Analysis and application of evolutionary markers in the epidemiology of *Mycobacterium tuberculosis*

Gian Dreyer van der Spuy

Dissertation presented for the degree of Doctor of Philosophy at
Stellenbosch University

Promoters: *Prof. RM Warren*
Prof. PD van Helden

December 2008

Declaration

By submitting this thesis electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the owner of the copyright thereof (unless to the extent explicitly otherwise stated) and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

Date: December 2008

Abstract

This series of studies includes both methodological analyses, aimed at furthering our understanding of, and improving the tools used in molecular epidemiology, and investigative projects which have used these tools to add to our knowledge of the *M. tuberculosis* epidemic.

Using serial isolates from tuberculosis patients, we have investigated the evolutionary rate of the IS*6110* RFLP pattern. In accordance with other studies, we determined a ½-life for this epidemiological marker of 10.69 years, confirming its appropriateness for this purpose. We also identified an initial, much higher apparent rate which we proposed was the result of pre-diagnostic evolution. In support of this, our investigations in the context of household transmission of *M. tuberculosis* revealed that IS*6110*-based evolution is closely associated with transmission of the organism, resulting in a strain population rate of change of 2.9% per annum.

To accommodate evolution within estimates of transmission, we proposed that calculations incorporate the concept of Nearest Genetic Distance (cases most similar in RFLP pattern and most closely associated in time). We used this to create transmission chains which allowed for limited evolution of the IS*6110* marker. As a result, in our study community, the estimated level of disease attributable to ongoing transmission was increased to between 73 and 88% depending on the Genetic Distance allowed.

We identified the duration of a study as a further source of under-estimation of transmission. This results from the artefactual abridgement of transmission chains caused by the loss of cases at the temporal boundaries of a study. Using both real and simulated data, we showed that viewing a 12-year study through shorter window periods dramatically lowered estimates of transmission. This effect was negatively correlated with the size of a cluster.

Various combinations of MIRU-VNTR loci have been proposed as an alternative epidemiological marker. Our investigations showed that, while this method yielded estimates of transmission similar to those of IS*6110*, there was discordance between the two markers in the epidemiological linking of cases as a result of their independent evolution. Attempting to compensate for this by allowing for evolution during transmission improved the performance of IS*6110*, but generally had a deleterious effect of that of MIRU-VNTR. However, this marker remains a valuable tool for higher phylogenetic analysis and we used it to demonstrate a correlation between sublineages of the Beijing clade and the regions in which they are found. We proposed that, either the host population had selected for a particular sublineage, or that specific sublineages had adapted to be more successful in particular human populations.

We further explored the dynamics of the epidemic over a 12-year period in terms of the five predominant *M. tuberculosis* clades. We found that, while four of these clades remained relatively stable, the incidence of cases from the Beijing clade increased exponentially. This growth was attributed to drug-sensitive cases although drug-resistant Beijing cases also appeared to be more successful than their non-Beijing counterparts. Possible factors contributing to this clade's success were a greater proportion of positive sputum smears and a lower rate of successful treatment.

Oorsig

Hierdie reeks studies bevat beide metodologiese analises, gerig op die uitbreiding van ons kennis van die metodes in gebruik in molekulêre epidemiologie en die verbetering daarvan, en ondersoekende projekte wat hierdie metodes gebruik om by te dra tot ons verstaan van die *M. tuberculosis* epidemie.

Ons het die evolusionêre tempo van die IS*6110* RFLP patroon ondersoek deur gebruik te maak van opeenvolgende isolate vanaf tuberkulose pasiënte. In ooreenstemming met ander studies, het ons die ½-leeftyd van hierdie epidemiologiese merker bepaal as 10.69 jaar, wat die geskiktheid daarvan vir hierdie doel bevestig het. Ons het ook `n aanvanklike, veel groter klaarblyklike tempo geïdentifiseer, wat ons voorgestel het afkomstig was van pre-diagnostiese evolusie. Ter ondersteuning hiervan, het ons ondersoeke in die konteks van huishoudelike oordrag van *M. tuberculosis* getoon dat IS*6110*-gebaseerde evolusie sterk geassosieer is met die oordrag van die organisme, wat lei tot `n raspopulasie tempo van verandering van 2.9% per jaar.

