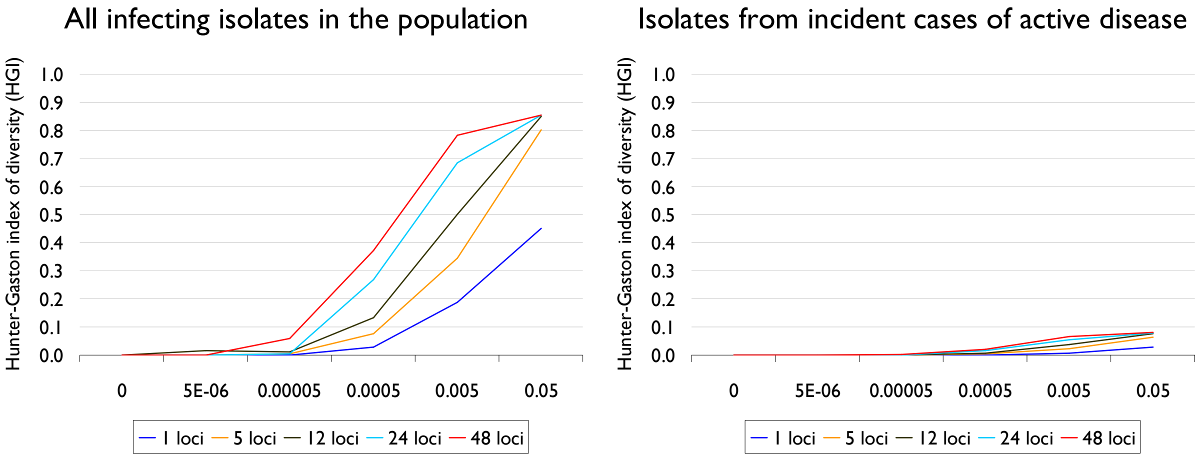


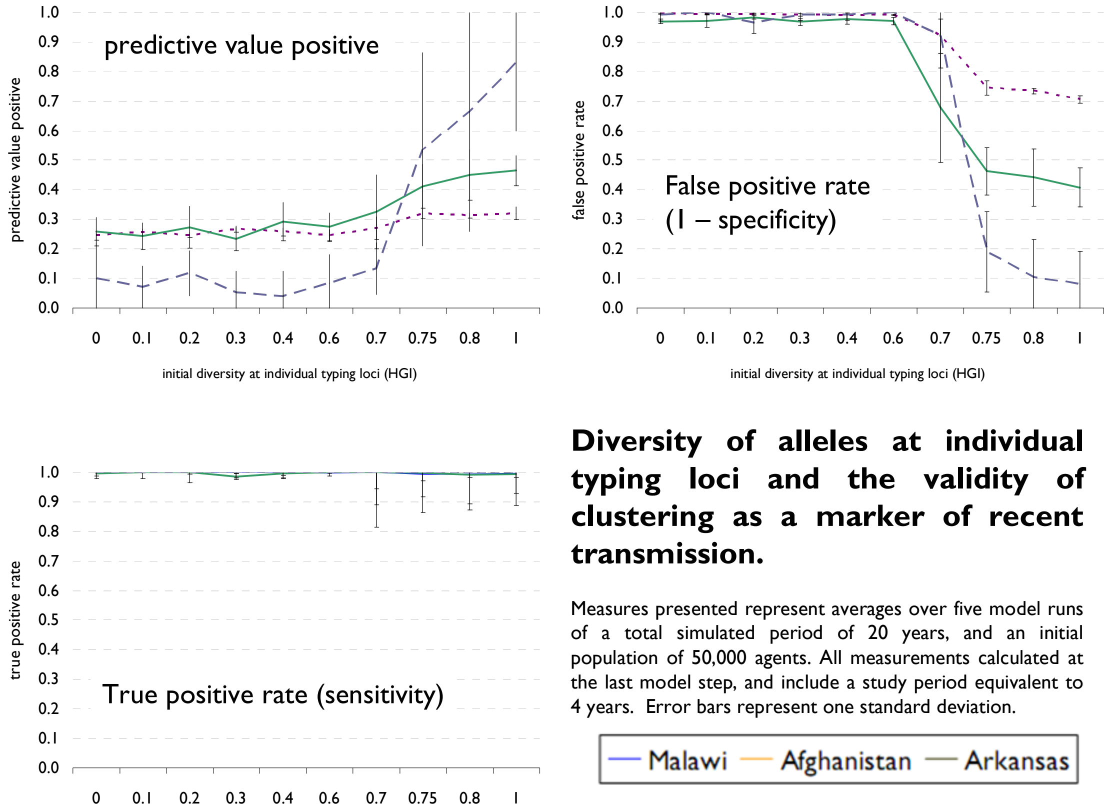
**Figure 4.6.** Rate of allele change and typing pattern diversity. Allele stability and the diversity of molecular typing patterns observed in a hypothetical transmission scenario represented in an agent-based computer simulation.



## Rate of allele change and the diversity of "typing patterns" observed, using a 12 loci typing pattern.

Rate of allele change corresponds to the weekly probability of marker change in at individual typing loci over the course of active infection. Measures presented represent averages over five model runs of a total simulated period of 20 years, and an initial population of 50,000 agents. All measurements calculated at the last model step, over a study period equivalent to 4 years. Error bars represent one standard deviation.

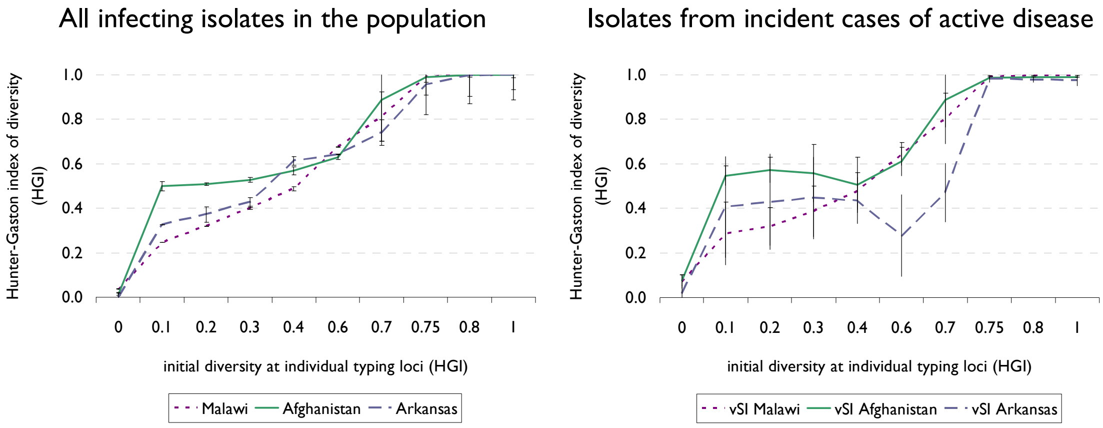
**Figure 4.7.** Allele diversity and the validity of clustering. Average individual allele diversity and the validity of clustering as a marker of recent transmission in three population-specific transmission scenarios represented in an agent-based computer simulation.



**Diversity of alleles at individual typing loci and the validity of clustering as a marker of recent transmission.**

Measures presented represent averages over five model runs of a total simulated period of 20 years, and an initial population of 50,000 agents. All measurements calculated at the last model step, and include a study period equivalent to 4 years. Error bars represent one standard deviation.

**Figure 4.8.** Allele diversity and typing pattern diversity. Average individual allele diversity and the diversity of typing patterns seen in three population-specific transmission scenarios represented in an agent-based computer simulation.



All infecting isolates in the population

Isolates from incident cases of active disease

## Diversity of alleles at individual typing loci and the diversity of "typing patterns" observed.

Measures presented represent averages over five model runs of a total simulated period of 20 years, and an initial population of 50,000 agents. All measurements calculated at the last model step, over a study period equivalent to 4 years. Error bars represent one standard deviation.

**Appendix.** Explicit description of model rules in outline form

The Model has the following components:

(1) The entities of the model, consisting of discrete human agents and discrete *Mycobacterium tuberculosis* agents (with each *M. tuberculosis* agent representing a clonal population of bacteria infecting one human case).

(2) The contact structure that governs the interaction of the agents, as represented by a group of neighborhoods, and households within those neighborhoods.

(3) The rules that govern the dynamics of the system, representing the social and biological interactions of the entities

(4) The time-scales on which these rules are executed

## I. Initialization: conditions at the start of a simulation

a. Create a population of discrete human hosts (default = 20,000)

   i. Randomly assign each host an "age" based on the age structure of the state of Arkansas at the 2000 Census, or the age distribution specified.

   ii. Create a number of neighborhoods sufficient to divide the initial number of human hosts into approximately evenly sized neighborhoods of 500 agents each, randomly assign agents to a neighborhood (neighborhood size based on prior studies of neighborhood effects on health[159]).

   iii. Create households within each neighborhood with an average household size normally distributed around the specified average household size (default = 5 agents) (Mean = average household size, SD = average household size/3), until each agent assigned to a household.

   iv. Check age distributions within households, re-distributing as necessary to ensure no household is made up entirely of children.

   v. Assign each host a group of friends, randomly chosen from the host's neighborhood and age-group (ages 0-9, 10-19, 20-44, 45-64, and 65+).

b. Based on the specified age of each initial agent, calculate the probability of infection for each calendar year of life experienced by that agent (according to ARI% and ARI % trend). If multiple pre-model infection events occur, consider only the most recent infection. For each initially infected agent:

i. Infect with a unique *Mtb* "isolate" (one discrete entity representing the clonal population of infecting bacteria).

ii. Randomly assign each *Mtb* isolate to a strain: if the initial level of clustering is 1, all isolates are assigned to the same strain, if 0, all isolates are assigned to a different strain. The range of iClust values between 0 and 1 assigns some proportion of isolates to the same strain and some proportion to unique strains.

iii. If a given individual is selected to be initially infected with a clustered isolate, check if any other agents in the same household have already been infected. If yes, infect the first agent with the same strain already present in the household. If no, infect the first agent with the most recently assigned strain.

## II. Overview: Timing and Order of Events

a. All rules that govern model events are executed in 1-week time intervals:

   i. Vital Dynamics of the human host population

      1. Aging
      2. Births
      3. Deaths
      4. Immigration
      5. Emigration
      6. Maintenance of friendships

   ii. Transmission of *M. tuberculosis*

      1. Contact events between human hosts

         a. Contact within households
         b. Contact between friends in a neighborhood
         c. Casual contact within a neighborhood
         d. Casual Contact outside of neighborhood

      2. Transmission according to specified probability

   iii. Disease progression of infected individuals

   iv. Marker mutation at bacterial loci specified by the typing system

   v. Incident infections and cases of active disease are recorded, and infection, disease, and strain-clustering measures are calculated.

   vi. HIV prevalence is maintained (by brute force, not transmission: if the HIV prevalence dips below the level specified, random non-HIV infected individuals are converted to HIV positive status until the desired prevalence is achieved.

   vii. All time-keeping variables are updated.

      1. time since infection
      2. time since development of disease

**III. Rules governing vital dynamics of human host population**

a. Aging
   i. After 52 steps in one age (measured by year), human host moves to next age.

b. Births
   i. Every human host aged 15-44 may generate a new human host according to the specified fertility rate
      1. Based on general fertility rate statistics, divided by 2 as gender is not represented in the model.
      2. New human hosts are
         a. Uninfected
         b. Age 0
         c. Same household and neighborhood as parent

c. Deaths (general)
   i. Every human host may die at any time step, according to an age-specific probability
      1. Removed from all lists in model

d. Immigration
   i. For each human host currently in the model, a new agent will enter the model according to the specified immigration probability.
      1. Move to neighborhood of the human host whose presence ‘triggered’ the immigration event
         a. if no other immigrants in the neighborhood, start a new household
         b. if other immigrant households in the neighborhood, 50% probability of starting a new household, 50% probability of moving into an already established immigrant household.
      2. Latently infected at entry according to the specified probability that an immigrant is infected
         a. Infecting *M.tuberculosis* is always of a unique strain
         b. Allele diversity is the same as that specified for the simulated population.

c. Time since infection assigned as a function of age, randomly assigned as a function of the immigrant host's age from a normal distribution with a mean of ¼ (age) and a standard deviation of ¼ (age).

e. Emigration

i. Every human host may leave the model at any time step, according to the specified probability of emigration.

f. Maintenance of friendships

i. Friendships are maintained throughout the lifetime of the human host, ending only at death or immigration.

ii. All friendships are bi-directional.

iii. Any host whose # of friends dips below a minimum allowed range (set for each agent) will make a new friend with another agent in the same neighborhood.

iv. Any host below a maximum allowed range will become friends with any agent who "tries" to make friends.

v. New human hosts (immigrants and births) randomly make friends with other hosts in the same age group and neighborhood that have not exceeded their maximum range of friends.