Om evolusie binne die berekening van oordrag in te sluit, het ons voorgestel dat berekeninge die konsep van Naaste Genetiese Afstand (gevalle wat die mees soortgelyk is in RFLP patroon en naaste geassosieer is in tyd) moet inkorporeer. Ons het dit gebruik om oordragkettings te skep wat beperkte evolusie van die IS*6110* merker toelaat. Die resultaat daarvan, in ons studie gemeenskap, is dat die beraamde vlak van siekte wat toegeskryf kan word aan deurlopende oordrag, verhoog is na tussen 73 en 88%, afhangende van die Genetiese Afstand wat toegelaat is.

Ons het die lengte van `n studie as `n verdere bron van onderberaming van die hoeveelheid oordrag geïdentifiseer. Dit is die resultaat van die artefaktuele verkorting van oordragkettings wat veroorsaak is deur die verlies van gevalle by die temporale grense van `n studie. Deur gebruik te maak van beide werklike en nagebootsde data, kon ons aantoon dat berekeninge van oordrag dramaties verlaag is deur `n 12-jaar studie in korter vensterperiodes te besigtig. Hierdie effek is negatief gekorreleer met die grootte van `n stam-groep.

Verskeie kombinasies van MIRU-VNTR lokusse is voorgestel as `n alternatiewe epidemiologiese merker. Ons ondersoeke het getoon dat, alhoewel hierdie metode bermings van oordrag soortgelyk aan die van IS*6110* gelewer het, daar tog onenigheid tussen die twee merkers in die epidemiologiese verbinding van gevalle was as gevolg van hul onafhanklike evolusie.

In `n poging om hiervoor te kompenseer, is daar vir evolusie gedurende oordrag voorsiening gemaak. Alhoewel die prestasie van IS*6110* hierdeur verbeter is, het dit oor die algemeen `n nadelige effek op die prestasie van die MIRU-VNTR gehad. Ondanks dit, bly hierdie merker `n waardevolle metode vir hoër filogenetiese analises en het ons daarvan gebruik gemaak om `n korrelasie tussen sublyne van die Beijing groep en die areas waarin dit voorkom, te maak. Ons stel twee moontlike verklaarings voor: dat die gasheerpopulasie het vir `n spesifieke sublyn geselekteer, óf dat spesifieke sublyne aangepas het om meer suksesvol in die onderskeie menspopulasies te wees.

Ons het die dinamika van die epidemie verder oor `n 12-jaar periode ondersoek in terme van die vyf oorheersende *M. tuberculosis* groepe. Ons het gevind dat, terwyl vier van hierdie groepe relatief stabiel gebly het, die voorkoms van gevalle van die Beijing groep eksponensieël vermeerder het. Hierdie groei was te wyte aan geneesmiddel-sensitiewe gevalle, alhoewel middelweerstandige Beijing-gevalle ook geblyk het om meer suksesvol te wees as die nie-Bejing groepe. `n Groter aandeel van positiewe sputum smere en `n laer koers van suksesvolle behandeling is as moontlike faktore wat bydra tot hierdie groep se sukses geïdentifiseer.

Acknowledgements

I would like to express my sincere thanks and appreciation to the following people, without whom this thesis, and the studies comprising it would not have been possible.

Firstly, to Prof. Rob Warren, who acted my promoter for this degree, for his keen insight and the guidance and motivation he provided during the course of these studies, all of which have helped me to become a better scientist.

Secondly, to Prof. Paul van Helden, in whose department this work took place, for the opportunity to study for this degree, for his support as co-promoter and his encouragement to grow and the space to do so.