**IV. Rules governing *M. tuberculosis* transmission**

a. Effective contact events between human hosts

i. Contact within households

1. At each time step, each human host with active disease contacts with every individual in the same household

ii. Contact between friends

1. At each time step, each human host with active disease contacts each of its friends.

iii. Casual contact within neighborhoods

1. At each time step, each human host with active disease comes into contact with random members of the same neighborhood according the probability of random neighborhood contact.

iv. Casual contact between neighborhoods

1. At each time step, each human host with active disease may visit another neighborhood in the model with the specified probability of casual outside of neighborhood contact
   a. Given a visit to another neighborhood, each actively diseased host will come into contact with members of that neighborhood according to the same probability of casual contact within the hosts own neighborhood.

b. Probability of transmission given contact
   i. Given a contact event between a host with active disease and another host in the model, transmission may occur according to the following rules:
      1. Diseased hosts with sputum smear positive disease will transmit infection according to the specified transmission probability
      2. Diseased individuals with sputum smear negative disease will transmit infection at a reduced probability (20% of the full transmission probability)
      3. Hosts with active disease, or who have been infected within the last 5 years, may not be infected.
      4. Transmission may occur between diseased hosts and hosts in the uninfected, latently infected, or recovered states.
         a. Transmission between household contacts occurs at the full transmission probability.
         b. Transmission between friends and casual contacts (within and outside of the neighborhood) occurs at a specified fraction of the transmission probability.

c. Transmission
   i. If a transmission event occurs, a new *M. tuberculosis* entity is created. This entity:
      1. Is assigned to the same strain as the infecting strain
      2. Is assigned the same "molecular fingerprint", and the same allele value for each typing loci, as the infecting strain
      3. Variable identifying this isolate is passed to the infected agent
      4. "Time since infection" is set to 0

## V. Rules governing disease progression

a. Once infected, human hosts may develop respiratory tuberculosis disease (non-respiratory disease is not considered) according to the following rules:

    i. Individuals infected (or re-infected) within the last 5 years may:

        1. Develop primary disease

            a. Dependent on

                i. age,

                ii. type of infection (primary vs. re-infection),

                iii. year since infection.

            b. A multiply infected individual is only at risk of developing disease from the most recent infecting strain

            c. A newly diseased host will be assigned to a "sputum smear positive" or "sputum smear negative" state according to an age-specific probability.

            d. A newly diseased host will be infectious only with the most recently infecting strain.

    ii. Individuals infected 5 years, who did not develop primary disease in this interval, progress to latent disease at the end of the 5th year following infection.

    iii. Lantently infected individuals (or re-infected) 6 or more years in the past may:

        1. Develop reactivation disease

            a. Dependent on

                i. Age

                ii. HIV status

            b. A multiply infected individual is only at risk of developing disease from the most recent infecting strain

            c. A newly diseased host will be assigned to a "sputum smear positive" or "sputum smear negative" state according to an age and HIV status -specific probability.

            d. A newly diseased host will be infectious only with the most recently infecting strain.

        2. Become re-infected and move to the "recently infected" state.

iv. Individuals with active disease may:
    1. Recover from disease according to the specified recovery probability,
    2. Die from TB disease according to an age specific probability.

v. Individuals recovered from active disease may:
    1. Develop relapse disease (with the same strain that last caused disease) according to the specified relapse probability
    2. Become re-infected and move to the "recently infected" state.

## VI. Measures of infection, disease, and strain clustering

a. At each step of the model:

i. New infection and disease progression events are recorded

ii. Infection-related statistics are calculated;
    1. Annual risk of infection
    2. Prevalence of latent infection
    3. Total prevalence of infection

iii. Disease-related statistics are calculated:
    1. Incidence of disease (cases per 100,000 human hosts)
    2. Prevalence of active disease
    3. Risk of disease following infection
    4. Proportion of disease resulting from re-infection

iv. Strain-clustering statistics are calculated according to the specified study period:
    1. Proportion of cases that are "clustered" by strain.
        a. "n" method
        b. "n-1" method
    2. PPV, NPV, Sensitivity, and Specificity of "clustering" in the study period for a transmission linkage with a case in the same cluster.
        a. Also calculated by age group
    3. Diversity calculations, using the Hunter-Gaston index of diversity [143] are calculated for;
        a. All infections in the model
        b. All isolates that have caused active disease

c. Average individual allele diversity

# Chapter V

## Conclusion

Molecular typing is a powerful tool for the study of the transmission and epidemiology of TB, and is increasingly integral to TB control programs. While much of this dissertation work takes a critical eye towards the inferences that are drawn from molecular typing data, the ultimate goal of this work is to contribute to refined typing tools and analytic techniques that will allow investigators and public health practitioners to most effectively employ this tool. By identifying key factors that may compromise the assumptions on which molecular typing applications are based, this work may contribute to the development of yet more powerful typing strategies.

The interpretation of molecular typing is most often boiled down to a simple binary classification: a case isolate is determined to be either clustered, or unique. This simple categorization belies the complex interplay of systems that influence whether or not an individual case-isolate is clustered and, I believe, prevents many practitioners from questioning the validity of the inferences they draw from molecular typing data.

The patterns generated using molecular typing tools result from a complex interaction of host and pathogen populations. Bacterial population genetics and evolution, host demographics, and host social contact patterns will all influence whether or not a case is clustered – and this is before considering the molecular typing system itself. Added to this complexity are the myriad complications of TB epidemiology. Identifying chains of TB transmission in a population is at its best highly informed guesswork, as clear transmission linkages cannot be unambiguously identified. Given these limitations, it is perhaps less surprising that, despite increasingly sophisticated molecular methods, the relationship between molecular typing data and the epidemiologic relationships it is used to predict remains poorly understood.

Many of the key players in the development and application of TB typing tools focus entirely on one side of the molecular typing equation—as either microbiologists or public health practitioners, their loyalties are often clear. As a result, rigorous evaluations of TB typing tools typically present a narrow perspective. The hope of this dissertation, with its highly interdisciplinary approach, is that it will represent a first step towards bridging this divide.

Using three different analytic approaches, the investigations presented here all cast light on this tight interrelationship between host and pathogen populations, each from a slightly different angle. Taken together, these results demonstrate that patterns in one of these systems cannot be understood without considering the influence of the other.

All three papers included in this dissertation demonstrate the importance of historic patterns in the transmission of *M. tuberculosis* on contemporary patterns in TB disease, and suggests that historic patterns influence the validity of molecular typing measures of recent transmission in a population.

Prior investigators had applied molecular typing to the study of trends in the incidence of active TB disease [67, 69]. The first paper presented in this dissertation, however, was the first to apply molecular typing to the study of TB trends in a rural population, and the first to clearly demonstrate the impact of historic disease patterns on contemporary disease trends. These results have clear implications for the evaluation of TB control programs, where declining rates of disease are generally considered to reflect the successful containment of active transmission. While a highly effective TB control program in Arkansas had been very successful at containing transmission there, our analysis demonstrated that some key populations may still be slipping beneath the radar.
The observation that historic trends project considerable momentum, possibly driving TB disease patterns many years into the future, is an insight that may allow public health practitioners to more accurately project future disease patterns, and to design intervention strategies and allocate resources accordingly.

The second paper of this dissertation identified a number of factors that were correlated with the validity of molecular typing results. Among the most interesting of the factors we identified was the impact of a family of endemic strains, which may be closely related to the major finding from the first paper. A family of endemic strains, which appear to have been historically transmitted in Arkansas, showed an increasing prevalence with increasing age. At the same time, this strain family appeared to be independently associated with very low levels of typing diversity and a low predictive value of typing for recent transmission. It seems plausible that the same historic patterns responsible for the sharp cohort effect we observed in the rate of reactivation disease may also be associated with the distribution of this endemic strain family.

Finally, our observations related to diversity and the validity of typing results is likely to be one of the most significant contributions of this dissertation work. With the diversity that can be reasonably achieved by a typing system seemingly every increasing, it seems rational to clarify the level of diversity that an ideal typing tool should attain. Previously, the only metrics for typing diversity were arbitrary [143, 157], and these were only infrequently cited in the evaluation of TB typing markers, even when diversity was calculated [124, 137, 170]. While this dissertation work represents only a first step towards clarifying the relationship between diversity and typing resolution, we have identified what appears to be a critical value in the level of individual MIRU allele diversity. Our early model experiments suggest that this value may delineate markers that contribute only negligibly to the ability of a typing panel to discern epidemiologic relationships from markers that contribute substantially to this aim.

Along with our model observations of the impact of allele diversity, numerous observations of low levels of MIRU allele diversity in specific populations [92, 124, 170], suggest that it may not be possible to identify a single typing panel that is optimal for all populations. We did not have MIRU typing results for the isolates we analyzed from Arkansas. It would be interesting to see if the low levels of diversity we observed in the X2 Spoligotype family using IS*6110* RFLP corresponded to a low level of diversity by MIRU. As molecular typing is more widely used to guide TB control programs, it will be critically important to characterize the association of different strain families with diversity by available typing tools, as well as to identify the composition of the *M. tuberculosis* population in a given setting.

At this time, the use molecular TB typing tools is, with rare exception, limited to resource rich countries. There is a grim irony to this, as 99% of all TB deaths worldwide occur in the world's poorest countries [3]. While TB contact tracing is arguably less important in situations where more fundamental faults of resources and infrastructure may present obstacles to diagnosis and treatment, the availability of molecular typing tools may still provide a critical tool in resource-limited settings. For example, when XDR TB was identified in Uganda in September of 2007, it was not possible to discern whether these resistant strains had developed independently in Uganda, or had been imported from South Africa. Nor was it possible, after a handful of initial reports identified its presence in Uganda, to assess the extent to which XDR TB had spread throughout the country. A major promise of more rapid, efficient, and economic molecular typing tools, such as MIRU-VNTR, is that these technologies may someday be within the reach of the populations that need them the most.

# Appendices

# Appendix 1

## Transmission and natural history of *M. tuberculosis* infection

Representing the transmission of *Mycobacterium tuberculosis* in a simulation model requires a comprehensive understanding of the transmission and natural history of tuberculosis disease. Such an understanding, along with a consideration of the research questions the model will be asked to address, is essential to determine which key features of the transmission system must be represented in the model, and which features may be simplified.

This appendix briefly outlines the major processes important in the transmission of *M. tuberculosis,* as well as the progression to and recovery from active disease. For each process, the current understanding of the process will be described at a level relevant to the representation of this process in the model, and key literature will be briefly reviewed. In the "methods" section of the main section this paper, the implementation of these processes, along with relevant parameter estimates, are described.

### Transmission

The transmission of infection with *M. tuberculosis* is central to the epidemiology of tuberculosis disease. Despite its fundamental importance, the transmission event is perhaps the least understood component of the natural history of TB. This lack of understanding results from the tremendous complexity of the transmission event itself, which depends on a complex array of factors, and from the difficulty in studying infection events on a population level. The high proportion of subclinical *M. tuberculosis* infections, long latency period, and difficulties in unambiguously identifying recently infected individuals have long limited epidemiologic investigations of TB transmission. Assessments of infection risk have relied primarily upon estimation techniques applied to other, more accessible data, such as the age-specific prevalence of latent infection [173, 174].

TB transmission occurs via the airborne route, by the inhalation of microscopic (< 5 µm in diameter) droplet nuclei carrying viable *M. tuberculosis* bacilli [175]. Following expulsion into the environment by an individual with active, infectious TB disease, these droplet nuclei can remain airborne for time spans ranging from minutes to hours. Whether or not an infection occurs when an individual breathes air that has been contaminated by an infectious tuberculosis patient depends on a number of factors, including the infectiousness of the infected (determining the number of viable bacilli expelled into the air), the size, ventilation, and UV exposure of the airspace (determining the density of bacilli in the air and the length of time they remain present and viable), the ability of the bacterium to remain viable in the environment, the susceptibility of the exposed individual, and the duration of the exposure.

The above factors will impact the probability of a transmission event given contact between an infectious and susceptible individual.  Transmission can only occur, however, if a susceptible individual comes into contact with an infectious individual. The probability of such a contact event occurring will depend critically on both the prevalence of active, infectious disease in the population, as well as on the social contact structure which determines the interaction of individuals in the population.

### Exposure: Type and duration of contact

What constitutes an effective contact?

Tuberculosis has long been associated with overcrowded conditions [176], and this association continues to be observed in contemporary settings [177, 178].  Interestingly, while associations have been observed between tuberculosis disease and an increasing number of dwellers per bedroom and a small housing unit size, after adjusting for other factors [178], an association between tuberculosis disease and “district” or “neighborhood” level overcrowding has not been observed in a number of studies that investigated it [178, 179]. These results suggest the importance of close, prolonged contact in determining transmission.  This suggestion is consistent with the historic pattern of TB transmission which informed the current contact tracing protocols central to TB control programs in the United States [47, 58, 59, 180], in which most transmission events occurred within the home and family.

Transmission of *M. tuberculosis* through casual contact has been documented, however [41], and some studies of the molecular epidemiology of tuberculosis have suggested that casual transmission is responsible for a non-negligible proportion of incident cases [181].  As the incidence of tuberculosis in the United States and other developed countries continues to decline, transmission occurring within the home and family may be less important than transmission occurring outside the home [57, 121].

Quantifying the risk of transmission across different types of contact is not possible given the current understanding of the myriad factors influencing transmission. While the risk of transmission has been observed to depend on the duration, frequency, and intimacy of contact, and the setting in which contact occurs (including such factors of the volume of air space and the frequency of air exchange) [175], understanding of the net effect of these factors to determine the relative probability of infection is unclear.

Data from household contact studies provides among the most straightforward data on the risk of infection: the risk of transmission given non-household or casual contact is much more difficult to assess. Even data from household contact investigations can provide ambiguous results of the transmission risk, however: particularly in high-incidence settings, an individual who converts from a negative to positive skin test (suggesting acquisition of *M. tuberculosis* infection) in the course of exposure to an infectious case in the same household may have acquired infection from the household contact, or from an infectious source outside of the household.

In investigations of household contacts, investigators have noted an increasing risk of infection with increasingly intimate contact (defined by the average distance of contact), with 42% of close/intimate contacts, 34% of close/regular contacts, and 13% of a not close/sporadic household contacts having a positive TST at the time of the contact investigation, compared to 16% of a healthy population sample without any household contact. According to only "strongly positive" TST reactions ( ≥20mm), the same study found evidence of recent infection in 27% of very close/intimate contacts, 13% of close/regular contacts, 0% of not close/sporadic contacts, and 0% of the healthy population controls [161].

Data from a number of studies have fairly consistently found an estimated risk of infection from a household contact of approximately 30% over the course of disease [162, 163]. Using contact tracing data that identified non-household contacts as well as contacts, Gryzybowski and colleagues found a risk of infection to non-household contacts of approximately 5% over the course of disease [163].

### Annual Risk of Infection

While the risk of contact by type/duration of exposure is essentially impossible to estimate given current data, the overall population-level risk of tuberculosis infection can be estimated from data that, by comparison, is attainable and relatively unambiguous.

Central to this approach is the idea that the prevalence of tuberculin reactivity for a birth cohort reflects the accumulation of tuberculosis infection since birth. Given consecutive tuberculin survey data from a single cohort (or tuberculin survey data for a single year, with the assumption of a constant infection risk across the lifetime of that cohort), the "Annual Risk of Infection" (ARI) can be estimated [108, 165].

In the United States, available estimates of the ARI have come from tuberculin surveys in military recruits. Estimates for the most recent time period suggest an ARI between 0.04 (White Navy recruits, 1990), and 0.06 (All navy and marine recruits, 1986) [182].

### Re-infection

It was long held that, following a primary infection with *M. tuberculosis*, an individual acquired relative immunity against subsequent re-infection [18]. All subsequent episodes of active disease at any point in an infected individual's life, and at any anatomic site, were considered to be due the reactivation of dormant bacilli of the original infecting strain.

As early as 1976, however, evidence surfaced that called this traditional understanding into question. In that year, using phage typing, an early molecular strain typing technique, Bates [183] reported on a patient with TB infection at multiple sites, from which multiple distinct strains were isolated. Two decades later, as molecular typing techniques for *M. tuberculosis* were refined and came into wider use, evidence of exogenous re-infection, in which an individual previously treated for and apparently cured of an infection with one strain developed active disease with another, distinct strain of *M. tuberculosis*, was reported [12]. In the last decade, investigations of exogenous re-infection have been numerous, and have included both molecular epidemiologic investigations documenting the phenomenon [20, 21, 23], mathematical models illustrating the role that exogenous re-infection has played in temporal trends in the incidence of tuberculosis [81, 165, 184], and experiments in animal models [185, 186].

The relative importance of re-infection to the overall incidence of TB in a population depends upon the risk of infection in that population (itself a function of the prevalence of active, infectious tuberculosis cases), as well as the population prevalence of factors that increase the risk of progression from infection to active disease, such as HIV infection. In the United States and other low-incidence countries where the risk of infection is low, the role of infection also appears to be low – a study of recurrent TB in San Francisco found only 4% of recurrent cases of active TB disease to result from re-infection [17]. As the incidence of TB rises, however, the importance of re-infection rises in tandem: accounting for 10%, 16%, and 33% of recurrent cases in the moderate-incidence countries of Switzerland, Italy, and Spain, respectively, and accounting for 60% of all recurrent cases in the high-incidence country of South Africa [14].

Based on results from experiments in guinea-pigs, which showed that animals previously infected by *M. tuberculosis* that were re-inoculated with the bacilli consistently developed local lesions (suggesting successful infection), but were less likely to suffer haematogenous dissemination following a re-infection event (suggesting a decreased risk of disease), Vynnycky and Fine, in their model of the natural history of tuberculosis infection, assumed that a primary tuberculosis infection did not modify the risk of subsequent infection events, but did alter the risk of progressing to disease following such a re-infection event [81]. They note that this is consistent with an explanation for the protective effect of BCG vaccination given by Sutherland and colleagues [187], which attributed the protective effect not to the prevention of infection, but to protection against subsequent haematogenous dissemination. They did, however, assume that, in the first 5 years following infection where an individual is at high risk of developing subsequent disease, re-infection could not occur. This second assumption was made in order to simplify the model.

## Natural History of tuberculosis infection

### Progression from infection to active disease

Following successful infection with *M. tuberculosis*, the majority of individuals will mount a cell-mediated immune response that successfully contains the infection. These infected, disease-free individuals are then in a state of latent infection: viable organisms remain, and may "reactivate" to cause active disease at a later time. Individuals who are not able to mount a successful immune response following infection will progress directly to disease, often referred to as "primary progressive disease". Although the relative frequency of each of these disease types is difficult to measure, a number of studies have generated estimates of both the relative importance of each type of disease, and the risk of developing each following infection. It is commonly cited that the lifetime risk of developing disease following infection is 5-10%, with roughly half of that risk occurring in the first 2 years after infection. Indeed, these figures are quoted so widely that they seem to be considered common knowledge: the studies from which they were estimated [188, 189] are only infrequently cited in support.

### Annual risk of progression from infection to disease

The annual risk of development of active disease has been observed to vary by age [188], as well as by time sense infection and the host immunity. Sutherland drew on tuberculosis data from Dutch adult males in 1952 to 1970 to estimate age-group specific risks of progression by year since infection [189]. In addition, his was the first model to explore the role of exogenous re-infection, and in addition to risks of disease progression, he estimated the level of protection conferred by an initial infection against subsequent re-infection (Table A.1.1).

While it is thought that the risk of progression might increase in old age, the magnitude of this possible increased risk is unknown, as available data is largely from school age children and younger adult populations. Available estimates of age-specific risks, therefore, are limited to broad age groups.

**Table A.1.1.** Estimates of annual risks of development of tuberculosis and the percentage protection from distant primary infection in the Netherlands. From: Sutherland 1976 [190].

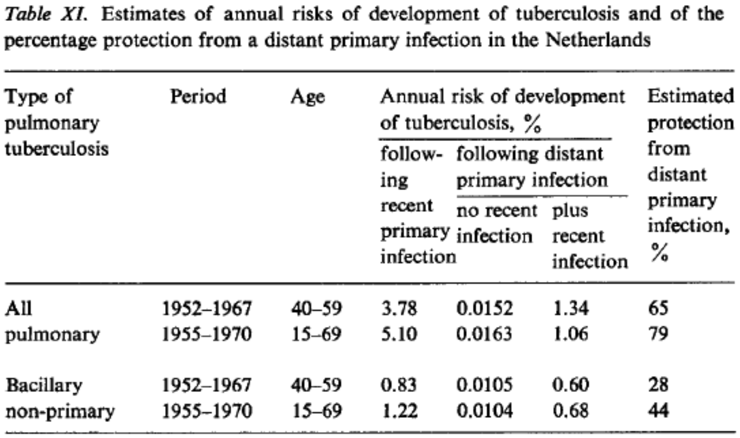


*Table XI.* Estimates of annual risks of development of tuberculosis and of the percentage protection from a distant primary infection in the Netherlands

| Type of pulmonary tuberculosis | Period | Age | Annual risk of development of tuberculosis, % | | | Estimated protection from distant primary infection, % |
|---|---|---|---|---|---|---|
| | | | following recent primary infection | following distant primary infection | | |
| | | | | no recent infection | plus recent infection | |
| All | 1952–1967 | 40–59 | 3.78 | 0.0152 | 1.34 | 65 |
| pulmonary | 1955–1970 | 15–69 | 5.10 | 0.0163 | 1.06 | 79 |
| Bacillary | 1952–1967 | 40–59 | 0.83 | 0.0105 | 0.60 | 28 |
| non-primary | 1955–1970 | 15–69 | 1.22 | 0.0104 | 0.68 | 44 |

Vynnycky and Fine extended Sutherland's original work to estimate the age-specific annual risks of developing 'primary', 'endogenous', and 'exogenous' disease using data on the incidence of TB in England and Wales from 1900 to 1990, and the relative risk of progression by time since infection [81]. This model is the most comprehensive and rigorously validated of the two: we therefore find its estimates the most compelling to inform our model.

Progression to disease following re-infection

By fitting their model to data on the incidence of respiratory forms of TB reported in England and Wales between 1900 and 1990, Vynnycky and Fine estimated that a previous TB infection confers a 16% protection against developing disease following a re-infection event in 15 year olds, and a 41% protection in those aged 20 and older [81].

While estimates from Vynnycky and Fine’s model are the most rigorous available on the risk of disease due to re-infection, their modeling approach presents some limitations. The equation-based model structure they used implicitly assumed that individuals in the population mix randomly. Further, they assumed that the risk of infection did not vary by age. Because social contact patterns may generate sub-populations where the risk of infection is high, the importance of re-infection might be greater than their estimates suggest [184]. The likely effect of these assumptions would be to underestimate the true level of protection against subsequent disease that is conferred by a primary infection.

**Table A.1.2.** Model estimates of the risk of developing disease following infection or re-infection, by age and year since infection. From: Vynnycky 1997[81].

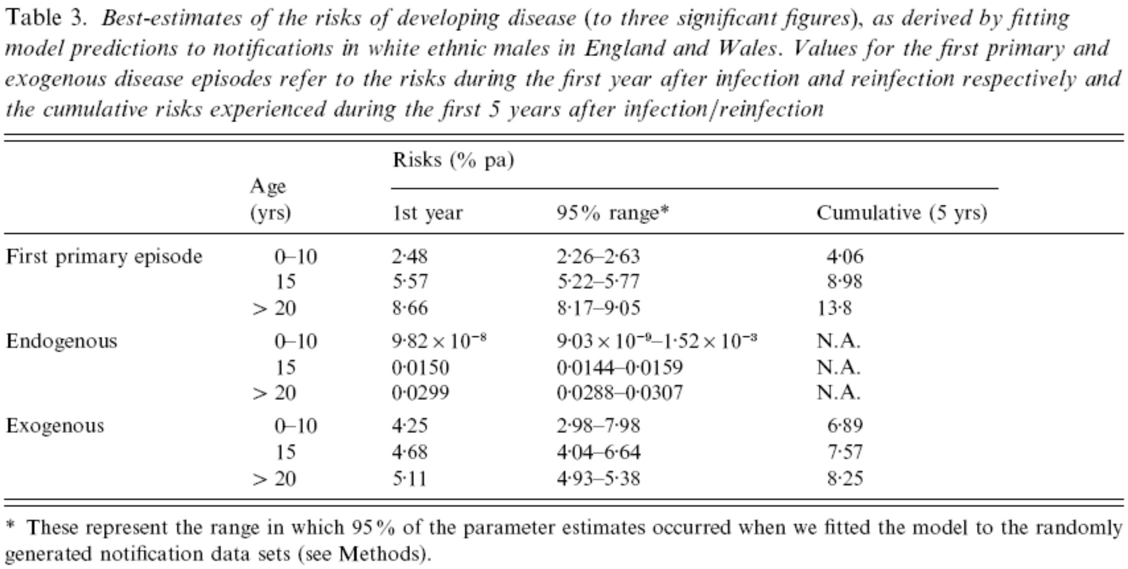
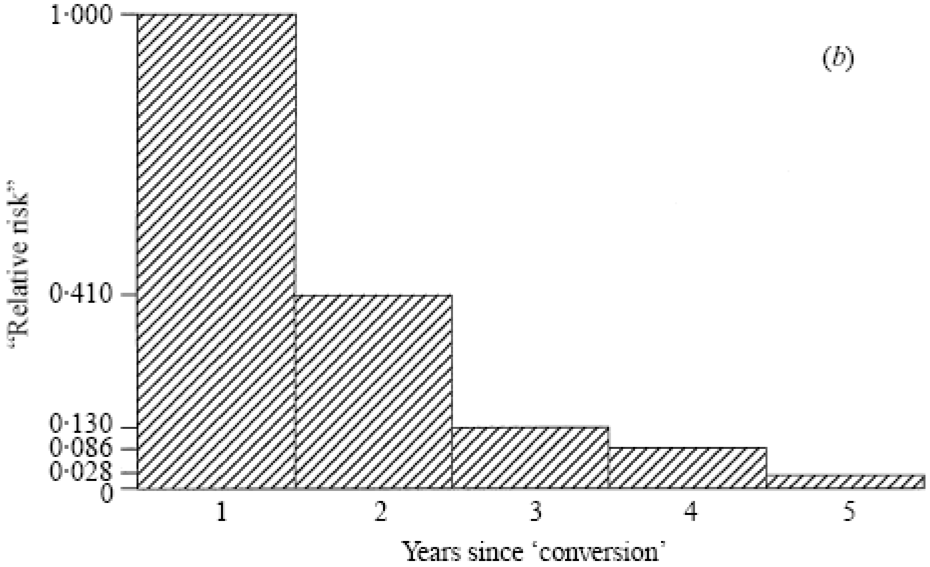


Table 3. *Best-estimates of the risks of developing disease (to three significant figures), as derived by fitting model predictions to notifications in white ethnic males in England and Wales. Values for the first primary and exogenous disease episodes refer to the risks during the first year after infection and reinfection respectively and the cumulative risks experienced during the first 5 years after infection/reinfection*

| | Age (yrs) | Risks (% pa) 1st year | 95% range* | Cumulative (5 yrs) |
|---|---|---|---|---|
| First primary episode | 0–10 | 2·48 | 2·26–2·63 | 4·06 |
| | 15 | 5·57 | 5·22–5·77 | 8·98 |
| | > 20 | 8·66 | 8·17–9·05 | 13·8 |
| Endogenous | 0–10 | $9{\cdot}82 \times 10^{-8}$ | $9{\cdot}03 \times 10^{-9}$–$1{\cdot}52 \times 10^{-3}$ | N.A. |
| | 15 | 0·0150 | 0·0144–0·0159 | N.A. |
| | > 20 | 0·0299 | 0·0288–0·0307 | N.A. |
| Exogenous | 0–10 | 4·25 | 2·98–7·98 | 6·89 |
| | 15 | 4·68 | 4·04–6·64 | 7·57 |
| | > 20 | 5·11 | 4·93–5·38 | 8·25 |

* These represent the range in which 95% of the parameter estimates occurred when we fitted the model to the randomly generated notification data sets (see Methods).

**Figure A.1.1.** Relationship between years since conversion and risk of progressing to active disease. From: Vynnycky 1997[81].



(b) Observed and assumed relationship between the rate at which individuals experience their first primary episode/exogenous disease in each year following infection/reinfection relative to that during the first year after infection/reinfection. These were estimated from the distribution of the time interval between `tuberculin conversion' and disease onset of those who were tuberculin-negative at the start of the UK MRC BCG trial.The 'relative risk' for a given year after 'conversion' is taken to be the ratio between: (i) the proportion of the total disease incidence among initially tuberculin-negative individuals which occurred in that year following 'conversion', and (ii) the corresponding proportion which occurred during the first year after 'conversion'

Site of disease

In the pre-AIDS era in the United States, approximately 85% of TB cases were limited to pulmonary involvement [191]. More recent studies describe a proportion of exclusively pulmonary disease ranging from 62% in the Netherlands between 1993 and 2001 [192] to 78% in Hong Kong in 1996 [193] to 88% in Arkansas between 1996 and 2000 [194]: in the United States in 2001, 80% of disease was exclusively pulmonary [195]. Differences in definitions may explain some of this variation: pleural disease, for example, was considered as a manifestation of pulmonary disease by Yang and colleagues, while a manifestation of extrapulmonary disease by both te Beek and colleagues and Noertjojo and colleagues. Additionally, the wide variety of clinical manifestations of extrapulmonary TB causes difficulty and delay in diagnosis [196, 197], presenting the potential for differential case ascertainment across populations. Beyond methodologic differences, however, population differences are likely largely responsible for the range in prevalence of exclusively pulmonary vs extrapulmonary disease. The risk of extrapulmonary disease has been observed to vary by nationality [192], age, gender, race/ethnicity, and HIV status [192, 193, 196]. In HIV infected individuals, extrapulmonary disease accounts for more than 50% of cases [13]. Among HIV negative individuals, extrapulmonary disease is particularly common among women and young children [14].

Infectiousness of an active case

Active tuberculosis disease can manifest in a number of anatomical sites, but, with rare exception, only patients with active pulmonary, laryngeal [7], or pleural [198] disease may transmit the infection.

Sputum Smear

Among individuals with respiratory forms of active tuberculosis disease, the detection of acid-fast bacilli (AFB) in the sputum provides a relative indicator of the infectiousness of the case [12]. Detection of AFB on sputum smear indicates the presence of a high quantity of bacilli in the sputum: the threshold for detecting AFB using light-microscopy is 5000-10,000 bacilli/mL [199].

Sputum smear has long been used as an indicator of the infectiousness of a patient with respiratory disease, with cases with a higher number of bacilli in the smear considered to present a higher risk of transmission [12]. The transmission risk presented by smear-negative cases has been considered to be low and, in the extreme, recommendations for some TB control programs have suggested that smear negative disease does not present a transmission risk [200].

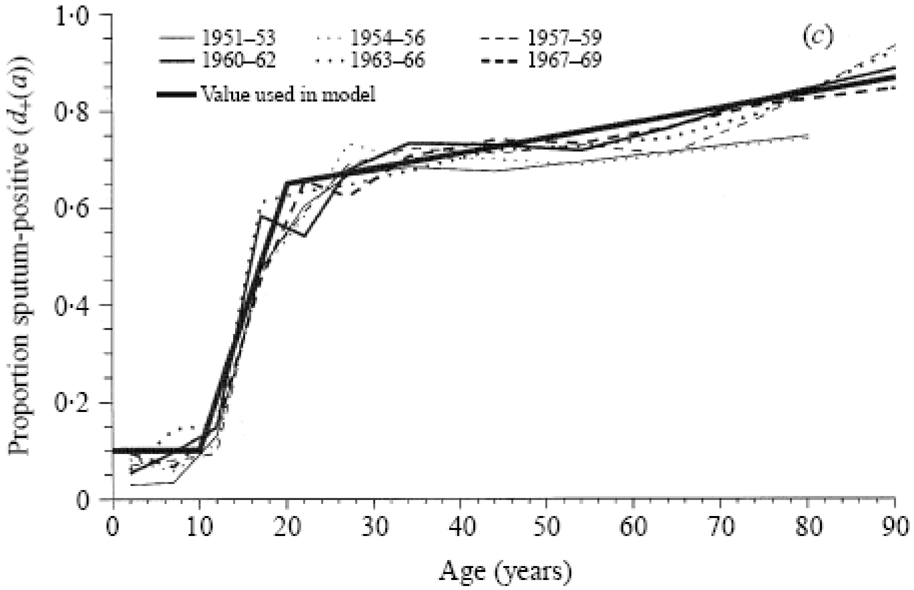
The minimum infecting dose for tuberculosis is very low, estimated to be fewer than 10 organisms [199]. The potential for transmission from smear-negative cases suggested by this low infecting dose (relative to the high quantity of bacilli necessary for AFB to be detected on sputum smear) has been corroborated by evidence from molecular epidemiologic investigations demonstrating transmission from smear negative cases. These epidemiologic studies of all culture confirmed cases of active tuberculosis in Vancouver [15] and San Francisco [164], estimated cases of smear-negative pulmonary tuberculosis to be at least 21% and 22% as likely as smear-positive cases to transmit infection, respectively: in the San Francisco study, the 95% confidence interval for this estimate was 0.16-0.32 [164].

Data from culture-confirmed cases of TB diagnosed between 1991 and 1996 in San Francisco showed that, in that population, 48% of pulmonary disease yielded a positive sputum smear. Pulmonary disease accounted for 93% of all positive sputum smears, suggesting that 7% of positive sputum smears in this study sample resulted from cases with extrapulmonary disease [164]. While the definitions of pulmonary and extrapulmonary TB used by the authors were not stated, it is reasonable to assume that extrapulmonary cases with thoracic involvement were included as ‘extrapulmonary’ cases, accounting for the contribution of this group to smear-positive cases.

These data are generalizable to few populations beyond San Francisco, however, particularly since 23% of the culture-confirmed TB cases in this study were HIV positive.  HIV status has been reported to be inversely related to the risk of having a positive sputum smear: after adjusting for other factors influencing smear status, HIV positive individuals diagnosed with active TB in Catalonia, Spain, between 1996 and 1997 were only half as likely to have a positive sputum smear as were HIV negative individuals diagnosed with active TB [201]. Beyond HIV status, age, gender, and alcohol abuse have all been observed to impact sputum smear status [201].

In estimating parameters for their equation-based model of the natural history of tuberculosis infection, Vynnycky and Fine [81] plotted data on the age-specific prevalence of sputum-smear positive disease from males in Norway between 1951 and 1969.  Similarly to the study from Spain, these data illustrate a substantially lower risk of developing smear positive disease in individuals in the youngest age groups.  However, while Godoy et al. found a higher risk of smear-positive disease among middle-age groups than among the oldest age group in their study (>44 years), the Norwegian data illustrate a steadily increasing risk of smear-positive disease from age 20 on.  The Spanish study included all forms of active disease, while Vynnycky and Fine plotted only respiratory forms of TB.  The conflicting results in the older age-groups may be due to this methodologic difference.

**Figure A.1.2** Relationship between Age and Proportion of Respiratory TB Cases that are Smear-Positive, Norway 1951-1969. From: Vynnycky 1997[81].



Observed and assumed proportion of total respiratory disease incidence among cases of age a attributable to sputum-positive forms, d+(a). All lines (excluding the heavy solid line) show the relative contribution of sputum-positive disease to age-specific notifications of pulmonary tuberculosis in males in Norway (1951±69). Source: Dr K. Styblo (TSRU) and Dr K. Bjartveit (Norwegian National Health Screening Service).

Duration of infectivity

The length of time for which an individual with active tuberculosis disease remains infectious is critically important to the risk of transmission from that individual. The duration of the infectious period will be influenced by three major components; i) the length of time between the development of active, infectious disease, and the initiation of treatment, ii) the length of time between the initiation of treatment and the conversion of the patient to a non-infectious state, and iii) the risk of mortality due to tuberculosis disease.

This framework assumes, of course, that all infectious individuals eventually initiate treatment, and that, once treatment is started, all patients successfully complete treatment. Neither of these assumptions is likely to be the case in most settings. Therefore, the proportion of active TB cases that go untreated, and the proportion of patients that successfully complete treatment, will both influence the average duration of infectivity in a population.

Delay to diagnosis

The delay between the onset of symptoms of active tuberculosis disease (assumed here to correspond to the start of the infectious period) and diagnosis/treatment is composed of two factors: i) the patient's delay between the onset of symptoms and seeking medical care ("patient delay"), and ii) the delay to accurate diagnosis once care is sought ("health care delay").

**Table A.1.3.** Delay to diagnosis of tuberculosis cases: selective summary of the literature. "Total delay" refers to the delay between symptom onset and diagnosis of TB. "Patient delay" refers to the delay between the onset of symptoms and first health care visit. "Health care delay" refers to the delay between the time the patient first sought care and the time the diagnosis of TB was made.

| **Author (Year)** | **Setting** | **Duration of delay (median or mean, as indicated)** |
|---|---|---|
| Sherman (1999)[166] | All culture-positive tuberculosis patients without previous treatment for tuberculosis (n = 184), New York City, April 1994. | Median delay of 57 days<br>(Range: 4 – 764) |
| Golub (2005)[202] | Tuberculosis (TB) patients reported to the Maryland Department of Health and Mental Hygiene from 1 June 2000 to 30 November 2001. (n = 158) | Median total delay of 89 days, composed of a median patient delay of 32 days (range 0–539 days), and a median health care delay of 26 days (range 0–519 days). |
| Sarmiento(2006)[203] | Harlem Hospital Directly Observed Therapy (DOT) Program, New York City.<br>Cross-sectional survey of the help-seeking behavior of TB patients within 2 months of their enrollment into DOT from May 2001 to December 2004.<br>(n=39) | Total delay between symptom onset and diagnosis of TB =18.0 weeks |

Time from initiation of treatment and resolution of infectiousness

Sputum smear status is generally regarded as the main indicator of infectiousness, and is tracked over the course of treatment as an indicator of contagiousness. Patients smear negative at the initiation of treatment are considered to be noninfectious after two weeks of treatment [204], while initially smear positive patients are considered to be noncontagious once a negative sputum smear is obtained on treatment [205]. In a hospital based study conducted between 1997 and 2001 in Alberta, Canada, the average time to sputum smear conversion among HIV seronegative patients with initially smear-positive tuberculosis was 46 days. Even under the optimal treatment conditions under which these cases were managed, however, the time to smear conversion varied widely, ranging from 8 to 115 days.

Sputum smear conversion is only a relative indicator of noninfectiousness, however --culture conversion is considered as an indicator of absolute noninfectiousness. In the hospital-based study in Alberta, only 34.4% of patients had converted by culture by the time they had converted by sputum smear [205]. Hospital-based studies conducted in both Ankara, Turkey, and in Madrid, Spain, both reported similar times to sputum and culture conversion: 59.4 ± 32.2 days to smear conversion and 57.1 ± 29.9 days to culture conversion in the Turkey study [206], and 34± 26 and 38 ±32 days in the Spain study [167]. In these studies, the time to smear and culture conversion was independently associated with the presence of diabetes [206], high bacillary counts in sputum smears at diagnosis, cavitary disease, and infection with a drug-resistant strain [167].

Recurrence of disease following successful treatment

Following treatment with standard short course chemotherapy, active tuberculosis recurs in 2 to 7% of cases [17]. Some proportion of these cases are likely to be due to re-infection events: this proportion likely varies widely across populations, however, as it will be dependent upon the risk of infection (itself a function of the prevalence of active, infectious tuberculosis cases in a population), as well as the population prevalence of factors that increase the risk of progression from infection to active disease.

In populations where the risk of infection is low, the majority of recurrent TB is likely due to recurrence of disease from a primary infection, rather than reinfection.  A molecular epidemiologic study of tuberculosis in the Netherlands between 1994 and 1997 suggested that, among patients having suffered active TB disease prior to 1981, 25% of new disease episodes be attributable to recent re-infection [207].  In cases from two prospective clinical trials of TB treatment regimens, a molecular epidemiologic analysis found a relapse rate in HIV negative individuals of 3.59 per 100 patient years, compared to a rate of disease due to reinfection of 0.12 per 100 patient years In HIV positive individuals, the relapse rate was 4.09 per 100 patient years, with a rate of disease due to reinfection of 0.27 per 100 patient years [17].

## Appendix 2

### Diversity and Discrimination: Molecular typing markers for *M. tuberculosis*

Fundamentally, all typing methods rely on diversity: without diversity in the characteristics or markers they assess, all organisms typed by a given method will appear the same. The 4.4 MB genome of *M. tuberculosis* is highly conserved, and the population is a largely clonal one, with little sequence polymorphism in structural genes[126]. Horizontal gene exchange, which occurs extensively in many bacterial pathogens, such as *Escherichia coli* and *Neisseria gonorrhea,* is rare in *M. tuberculosis* [208, 209]. Genetic regions known to be highly variable in other mycobacteria have been found to be invariant in *M. tuberculosis* and an early report found only four synonymous nucleotide substitutions among more 200,000 base pairs (bp) of DNA that was sequenced from diverse *M. tuberculosis* strains [210]. This lack of genetic variation surprised initial investigators, given the substantial phenotypic variability in the species. Research since these early reports has suggested that the *M. tuberculosis* genome is less homogeneous than originally suspected, with comparative genomic studies identifying variability in the form of genetic insertions, deletions, and repetitive elements [156, 172, 211].

A number of genotype-base molecular typing techniques have been developed for TB, each exploiting regions of variability in the *M. tuberculosis* genome. These techniques can be broadly categorized into two groups: those based on restriction-fragment length polymorphism (RFLP) analysis, which require culturing the very slow-growing *M. tuberculosis*, and those based on PCR amplification, which can be preformed on a smaller number of bacteria and therefore do not rely on culture. These techniques vary greatly in turnaround time, reproducibility, ease of communicating and comparing results, and the diversity of patterns generated [34].

Insertion elements

A number of insertion sequence (IS) elements, have been identified in the *M. tuberculosis* genome. IS elements are mobile genetic elements that encode genes that enable their transposition. While IS elements are characterized by their potential to translocate within the genome, the stability of the IS elements so far identified in *M. tuberculosis* appears to vary greatly, with the majority exhibiting little variation across strains [34]. One IS element, however, identified as IS*6110*, is characterized by a high mobility, and substantial diversity in the number and distribution of elements present in the *M. tuberculosis* genome [212].

A large amount of variation in the *M. tuberculosis* genome has been associated with the insertion (IS) element IS*6110*. This IS element is highly variable in number and distribution throughout the *M. tuberculosis* genome, with the number of copies in the genome varying between 0 and 25 [34]**.** IS*6110* appears to contribute to large sequence polymorphisms (LSPs), as recombination between IS*6110* elements can result in the deletion of the genetic region between the two elements [136]**.** IS*6110* has a higher level of transpositional activity than other IS elements identified in *M. tuberculosis*, and often inserts into coding regions [213]**.** While IS*6110* can integrate anywhere in the *M. tuberculosis* genome, "hot-spots" for insertion have been identified, indicating that the distribution of this element throughout the genome is not random [214].

### IS6110 RFLP typing

IS*6110* RFLP typing, which assesses diversity in the number and distribution of the IS*6110* IS element throughout the *M. tuberculosis* genome, is considered to be the "gold standard" in TB typing. In this RFLP-based method, typing patterns are generated by restriction digest of *M. tuberculosis* genomic DNA followed by Southern hybridization using an IS*6110* probe [215]. Despite a slow turnaround time, relatively high expense, and the difficulty in communicating and comparing the gel-based patterns that this technique generates, these patterns are highly reproducible, and the diversity of patterns generated is considered to distinguish well between epidemiologically related and unrelated isolates.

Discrimination based on this technique, however, cannot be considered ideal for the purposes of epidemiologic typing. The diversity of patterns generated by this method depends on the number of IS*6110* elements present, owing to insertion "hot-spots" which constrain diversity when fewer than six IS*6110* elements are present [153]. These low-copy number isolates are therefore less diverse by IS*6110* RFLP, and agree less well with epidemiologic data, than isolates with six or more IS*6110* elements [114]. Additionally, a study of serial isolates collected from patients over time found that changes in IS*6110* RFLP pattern were less likely in low-copy than high-copy isolates [216]. The rate of pattern change (molecular clock) that is best suited to epidemiologic typing of TB is unclear, and while IS*6110* is often described as stable enough to distinguish epidemiologically related from unrelated isolates [217], a number of reports have described changes in IS*6110* RFLP patterns occurring between isolates that were known to have been directly related by transmission [147] [35]. Different rates of pattern change between low copy and high copy isolates complicate the interpretation of epidemiologic typing data. Additionally, the molecular clock may vary according to the genetic background of the strain [218, 219], further complicating the interpretation of typing results based on this method.

<u>Repetitive polymorphic sequences</u>

DR locus

In an investigation to characterize a putative IS element (IS*987*), Hermans et al.[220] identified an unique locus in the *M. tuberculosis* genome, which is characterized by the presence of multiple direct-repeat (DR) sequences of 36 bp, with non-repetitive "spacer DNA" regions, of 35 to 41 bp in length, between each repeat. One of the DR sequences was split by the IS element that was the initial focus of their investigation, IS*987*. This locus, termed the "DR locus", was shown to be highly conserved within species of the *M. tuberculosis* complex, and also unique to this complex, as it was not identified in any other mycobacteria. Further investigation of the DR locus revealed that *M. tuberculosis* strains vary in the number of DRs and in the presence or absence of particular spacers [221].

Spoligotyping

Spoligotyping, a typing technique based on the PCR amplification of the unique spacer sequences between direct repeats of the DR locus, is a rapid, highly reproducible typing technique [217]. However, the level of discrimination provided by spoligotyping is low relative to other common techniques. While it is somewhat more discriminatory than IS*6110* RFLP for low-band isolates, this technique is rarely used for epidemiologic typing, other than in conjunction with IS*6110* RFLP as a secondary typing method for low-band isolates.

Although not well suited to epidemiologic typing, spoligotyping is widely used to characterize phylogenetic and geographic distribution of *M. tuberculosis* strains [28, 29], and to assess the diversity of the *M. tuberculosis* population in a given region [222]. Spoligotyping is limited in its use as a marker for phylogenetic studies, as the evolution of individual loci in the DR region is not independent (Contiguous blocks of spacers can be lost in single deletion events), transposition of insertion sequences can lead to convergence of spoligotype patterns, and evolution is unidirectional (spacers can be lost, but not gained) [136]. However, owing in part to the ease and economy of the technique, a global database of spoligotyping results is actively maintained [132], facilitating international comparisons and communication about spoligotype-defined strain families.

### PGRS

A whole-genome comparison of two sequenced strains of *M. tuberculosis,* CDC1551 and H37Rv, found high levels of polymorphism in genes of the PE/PPE gene family [32]. Genes of this family, some of which have been implicated in virulence or host immune response, comprise approximately 5% of the *M. tuberculosis* genome [33], and represent a major source of genetic variability across the species [223]. The polymorphic GC-rich tandem repeat sequence (PGRS), which is present in multiple genomic clusters throughout the genome, is associated with genes of the PE/PPE family, and possibly contributes to the antigenic variation that has been associated with this family of genes [34].

### pTBN12 typing

While not evaluated as a stand-alone typing technique, pTBN12 typing is one of the most common secondary typing techniques used alongside IS6110 RFLP typing to improve discrimination among low-band isolates. This RFLP-based technique employs a recombinant plasmid, pTBN12, which carries an insert of PGRS as a probe. RFLP patterns generated reflect the number and distribution of the PGRS sequence in the genome [34]. While time and labor intensive, this technique provides results that agree well with epidemiologic data for low band isolates [140, 224].

### Exact Tandem Repeats, Variable-Number Tandem Repeats, and Mycobacterial Interspersed Repeat Units

The publication of the complete sequence from *M. tuberculosis* strain H37Rv allowed for a more systematic investigation of genetic variability. Using this published sequence, Frothingham and colleagues [225] identified regions of tandemly repeated DNA sequence. Such regions had previously been identified in a diverse array of other organisms, ranging from humans to bacteria, and were noted to be highly variable. In humans, variability in this type of locus was already being exploited as a marker to facilitate genetic mapping as well as forensic and paternity testing. In their first report, Frothingham et al. [225] described 6 exact tandem repeat regions (ETR), each with a unique repeat sequence of between 53 and 79 bp, and which was highly polymorphic in their testing panel of 48 strains from diverse geographic locations. While initially only a limited number of these tandem repeat sequences were identified, in recent years more than 31 additional tandem repeat sequences of between 40 and 100 base pairs (bp) have been identified in the *M. tb* genome. These tandem repeat sequences, variously called *V*ariable *N*umber *T*andem *R*epeats (VNTR), *M*ycobacterial *I*nterspersed *R*epetitive *U*nits (MIRU), and *E*xact *T*andem *R*epeats, are thought to be the most variable structures in the *M. tuberculosis* genome families [156].

## MIRU-VNTR

Variability in number of repeats at tandem-repeat loci is exploited in an increasingly favored genetic typing approach, most commonly referred to as MIRU or MIRU-VNTR. This approach is highly analogous to microsatellite typing in higher eukaryotes, and the high-throughput methods that were originally developed for typing of these organisms has been adapted to use with *M. tuberculosis*. This PCR-based technique characterizes the number of repeats at each of a series of independent loci, resulting in a highly reproducible digital pattern that can be easily catalogued and communicated [215]. Initial typing sets for MIRU-VNTR and MIRU-VNTR like systems such as ETR were based on very limited sets of loci, and resulted in low levels of discrimination. More recent incarnations of this method use a combination of these original loci in addition to more recently identified loci to create typing panels of 12, 15, 25, or 29 loci [123, 226], which promise substantially higher levels of discrimination. As this rapid, economical, and highly flexible technique may achieve levels of discrimination comparable to IS*6110* RFLP (assuming the identification of an optimal panel of typing loci), MIRU-VNTR has been heralded as the successor to IS*6110* RFLP [74, 215]. MIRU-VNTR typing based on a 12-locus panel has already replaced IS*6110* as the primary typing method for routine TB surveillance in the United States [37], and a proposal has been made for the institution of a 15-locus typing panel as an international standard [123].

Considerable debate remains regarding the composition of the optimal MIRU-VNTR typing panel, particularly if a single panel is to be accepted as an international standard. As the diversity and stability of MIRU loci appears to vary according to genetic family [125], it is questionable if a single typing panel could provide optimal discrimination to differentiate epidemiologically related from unrelated isolates across diverse global populations. This issue is of particular concern in populations with a high prevalence of isolates belonging to a specific family known as the Beijing family, which is dominant across many countries in Asia and former Soviet Union, and a variant of which has been associated with multiple outbreaks of multi-drug resistant (MDR) TB in the United States ([169, 227, 228]. In the Beijing family, many of the MIRU loci included in standard typing panels exhibit very low levels of polymorphism [45, 46]. MIRU loci which exhibit higher levels of polymorphism in the Beijing family have been identified [137, 170], but these loci are not included in the MIRU-VNTR typing panel currently used for routine typing in the United States, nor in the optimized panel proposed as an international standard [38].

### Stability of MIRU loci

Variation in MIRU loci appears to exhibit stepwise variation in the number of repeats, with change occurring by the gain or loss of single repeat units [156]. The mechanism by which this change occurs has not been proven, but it has been suggested that, in eukaryotic organisms, slipped-strand mis-pairing (SSM) of the DNA polymerase might be the cause [229]. The absence of a mismatch repair system in *M. tuberculosis* [230] may foster variation by this mechanism. While variation in many MIRU loci is consistent with this model [11, 41], the rarity with which strand-slippage mutation events occur for repeats as large as those occurring at the MIRU loci, it has been suggested that homologous recombination may be a major mechanism for the generation of variation at MIRU loci [156]. The level of polymorphism at individual MIRU loci ranges substantially, and evidence suggests that the rate of change of individual loci varies, and that this variation is dependent on the genetic background of the isolate [125]. A direct calculation of the rate of change at these loci is not possible, and the best evidence *in vivo* comes from investigations of serial isolates in persistently infected patients. In a study of patients with persistent disease who remained infectious for up to 2,185 days, serially collected isolates from 55 of 56 patients exhibited identical MIRU profiles, using a standard 12-loci panel. The single patient with a variant MIRU profile had two serial isolates with 11 identical loci, but which differed at MIRU allele 26 by a single repeat difference. In this same study sample, 11 of 56 patients had serial isolates with minor variations in their IS*6110* banding patterns [160], suggesting that patterns based on this 12 loci MIRU typing panel are less stable than patterns based on IS*6110* RFLP.

Stability of typing patterns in latent infection

There is considerable uncertainty regarding the state of *M. tuberculosis* during latent infection. Latent infection has recently been characterized as a dynamic microenvironment, with continuous activation of the immune response to restrain replication of the bacteria [85]. Whether or not *M. tuberculosis* replicates during latent infection, and the extent of replication that may occur, is unknown. It has been suggested that both dormant and replicating *M. tuberculosis* might be present in latently infected individuals within different types of lesions [84].

Whether or not molecular typing patterns evolve during latency, a question with clear implications for the interpretation of molecular typing data, is similarly uncertain. Evidence from an investigation of epidemiologically linked TB cases in the Netherlands with years to decades separating the date of suspected infection from the date of disease onset suggest that IS*6110* RFLP patterns are stable over the course of latent infection[40]. While no similar investigations have been made using other molecular typing techniques, in vitro evidence suggests that MIRU loci evolve slowly in anoxic culture conditions that may be similar to the conditions *M. tuberculosis* experiences within a latent granuloma [156]

# Bibliography

1. Espinal, M.A., Raviglione, M.C., *Global Epidemiology of Tuberculosis*, in *Tuberculosis*, M.M. Madkour, Editor. 2004, Springer: Berlin ; New York. p. 34-43.
2. *Global Tuberculosis Control: Surveillance, Planning, Financing*. 2007, World Health Organization: Geneva. p. 280.
3. Cegielski, J.P., et al., *The global tuberculosis situation. Progress and problems in the 20th century, prospects for the 21st century.* Infect Dis Clin North Am, 2002. **16**(1): p. 1-58.
4. *Trends in tuberculosis incidence--United States, 2006.* MMWR Morb Mortal Wkly Rep, 2007. **56**(11): p. 245-50.
5. Bennett, D.E., et al., *Prevalence of Tuberculosis Infection in the U.S. Population.* Am J Respir Crit Care Med, 2007.
6. Haddad, M.B., et al., *Tuberculosis and homelessness in the United States, 1994-2003.* JAMA, 2005. **293**(22): p. 2762-6.
7. Frieden, T.R., et al., *Tuberculosis.* Lancet, 2003. **362**(9387): p. 887-99.
8. Cantwell, M.F., et al., *Epidemiology of tuberculosis in the United States, 1985 through 1992.* JAMA, 1994. **272**(7): p. 535-9.
9. *HIV/AIDS Surveillance Report*. 2000, Centers for Disease Control and Prevention: Atlanta.
10. Skuce, R., *Molecular Epidemiology of Mycobacterium bovis*, in *Tuberculosis*, M.M. Madkour, Editor. 2004, Springer: Berlin ; New York. p. 75-92.
11. Osoba, O.A., *Microbiology of Tuberculosis*, in *Tuberculosis*, M.M. Madkour, Editor. 2004, Springer: Berlin ; New York. p. 115-132.
12. *American Thoracic Society/Centers for Disease Control and Prevention/Infectious Diseases Society of America: controlling tuberculosis in the United States.* Am J Respir Crit Care Med, 2005. **172**(9): p. 1169-227.
13. Sharma, S.K. and A. Mohan, *Multidrug-resistant tuberculosis.* Indian J Med Res, 2004. **120**(4): p. 354-76.
14. Ong, A., et al., *A molecular epidemiological assessment of extrapulmonary tuberculosis in San Francisco.* Clin Infect Dis, 2004. **38**(1): p. 25-31.
15. Hernandez-Garduno, E., et al., *Transmission of tuberculosis from smear negative patients: a molecular epidemiology study.* Thorax, 2004. **59**(4): p. 286-90.
16. Maartens, G. and R.J. Wilkinson, *Tuberculosis.* Lancet, 2007. **370**(9604): p. 2030-43.
17. Jasmer, R.M., et al., *Recurrent tuberculosis in the United States and Canada: relapse or reinfection?* Am J Respir Crit Care Med, 2004. **170**(12): p. 1360-6.
18. Stead, W.W., *Pathogenesis of a first episode of chronic pulmonary tuberculosis in man: recrudescence of residuals of the primary infection or exogenous reinfection?* Am Rev Respir Dis, 1967. **95**(5): p. 729-45.
19. van Rie, A., et al., *Exogenous reinfection as a cause of recurrent tuberculosis after curative treatment.* N Engl J Med, 1999. **341**(16): p. 1174-9.
20. Nardell, E., et al., *Exogenous reinfection with tuberculosis in a shelter for the homeless.* N Engl J Med, 1986. **315**(25): p. 1570-5.
21. Caminero, J.A., et al., *Exogenous reinfection with tuberculosis on a European island with a moderate incidence of disease.* Am J Respir Crit Care Med, 2001. **163**(3 Pt 1): p. 717-20.
22. Bandera, A., et al., *Molecular epidemiology study of exogenous reinfection in an area with a low incidence of tuberculosis.* J Clin Microbiol, 2001. **39**(6): p. 2213-8.
23. Garcia de Viedma, D., et al., *Tuberculosis recurrences: reinfection plays a role in a population whose clinical/epidemiological characteristics do not favor reinfection.* Arch Intern Med, 2002. **162**(16): p. 1873-9.
24. Foxman, B. and L. Riley, *Molecular epidemiology: focus on infection.* Am J Epidemiol, 2001. **153**(12): p. 1135-41.

25. Lancefield, R.C., *A serological differentiation of human and other groups of hemolytic streptococci.* Journal of Experimental Medicine, 1933. **57**(4): p. 571–595.

26. Jones WD Jr, K.G., *Fluorescent antibody techniques with mycobacteria. 3. Investigation of five serologically homogenous groups of mycobacteria* Zentralbl Bakteriol, 1968. **207**(1): p. 58–62.

27. Bates, J.H., Fitzhugh, J.K., *Subdivision of the species M. tuberculosis.* Am Rev Respir Dis, 1967. **96**: p. 7-10.

28. Jones, W.D., Jr. and J. Greenberg, *Use of phage F-phi WJ-1 of Mycobacterium fortuitum to discern more phage types of Mycobacterium tuberculosis.* J Clin Microbiol, 1976. **3**(3): p. 324-6.

29. Snider, D.E., Jr., W.D. Jones, and R.C. Good, *The usefulness of phage typing Mycobacterium tuberculosis isolates.* Am Rev Respir Dis, 1984. **130**(6): p. 1095-9.

30. Jones, W.D., Jr., *Bacteriophage typing of Mycobacterium tuberculosis cultures from incidents of suspected laboratory cross-contamination.* Tubercle, 1988. **69**(1): p. 43-6.

31. Kanduma, E., T.D. McHugh, and S.H. Gillespie, *Molecular methods for Mycobacterium tuberculosis strain typing: a users guide.* J Appl Microbiol, 2003. **94**(5): p. 781-91.

32. Cave, M.D., et al., *IS6110: conservation of sequence in the Mycobacterium tuberculosis complex and its utilization in DNA fingerprinting.* Mol Cell Probes, 1991. **5**(1): p. 73-80.

33. Kamerbeek, J., et al., *Simultaneous detection and strain differentiation of Mycobacterium tuberculosis for diagnosis and epidemiology.* J Clin Microbiol, 1997. **35**(4): p. 907-14.

34. Yang, Z., *Molecular epidemiology of tuberculosis.* Front Biosci, 2003. **8**: p. d440-50.

35. Cave, M.D., et al., *Epidemiologic import of tuberculosis cases whose isolates have similar but not identical IS6110 restriction fragment length polymorphism patterns.* J Clin Microbiol, 2005. **43**(3): p. 1228-33.

36. Mazars, E., et al., *High-resolution minisatellite-based typing as a portable approach to global analysis of Mycobacterium tuberculosis molecular epidemiology.* Proc Natl Acad Sci U S A, 2001. **98**(4): p. 1901-6.

37. Cowan, L.S., et al., *Evaluation of a two-step approach for large-scale, prospective genotyping of Mycobacterium tuberculosis isolates in the United States.* J Clin Microbiol, 2005. **43**(2): p. 688-95.

38. Yeh, R.W., P.C. Hopewell, and C.L. Daley, *Simultaneous infection with two strains of Mycobacterium tuberculosis identified by restriction fragment length polymorphism analysis.* Int J Tuberc Lung Dis, 1999. **3**(6): p. 537-9.

39. Shamputa, I.C., et al., *Mixed infection and clonal representativeness of a single sputum sample in tuberculosis patients from a penitentiary hospital in Georgia.* Respir Res, 2006. **7**: p. 99.

40. Lillebaek, T., et al., *Stability of DNA patterns and evidence of Mycobacterium tuberculosis reactivation occurring decades after the initial infection.* J Infect Dis, 2003. **188**(7): p. 1032-9.

41. Golub, J.E., et al., *Transmission of Mycobacterium tuberculosis through casual contact with an infectious case.* Arch Intern Med, 2001. **161**(18): p. 2254-8.

42. Small, P.M., et al., *The epidemiology of tuberculosis in San Francisco. A population-based study using conventional and molecular methods.* N Engl J Med, 1994. **330**(24): p. 1703-9.

43. Ellis, B.A., et al., *Molecular epidemiology of tuberculosis in a sentinel surveillance population.* Emerg Infect Dis, 2002. **8**(11): p. 1197-209.

44. van Soolingen, D., et al., *Molecular epidemiology of tuberculosis in the Netherlands: a nationwide study from 1993 through 1997.* J Infect Dis, 1999. **180**(3): p. 726-36.

45. Dye, C., et al., *Consensus statement. Global burden of tuberculosis: estimated incidence, prevalence, and mortality by country. WHO Global Surveillance and Monitoring Project.* JAMA, 1999. **282**(7): p. 677-86.

46. Institute of Medicine (U.S.). Committee on the Elimination of Tuberculosis in the United States. and L. Geiter, *Ending neglect : the elimination of tuberculosis in the United States*. 2000, Washington, D.C.: National Academy Press. xvi, 269 p.

47. Weis, S., *Contact investigations: how do they need to be designed for the 21st century?* Am J Respir Crit Care Med, 2002. **166**(8): p. 1016-7.

48. Weis, S.E., et al., *Transmission dynamics of tuberculosis in Tarrant county, Texas.* Am J Respir Crit Care Med, 2002. **166**(1): p. 36-42.

49. Cronin, W.A., et al., *Molecular epidemiology of tuberculosis in a low- to moderate-incidence state: are contact investigations enough?* Emerg Infect Dis, 2002. **8**(11): p. 1271-9.

50. Frieden, T.R., et al., *A multi-institutional outbreak of highly drug-resistant tuberculosis: epidemiology and clinical outcomes.* JAMA, 1996. **276**(15): p. 1229-35.

51. Frieden, T.R., T.R. Sterling, and P.M. Simone, *Tuberculosis in a neighborhood bar.* N Engl J Med, 1996. **334**(5): p. 334.

52. Tiruviluamala, P. and L.B. Reichman, *Tuberculosis.* Annu Rev Public Health, 2002. **23**: p. 403-26.

53. Ijaz, K., et al., *Persistence of a strain of Mycobacterium tuberculosis in a prison system.* Int J Tuberc Lung Dis, 2004. **8**(8): p. 994-1000.

54. Barnes, P.F., et al., *Patterns of tuberculosis transmission in Central Los Angeles.* JAMA, 1997. **278**(14): p. 1159-63.

55. McNabb, S.J., et al., *Added epidemiologic value to tuberculosis prevention and control of the investigation of clustered genotypes of Mycobacterium tuberculosis isolates.* Am J Epidemiol, 2004. **160**(6): p. 589-97.

56. Bishai, W.R., et al., *Molecular and geographic patterns of tuberculosis transmission after 15 years of directly observed therapy.* JAMA, 1998. **280**(19): p. 1679-84.

57. Malakmadze, N., et al., *Unsuspected recent transmission of tuberculosis among high-risk groups: implications of universal tuberculosis genotyping in its detection.* Clin Infect Dis, 2005. **40**(3): p. 366-73.

58. Veen, J., *Microepidemics of tuberculosis: the stone-in-the-pond principle.* Tuber Lung Dis, 1992. **73**(2): p. 73-6.

59. Etkind, S.C., *The role of the public health department in tuberculosis.* Med Clin North Am, 1993. **77**(6): p. 1303-14.

60. van Deutekom, H., et al., *A molecular epidemiological approach to studying the transmission of tuberculosis in Amsterdam.* Clin Infect Dis, 1997. **25**(5): p. 1071-7.

61. Pfyffer, G.E., et al., *Transmission of tuberculosis in the metropolitan area of Zurich: a 3 year survey based on DNA fingerprinting.* Eur Respir J, 1998. **11**(4): p. 804-8.

62. Daley, C.L. and L.M. Kawamura, *The role of molecular epidemiology in contact investigations: a US perspective.* Int J Tuberc Lung Dis, 2003. **7**(12 Suppl 3): p. S458-62.

63. Rosenblum, L.S., T.R. Navin, and J.T. Crawford, *Molecular epidemiology of tuberculosis.* N Engl J Med, 2003. **349**(24): p. 2364.

64. Templeton, G.L., et al., *The risk for transmission of Mycobacterium tuberculosis at the bedside and during autopsy.* Ann Intern Med, 1995. **122**(12): p. 922-5.

65. Nakamura, Y., et al., *A small outbreak of pulmonary tuberculosis in non-close contact patrons of a bar.* Intern Med, 2004. **43**(3): p. 263-7.

66. Mangura, B.T., et al., *Mycobacterium tuberculosis miniepidemic in a church gospel choir.* Chest, 1998. **113**(1): p. 234-7.

67. Jasmer, R.M., et al., *A molecular epidemiologic analysis of tuberculosis trends in San Francisco, 1991-1997.* Ann Intern Med, 1999. **130**(12): p. 971-8.

68. Cattamanchi, A., et al., *A 13-year molecular epidemiological analysis of tuberculosis in San Francisco.* Int J Tuberc Lung Dis, 2006. **10**(3): p. 297-304.

69. Borgdorff, M.W., et al., *Tuberculosis elimination in the Netherlands.* Emerg Infect Dis, 2005. **11**(4): p. 597-602.

70. Braden, C.R., et al., *Interpretation of restriction fragment length polymorphism analysis of Mycobacterium tuberculosis isolates from a state with a large rural population.* J Infect Dis, 1997. **175**(6): p. 1446-52.

71. Lambregts-van Weezenbeek, C.S., et al., *Tuberculosis contact investigation and DNA fingerprint surveillance in The Netherlands: 6 years' experience with nation-wide cluster feedback and cluster monitoring.* Int J Tuberc Lung Dis, 2003. **7**(12 Suppl 3): p. S463-70.

72. Diel, R., et al., *Epidemiology of tuberculosis in Hamburg, Germany: long-term population-based analysis applying classical and molecular epidemiological techniques.* J Clin Microbiol, 2002. **40**(2): p. 532-9.

73. Oelemann, M.C., et al., *Assessment of an optimized mycobacterial interspersed repetitive- unit-variable-number tandem-repeat typing system combined with spoligotyping for population-based molecular epidemiology studies of tuberculosis.* J Clin Microbiol, 2007. **45**(3): p. 691-7.

74. Blackwood, K.S., J.N. Wolfe, and A.M. Kabani, *Application of mycobacterial interspersed repetitive unit typing to Manitoba tuberculosis cases: can restriction fragment length polymorphism be forgotten?* J Clin Microbiol, 2004. **42**(11): p. 5001-6.

75. Scott, A.N., et al., *Sensitivities and specificities of spoligotyping and mycobacterial interspersed repetitive unit-variable-number tandem repeat typing methods for studying molecular epidemiology of tuberculosis.* J Clin Microbiol, 2005. **43**(1): p. 89-94.

76. Sola, C., et al., *Genotyping of the Mycobacterium tuberculosis complex using MIRUs: association with VNTR and spoligotyping for molecular epidemiology and evolutionary genetics.* Infect Genet Evol, 2003. **3**(2): p. 125-33.

77. Altman, D.G. and J.M. Bland, *Diagnostic tests 2: Predictive values.* BMJ, 1994. **309**(6947): p. 102.

78. Alland, D., et al., *Transmission of tuberculosis in New York City. An analysis by DNA fingerprinting and conventional epidemiologic methods.* N Engl J Med, 1994. **330**(24): p. 1710-6.

79. Glynn, J.R., E. Vynnycky, and P.E. Fine, *Influence of sampling on estimates of clustering and recent transmission of Mycobacterium tuberculosis derived from DNA fingerprinting techniques.* Am J Epidemiol, 1999. **149**(4): p. 366-71.

80. Murphy, B.M., et al., *Comparing epidemic tuberculosis in demographically distinct heterogeneous populations.* Math Biosci, 2002. **180**: p. 161-85.

81. Vynnycky, E. and P.E. Fine, *The natural history of tuberculosis: the implications of age-dependent risks of disease and the role of reinfection.* Epidemiol Infect, 1997. **119**(2): p. 183-201.

82. Chang, S.T., J.J. Linderman, and D.E. Kirschner, *Multiple mechanisms allow Mycobacterium tuberculosis to continuously inhibit MHC class II-mediated antigen presentation by macrophages.* Proc Natl Acad Sci U S A, 2005. **102**(12): p. 4530-5.

83. Vynnycky, E., et al., *Annual Mycobacterium tuberculosis infection risk and interpretation of clustering statistics.* Emerg Infect Dis, 2003. **9**(2): p. 176-83.

84. Zhang, Y., *Persistent and dormant tubercle bacilli and latent tuberculosis.* Front Biosci, 2004. **9**: p. 1136-56.

85. Saunders, B.M. and W.J. Britton, *Life and death in the granuloma: immunopathology of tuberculosis.* Immunol Cell Biol, 2007. **85**(2): p. 103-11.

86. Vynnycky, E., et al., *The effect of age and study duration on the relationship between 'clustering' of DNA fingerprint patterns and the proportion of tuberculosis disease attributable to recent transmission.* Epidemiol Infect, 2001. **126**(1): p. 43-62.

87. Grimm, V. and S.F. Railsback, *Individual-based modeling and ecology*. Princeton series in theoretical and computational biology. 2005, Princeton: Princeton University Press. xvi, 428 p.

88. Segovia-Juarez, J.L., S. Ganguli, and D. Kirschner, *Identifying control mechanisms of granuloma formation during M. tuberculosis infection using an agent-based model.* J Theor Biol, 2004. **231**(3): p. 357-76.

89. Murray, M., *Determinants of cluster distribution in the molecular epidemiology of tuberculosis.* Proc Natl Acad Sci U S A, 2002. **99**(3): p. 1538-43.

90. Nguyen, D., et al., *Genomic characterization of an endemic Mycobacterium tuberculosis strain: evolutionary and epidemiologic implications.* J Clin Microbiol, 2004. **42**(6): p. 2573-80.

91. Kempf, M.C., et al., *Long-term molecular analysis of tuberculosis strains in alabama, a state characterized by a largely indigenous, low-risk population.* J Clin Microbiol, 2005. **43**(2): p. 870-8.

92. Nikolayevskyy, V., et al., *Differentiation of tuberculosis strains in a population with mainly Beijing-family strains.* Emerg Infect Dis, 2006. **12**(9): p. 1406-13.

93. Richardson, M., et al., *Historic and recent events contribute to the disease dynamics of Beijing-like Mycobacterium tuberculosis isolates in a high incidence region.* Int J Tuberc Lung Dis, 2002. **6**(11): p. 1001-11.

94. Taylor, Z., C.M. Nolan, and H.M. Blumberg, *Controlling tuberculosis in the United States. Recommendations from the American Thoracic Society, CDC, and the Infectious Diseases Society of America.* MMWR Recomm Rep, 2005. **54**(RR-12): p. 1-81.

95. *Trends in tuberculosis--United States, 2005.* MMWR Morb Mortal Wkly Rep, 2006. **55**(11): p. 305-8.

96. Geng, E., et al., *Clinical and radiographic correlates of primary and reactivation tuberculosis: a molecular epidemiology study.* JAMA, 2005. **293**(22): p. 2740-5.

97. Yang, Z.H., et al., *Molecular epidemiology of tuberculosis in Denmark in 1992.* J Clin Microbiol, 1995. **33**(8): p. 2077-81.

98. *DP-1. Profile of general demographic characteristics: 2000 data set: Census 2000 summary file 1 (SF 1) 100-percent data geographic area: Arkansas.*

99. *Reported tuberculosis in the United States, 2003.* 2004, U.S. Department of Health and Human Services, CDC: Atlanta, Ga.

100. *Reported tuberculosis in the United States, 2002.* 2003, U.S. Department of Health and Human Services, CDC: Atlanta, Ga.

101. *Reported tuberculosis in the United States, 1996.* 1997, U.S. Department of Health and Human Services, CDC: Atlanta, Ga.

102. *Reported tuberculosis in the United States, 2001.* 2002, U.S. Department of Health and Human Services, CDC: Atlanta, Ga.

103. *Reported tuberculosis in the United States, 2000.* 2001, U.S. Department of Health and Human Services, CDC: Atlanta, Ga.

104. *Reported tuberculosis in the United States, 1998.* 1999, U.S. Department of Health and Human Services, CDC: Atlanta, Ga.

105. *Reported tuberculosis in the United States, 1999.* 2000, U.S. Department of Health and Human Services, CDC: Atlanta, Ga.

106. van Embden, J.D., et al., *Strain identification of Mycobacterium tuberculosis by DNA fingerprinting: recommendations for a standardized methodology.* J Clin Microbiol, 1993. **31**(2): p. 406-9.

107. Nguyen, D., et al., *Widespread pyrazinamide-resistant Mycobacterium tuberculosis family in a low-incidence setting.* J Clin Microbiol, 2003. **41**(7): p. 2878-83.

108. Sutherland, I., *On the risk of infection.* Bull Int Union Tuberc Lung Dis, 1991. **66**(4): p. 189-91.

109. *SAS*. 2003, SAS Institute Inc.: Cary NC.

110. Buchanan, I. *Calculating Poisson Confidence Intervals in Excel.* [cited 7/28/3006]; Available from: http://www.nwpho.org.uk/sadb/Poisson%20CI%20in%20spreadsheets.pdf

111. *Comparison of MSA Population (AR vs. US).* [cited 7/28/2006]; Available from: http://www.aiea.ualr.edu/research/demographic/population/msa.html.

112. *Selected place of birth and migration statistics for states. 1990 census special tabulations.* 1993, US Bureau of the Census, Journey to Work and Migration Statistics Branch.

113. *Racial disparities in tuberculosis--selected southeastern states, 1991-2002.* MMWR Morb Mortal Wkly Rep, 2004. **53**(25): p. 556-9.

114. Yang, Z.H., et al., *Secondary typing of Mycobacterium tuberculosis isolates with matching IS6110 fingerprints from different geographic regions of the United States.* J Clin Microbiol, 2001. **39**(5): p. 1691-5.

115. Dillaha, J.A., et al., *Transmission of Mycobacterium tuberculosis in a rural community, Arkansas, 1945-2000.* Emerg Infect Dis, 2002. **8**(11): p. 1246-8.

116. Agerton, T., et al., *Transmission of a highly drug-resistant strain (strain W1) of Mycobacterium tuberculosis. Community outbreak and nosocomial transmission via a contaminated bronchoscope.* JAMA, 1997. **278**(13): p. 1073-7.

117. Easterbrook, P.J., et al., *High rates of clustering of strains causing tuberculosis in Harare, Zimbabwe: a molecular epidemiological study.* J Clin Microbiol, 2004. **42**(10): p. 4536-44.

118. Diel, R., et al., *Ongoing outbreak of tuberculosis in a low-incidence community: a molecular-epidemiological evaluation.* Int J Tuberc Lung Dis, 2004. **8**(7): p. 855-61.

119. Freeman, R., et al., *Use of rapid genomic deletion typing to monitor a tuberculosis outbreak within an urban homeless population.* J Clin Microbiol, 2005. **43**(11): p. 5550-4.

120. Frieden, T.R., et al., *The emergence of drug-resistant tuberculosis in New York City.* N Engl J Med, 1993. **328**(8): p. 521-6.

121. Genewein, A., et al., *Molecular approach to identifying route of transmission of tuberculosis in the community.* Lancet, 1993. **342**(8875): p. 841-4.

122. Sola, C., et al., *Genetic diversity of Mycobacterium tuberculosis in Sicily based on spoligotyping and variable number of tandem DNA repeats and comparison with a spoligotyping database for population-based analysis.* J Clin Microbiol, 2001. **39**(4): p. 1559-65.

123. Supply, P., et al., *Proposal for standardization of optimized mycobacterial interspersed repetitive unit-variable-number tandem repeat typing of Mycobacterium tuberculosis.* J Clin Microbiol, 2006. **44**(12): p. 4498-510.

124. Kam, K.M., et al., *Utility of mycobacterial interspersed repetitive unit typing for differentiating multidrug-resistant Mycobacterium tuberculosis isolates of the Beijing family.* J Clin Microbiol, 2005. **43**(1): p. 306-13.

125. Thorne, N., et al., *Evolutionary clues from comparative analysis of Mycobacterium tuberculosis variable-number tandem repeat sequences within genetic families.* Infect Genet Evol, 2007. **7**(2): p. 239-46.

126. Sreevatsan, S., et al., *Restricted structural gene polymorphism in the Mycobacterium tuberculosis complex indicates evolutionarily recent global dissemination.* Proc Natl Acad Sci U S A, 1997. **94**(18): p. 9869-74.

127. Hirsh, A.E., et al., *Stable association between strains of Mycobacterium tuberculosis and their human host populations.* Proc Natl Acad Sci U S A, 2004. **101**(14): p. 4871-6.

128. Baker, L., et al., *Silent nucleotide polymorphisms and a phylogeny for Mycobacterium tuberculosis.* Emerg Infect Dis, 2004. **10**(9): p. 1568-77.

129. Gagneux, S., et al., *Variable host-pathogen compatibility in Mycobacterium tuberculosis.* Proc Natl Acad Sci U S A, 2006. **103**(8): p. 2869-73.

130. Filliol, I., et al., *Global phylogeny of Mycobacterium tuberculosis based on single nucleotide polymorphism (SNP) analysis: insights into tuberculosis evolution, phylogenetic accuracy of other DNA fingerprinting systems, and recommendations for a minimal standard SNP set.* J Bacteriol, 2006. **188**(2): p. 759-72.

131. Gutacker, M.M., et al., *Single-nucleotide polymorphism-based population genetic analysis of Mycobacterium tuberculosis strains from 4 geographic sites.* J Infect Dis, 2006. **193**(1): p. 121-8.

132. Brudey, K., et al., *Mycobacterium tuberculosis complex genetic diversity: mining the fourth international spoligotyping database (SpolDB4) for classification, population genetics and epidemiology.* BMC Microbiol, 2006. **6**: p. 23.

133. Mokrousov, I., et al., *Origin and primary dispersal of the Mycobacterium tuberculosis Beijing genotype: clues from human phylogeography.* Genome Res, 2005. **15**(10): p. 1357-64.

134. Lok, K.H., et al., *Molecular typing of Mycobacterium tuberculosis strains with a common two-band IS6110 pattern.* Emerg Infect Dis, 2002. **8**(11): p. 1303-5.

135. Yang, Z.H., et al., *Restriction fragment length polymorphism Mycobacterium tuberculosis strains isolated from Greenland during 1992: evidence of tuberculosis transmission between Greenland and Denmark.* J Clin Microbiol, 1994. **32**(12): p. 3018-25.

136. Warren, R.M., et al., *Microevolution of the direct repeat region of Mycobacterium tuberculosis: implications for interpretation of spoligotyping data.* J Clin Microbiol, 2002. **40**(12): p. 4457-65.

137. Iwamoto, T., et al., *Hypervariable loci that enhance the discriminatory ability of newly proposed 15-loci and 24-loci variable-number tandem repeat typing method on Mycobacterium tuberculosis strains predominated by the Beijing family.* FEMS Microbiol Lett, 2007. **270**(1): p. 67-74.

138. Filliol, I., et al., *Snapshot of moving and expanding clones of Mycobacterium tuberculosis and their global distribution assessed by spoligotyping in an international study.* J Clin Microbiol, 2003. **41**(5): p. 1963-70.

139. Cowan, L.S. and J.T. Crawford, *Genotype analysis of Mycobacterium tuberculosis isolates from a sentinel surveillance population.* Emerg Infect Dis, 2002. **8**(11): p. 1294-302.

140. Yang, Z., et al., *Evaluation of method for secondary DNA typing of Mycobacterium tuberculosis with pTBN12 in epidemiologic study of tuberculosis.* J Clin Microbiol, 1996. **34**(12): p. 3044-8.

141. Castro, K.G. and H.W. Jaffe, *Rationale and methods for the National Tuberculosis Genotyping and Surveillance Network.* Emerg Infect Dis, 2002. **8**(11): p. 1188-91.

142. Vitol, I., et al., *Identifying Mycobacterium tuberculosis complex strain families using spoligotypes.* Infect Genet Evol, 2006. **6**(6): p. 491-504.

143. Hunter, P.R. and M.A. Gaston, *Numerical index of the discriminatory ability of typing systems: an application of Simpson's index of diversity.* J Clin Microbiol, 1988. **26**(11): p. 2465-6.

144. Center for Disease Control., et al., *Reported tuberculosis data*, Atlanta, Ga. v.

145. Blethen, T., C. Wood, and NetLibrary Inc., *Ulster and North America transatlantic perspectives on the Scotch-Irish.* 1997, University of Alabama Press: Tuscaloosa, Ala. p. xii, 283 p.

146. Crawford, J.T., *Genotyping in contact investigations: a CDC perspective.* Int J Tuberc Lung Dis, 2003. **7**(12 Suppl 3): p. S453-7.

147. Alito, A., et al., *The IS6110 restriction fragment length polymorphism in particular multidrug-resistant Mycobacterium tuberculosis strains may evolve too fast for reliable use in outbreak investigation.* J Clin Microbiol, 1999. **37**(3): p. 788-91.

148. Warren, R.M., et al., *Evolution of the IS6110-based restriction fragment length polymorphism pattern during the transmission of Mycobacterium tuberculosis.* J Clin Microbiol, 2002. **40**(4): p. 1277-82.

149. Godfrey-Faussett, P., et al., *Evidence of transmission of tuberculosis by DNA fingerprinting.* BMJ, 1992. **305**(6847): p. 221-3.

150. Verver, S., et al., *Transmission of tuberculosis in a high incidence urban community in South Africa.* Int J Epidemiol, 2004. **33**(2): p. 351-7.

151. Godfrey-Faussett, P., et al., *Tuberculosis control and molecular epidemiology in a South African gold-mining community.* Lancet, 2000. **356**(9235): p. 1066-71.

152. Glynn, J.R., et al., *The importance of recent infection with Mycobacterium tuberculosis in an area with high HIV prevalence: a long-term molecular epidemiological study in Northern Malawi.* J Infect Dis, 2005. **192**(3): p. 480-7.

153. Fang, Z., et al., *Molecular evidence for independent occurrence of IS6110 insertions at the same sites of the genome of Mycobacterium tuberculosis in different clinical isolates.* J Bacteriol, 2001. **183**(18): p. 5279-84.

154. Yokoyama, E., et al., *Improved differentiation of Mycobacterium tuberculosis strains, including many Beijing genotype strains, using a new combination of variable number of tandem repeats loci.* Infect Genet Evol, 2007. **7**(4): p. 499-508.

155. Kremer, K., et al., *Discriminatory power and reproducibility of novel DNA typing methods for Mycobacterium tuberculosis complex strains.* J Clin Microbiol, 2005. **43**(11): p. 5628-38.

156. Supply, P., et al., *Variable human minisatellite-like regions in the Mycobacterium tuberculosis genome.* Mol Microbiol, 2000. **36**(3): p. 762-71.

157. Supply, P., et al., *Automated high-throughput genotyping for study of global epidemiology of Mycobacterium tuberculosis based on mycobacterial interspersed repetitive units.* J Clin Microbiol, 2001. **39**(10): p. 3563-71.

158. Uplekar, M.W., et al., *Attention to gender issues in tuberculosis control.* Int J Tuberc Lung Dis, 2001. **5**(3): p. 220-4.

159. Propper, C., et al., *Local neighbourhood and mental health: evidence from the UK.* Soc Sci Med, 2005. **61**(10): p. 2065-83.

160. Savine, E., et al., *Stability of variable-number tandem repeats of mycobacterial interspersed repetitive units from 12 loci in serial isolates of Mycobacterium tuberculosis.* J Clin Microbiol, 2002. **40**(12): p. 4561-6.

161. Lutong, L. and Z. Bei, *Association of prevalence of tuberculin reactions with closeness of contact among household contacts of new smear-positive pulmonary tuberculosis patients.* Int J Tuberc Lung Dis, 2000. **4**(3): p. 275-7.

162. Lemos, A.C., et al., *Risk of tuberculosis among household contacts in Salvador, Bahia.* Braz J Infect Dis, 2004. **8**(6): p. 424-30.

163. Grzybowski, S., G.D. Barnett, and K. Styblo, *Contacts of cases of active pulmonary tuberculosis.* Bull Int Union Tuberc, 1975. **50**(1): p. 90-106.

164. Behr, M.A., et al., *Transmission of Mycobacterium tuberculosis from patients smear-negative for acid-fast bacilli.* Lancet, 1999. **353**(9151): p. 444-9.

165. Sutherland, I., E. Svandova, and S. Radhakrishna, *The development of clinical tuberculosis following infection with tubercle bacilli. 1. A theoretical model for the development of clinical tuberculosis following infection, linking from data on the risk of tuberculous infection and the incidence of clinical tuberculosis in the Netherlands.* Tubercle, 1982. **63**(4): p. 255-68.

166. Sherman, L.F., et al., *Patient and health care system delays in the diagnosis and treatment of tuberculosis.* Int J Tuberc Lung Dis, 1999. **3**(12): p. 1088-95.

167. Fortun, J., et al., *Sputum conversion among patients with pulmonary tuberculosis: are there implications for removal of respiratory isolation?* J Antimicrob Chemother, 2007. **59**(4): p. 794-8.

168. Jereb, J.A., et al., *Tuberculosis morbidity in the United States: final data, 1990.* MMWR CDC Surveill Summ, 1991. **40**(3): p. 23-7.

169. Glynn, J.R., et al., *Worldwide occurrence of Beijing/W strains of Mycobacterium tuberculosis: a systematic review.* Emerg Infect Dis, 2002. **8**(8): p. 843-9.

170. Millet, J., et al., *Assessment of mycobacterial interspersed repetitive unit-QUB markers to further discriminate the Beijing genotype in a population-based study of the genetic diversity of Mycobacterium tuberculosis clinical isolates from Okinawa, Ryukyu Islands, Japan.* J Clin Microbiol, 2007. **45**(11): p. 3606-15.

171. Dormans, J., 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.* Clin Exp Immunol, 2004. **137**(3): p. 460-8.

172. Frothingham, R. and W.A. Meeker-O'Connell, *Genetic diversity in the Mycobacterium tuberculosis complex based on variable numbers of tandem DNA repeats.* Microbiology, 1998. **144 ( Pt 5)**: p. 1189-96.

173. Styblo, K., *The relationship between the risk of tuberculosis and the impact of control measures.* Bull Int Union Against Tuberculosis, 1985. **60**(117-119).

174. Rieder, H., *Annual risk of infection with Mycobacterium tuberculosis.* Eur Respir J, 2005. **25**(1): p. 181-5.

175. Beggs, C.B., et al., *The transmission of tuberculosis in confined spaces: an analytical review of alternative epidemiological models.* Int J Tuberc Lung Dis, 2003. **7**(11): p. 1015-26.

176. Horsfall, T.C., *The Improvement of the Dwellings and Surroundings of the People: The Example of Germany*. 1905, Manchester: University Press.

177. Hawker, J.I., et al., *Ecological analysis of ethnic differences in relation between tuberculosis and poverty.* BMJ, 1999. **319**(7216): p. 1031-4.

178. Antunes, J.L. and E.A. Waldman, *The impact of AIDS, immigration and housing overcrowding on tuberculosis deaths in Sao Paulo, Brazil, 1994-1998.* Soc Sci Med, 2001. **52**(7): p. 1071-80.

179. Myers, W.P., et al., *An ecological study of tuberculosis transmission in California.* Am J Public Health, 2006. **96**(4): p. 685-90.

180. Moonan, P.K., et al., *What is the outcome of targeted tuberculosis screening based on universal genotyping and location?* Am J Respir Crit Care Med, 2006. **174**(5): p. 599-604.

181. Chan-Yeung, M., et al., *Population-based prospective molecular and conventional epidemiological study of tuberculosis in Hong Kong.* Respirology, 2006. **11**(4): p. 442-8.

182. Daniel, T.M. and S.M. Debanne, *Estimation of the annual risk of tuberculosis infection for white men in the United States.* J Infect Dis, 1997. **175**(6): p. 1535-7.

183. Bates, J.H., W.W. Stead, and T.A. Rado, *Phage type of tubercle bacilli isolated from patients with two or more sites of organ involvement.* Am Rev Respir Dis, 1976. **114**(2): p. 353-8.

184. Cohen, T., et al., *Exogenous re-infection and the dynamics of tuberculosis epidemics: local effects in a network model of transmission.* J R Soc Interface, 2007. **4**(14): p. 523-31.

185. Ziegler, J.E., M.L. Edwards, and D.W. Smith, *Exogenous reinfection in experimental airborne tuberculosis.* Tubercle, 1985. **66**(2): p. 121-8.

186. McMurray, D.N., R.A. Bartow, and C.L. Mintzer, *Impact of protein malnutrition on exogenous reinfection with Mycobacterium tuberculosis.* Infect Immun, 1989. **57**(6): p. 1746-9.

187. Sutherland, I. and I. Lindgren, *The protective effect of BCG vaccination as indicated by autopsy studies.* Tubercle, 1979. **60**(4): p. 225-31.

188. Comstock, G.W., V.T. Livesay, and S.F. Woolpert, *The prognosis of a positive tuberculin reaction in childhood and adolescence.* Am J Epidemiol, 1974. **99**(2): p. 131-8.

189. Sutherland, I., *Recent studies in the epidemiology of tuberculosis, based on the risk of being infected with tubercle bacilli.* Adv Tuberc Res, 1976. **19**: p. 1-63.

190. Sutherland, I., E. Svandova, and S. Radhakrishna, *Alternative models for the development of tuberculosis disease following infection with tubercle bacilli.* Bull Int Union Tuberc, 1976. **51**(1): p. 171-9.

191. Farer, L.S., A.M. Lowell, and M.P. Meador, *Extrapulmonary tuberculosis in the United States.* Am J Epidemiol, 1979. **109**(2): p. 205-17.

192. te Beek, L.A., et al., *Extrapulmonary tuberculosis by nationality, The Netherlands, 1993-2001.* Emerg Infect Dis, 2006. **12**(9): p. 1375-82.

193. Noertjojo, K., et al., *Extra-pulmonary and pulmonary tuberculosis in Hong Kong.* Int J Tuberc Lung Dis, 2002. **6**(10): p. 879-86.

194. Yang, Z., et al., *Identification of risk factors for extrapulmonary tuberculosis.* Clin Infect Dis, 2004. **38**(2): p. 199-205.

195. National Center for Prevention Services (U.S.). Division of Tuberculosis Elimination., *Reported tuberculosis in the United States*, Atlanta, Ga.: U.S. Dept. of Health and Human Services, Public Health Service, Centers for Disease Control and Prevention, National Center for Prevention Services. v.

196. Mehta, J.B., et al., *Epidemiology of extrapulmonary tuberculosis. A comparative analysis with pre-AIDS era.* Chest, 1991. **99**(5): p. 1134-8.

197. Gonzalez, O.Y., et al., *Extra-pulmonary manifestations in a large metropolitan area with a low incidence of tuberculosis.* Int J Tuberc Lung Dis, 2003. **7**(12): p. 1178-85.

198. Conde, M.B., et al., *Yield of sputum induction in the diagnosis of pleural tuberculosis.* Am J Respir Crit Care Med, 2003. **167**(5): p. 723-5.

199. Yeager, H., Jr., et al., *Quantitative studies of mycobacterial populations in sputum and saliva.* Am Rev Respir Dis, 1967. **95**(6): p. 998-1004.

200. *Control and prevention of tuberculosis in Britain: an updated code of practice. Subcommittee of the Joint Tuberculosis Committee of the British Thoracic Society.* BMJ, 1990. **300**(6730): p. 995-9.

201. Godoy, P., et al., *Characteristics of tuberculosis patients with positive sputum smear in Catalonia, Spain.* Eur J Public Health, 2004. **14**(1): p. 71-5.

202. Golub, J.E., et al., *Patient and health care system delays in pulmonary tuberculosis diagnosis in a low-incidence state.* Int J Tuberc Lung Dis, 2005. **9**(9): p. 992-8.

203. Sarmiento, K., et al., *Help-seeking behavior of marginalized groups: a study of TB patients in Harlem, New York.* Int J Tuberc Lung Dis, 2006. **10**(10): p. 1140-5.

204. Menzies, D., *Effect of treatment on contagiousness of patients with active pulmonary tuberculosis.* Infect Control Hosp Epidemiol, 1997. **18**(8): p. 582-6.

205. Long, R., et al., *Relative versus absolute noncontagiousness of respiratory tuberculosis on treatment.* Infect Control Hosp Epidemiol, 2003. **24**(11): p. 831-8.

206. Guler, M., et al., *Factors influencing sputum smear and culture conversion time among patients with new case pulmonary tuberculosis.* Int J Clin Pract, 2007. **61**(2): p. 231-5.

207. de Boer, A.S., et al., *Exogenous re-infection as a cause of recurrent tuberculosis in a low-incidence area.* Int J Tuberc Lung Dis, 2003. **7**(2): p. 145-52.

208. Gutacker, M.M., et al., *Genome-wide analysis of synonymous single nucleotide polymorphisms in Mycobacterium tuberculosis complex organisms: resolution of genetic*

*relationships among closely related microbial strains.* Genetics, 2002. **162**(4): p. 1533-43.

209. Gutierrez, M.C., et al., *Ancient origin and gene mosaicism of the progenitor of Mycobacterium tuberculosis.* PLoS Pathog, 2005. **1**(1): p. e5.

210. Kapur, V., T.S. Whittam, and J.M. Musser, *Is Mycobacterium tuberculosis 15,000 years old?* J Infect Dis, 1994. **170**(5): p. 1348-9.

211. Tsolaki, A.G., et al., *Functional and evolutionary genomics of Mycobacterium tuberculosis: insights from genomic deletions in 100 strains.* Proc Natl Acad Sci U S A, 2004. **101**(14): p. 4865-70.

212. Thierry, D., et al., *IS6110, an IS-like element of Mycobacterium tuberculosis complex.* Nucleic Acids Res, 1990. **18**(1): p. 188.

213. Sampson, S.L., et al., *Disruption of coding regions by IS6110 insertion in Mycobacterium tuberculosis.* Tuber Lung Dis, 1999. **79**(6): p. 349-59.

214. McHugh, T.D. and S.H. Gillespie, *Nonrandom association of IS6110 and Mycobacterium tuberculosis: implications for molecular epidemiological studies.* J Clin Microbiol, 1998. **36**(5): p. 1410-3.

215. Barnes, P.F. and M.D. Cave, *Molecular epidemiology of tuberculosis.* N Engl J Med, 2003. **349**(12): p. 1149-56.

216. Warren, R.M., et al., *Calculation of the stability of the IS6110 banding pattern in patients with persistent Mycobacterium tuberculosis disease.* J Clin Microbiol, 2002. **40**(5): p. 1705-8.

217. Nguyen, L.N., G.L. Gilbert, and G.B. Marks, *Molecular epidemiology of tuberculosis and recent developments in understanding the epidemiology of tuberculosis.* Respirology, 2004. **9**(3): p. 313-9.

218. Yeh, R.W., et al., *Stability of Mycobacterium tuberculosis DNA genotypes.* J Infect Dis, 1998. **177**(4): p. 1107-11.

219. Wall, S., et al., *Context-sensitive transposition of IS6110 in mycobacteria.* Microbiology, 1999. **145 ( Pt 11)**: p. 3169-76.

220. Hermans, P.W., et al., *Insertion element IS987 from Mycobacterium bovis BCG is located in a hot-spot integration region for insertion elements in Mycobacterium tuberculosis complex strains.* Infect Immun, 1991. **59**(8): p. 2695-705.

221. Groenen, P.M., et al., *Nature of DNA polymorphism in the direct repeat cluster of Mycobacterium tuberculosis; application for strain differentiation by a novel typing method.* Mol Microbiol, 1993. **10**(5): p. 1057-65.

222. Eldholm, V., et al., *A first insight into the genetic diversity of Mycobacterium tuberculosis in Dar es Salaam, Tanzania, assessed by spoligotyping.* BMC Microbiol, 2006. **6**: p. 76.

223. Fleischmann, R.D., et al., *Whole-genome comparison of Mycobacterium tuberculosis clinical and laboratory strains.* J Bacteriol, 2002. **184**(19): p. 5479-90.

224. Burman, W.J., et al., *DNA fingerprinting with two probes decreases clustering of Mycobacterium tuberculosis.* Am J Respir Crit Care Med, 1997. **155**(3): p. 1140-6.

225. Frothingham, R., *Differentiation of strains in Mycobacterium tuberculosis complex by DNA sequence polymorphisms, including rapid identification of M. bovis BCG.* J Clin Microbiol, 1995. **33**(4): p. 840-4.

226. Gopaul, K.K., et al., *Progression toward an improved DNA amplification-based typing technique in the study of Mycobacterium tuberculosis epidemiology.* J Clin Microbiol, 2006. **44**(7): p. 2492-8.

227. Kremer, K., et al., *Definition of the Beijing/W lineage of Mycobacterium tuberculosis on the basis of genetic markers.* J Clin Microbiol, 2004. **42**(9): p. 4040-9.

228. Milan, S.J., et al., *Expanded geographical distribution of the N family of Mycobacterium tuberculosis strains within the United States.* J Clin Microbiol, 2004. **42**(3): p. 1064-8.

229. Levinson, G. and G.A. Gutman, *Slipped-strand mispairing: a major mechanism for DNA sequence evolution.* Mol Biol Evol, 1987. **4**(3): p. 203-21.

230. Mizrahi, V. and S.J. Andersen, *DNA repair in Mycobacterium tuberculosis. What have we learnt from the genome sequence?* Mol Microbiol, 1998. **29**(6): p. 1331-9.