Thanks are also due to the many, often unsung field- and lab-workers who must necessarily form part of any epidemiological team and without whom there would be nothing to analyse.

To the various funders of these projects: GlaxoSmithKline Action TB Initiative, the Sequella Global Tuberculosis Foundation, the National Research Foundation, the European Commission 6th framework programme for research and technological development and the Harry Crossley Foundation.

Finally, to my colleagues in the MRC Centre for Molecular and Cellular Biology, in particular, Eileen van Helden and Cedric Werely, for advice and encouragement along the way and for diversionary discussions which helped to keep me sane.

Shun no toil to make yourself remarkable by some talent or other; yet do not devote yourself to one branch exclusively. Strive to get clear notions about all. Give up no science entirely; for science is but one.

Lucius Annaeus Seneca

Table of Contents

For Dorothy, who makes it all worthwhile.

Preface

It is estimated by the WHO that roughly 1/3 of the world's population is infected with *Mycobacterium tuberculosis* and that, currently, 2 million people die annually as a result of this disease. Evidence for the existence of tuberculosis as a human pathogen dates back as far as 4000 BC, tubercular decay having been detected in skeletal remains and in the spines of Egyptian mummies. Despite its long history, it was only with the discovery of *M. tuberculosis* as the causative agent of tuberculosis in 1882 by Robert Koch, made possible by advances in bacteriological techniques, that progress began to be made in understanding the disease. However, it was only with the development of molecular epidemiological tools, subsequent to the 'molecular biology revolution', that a clearer picture began to emerge as to the dynamics of the tuberculosis within its host population.

Chapter 1 gives an overview of the various molecular markers currently used in tuberculosis epidemiology in different contexts, as well as a number that have been previously used and some prospective future markers which have recently been proposed.

The subsequent seven chapters describe methodological and analytical studies using the tools currently available to molecular epidemiology as it applies to tuberculosis. The community from which the data used in this series of studies is derived, comprises a population of approximately 36 300 according to census data provided by Statistics South Africa. There is a high incidence of tuberculosis, with an average of 320 new, bacteriologically-confirmed, adult cases per 100 000 population reported per annum.

As a thorough and accurate understanding of any scientific tool is essential, the first four studies presented in this thesis comprise a number of investigations into the nature of the most commonly used epidemiological marker in tuberculosis: the insertion element, IS*6110.* The first two (Chapters 2 and 3) examine the stability of this marker in the context of individual patients and during transmission respectively. Having established an evolutionary rate for IS*6110,* Chapter 4 proposes a method whereby minor changes in the marker, in the context of recently transmitted disease, might be incorporated into the concept of a chain of transmission for the purposes of epidemiological calculations.

A number of logistical issues beset any epidemiological investigation. One of these is the difficulty in describing an epidemic of a complex disease like tuberculosis, which has a lengthy incubation interval and may remain dormant for long periods, from data derived from a study of finite duration. Chapter 5 examines the effect of study duration on the standard calculations of ongoing or recent transmission of tuberculosis and the identification of cases having unique strain-types.

With the nature and use of the IS*6110* marker established by the preceding studies and those of other investigators in this field, Chapter 6 examines the structure and dynamics of the epidemic in the aforementioned study community in terms of the prevalent strain-clades. The Beijing clade is noted as being particularly successful and possible reasons for this are presented.

The origins of tuberculosis in South Africa are various, however, the Beijing clade is generally accepted to have arrived with the importation of slave labour from the Far-East. As noted in the preceding chapter, this clade has been particularly successful the study community, which consists largely of people of the Cape Coloured ethnic group. Chapter 7 addresses this observation using the more recent Mycobacterial Interspersed Repetitive Unit (MIRU) marker to show an association between the frequency of occurrence of strains from defined Beijing clade sublineages and the human host population.

Use of the MIRU marker is increasing and, while comparative studies between it and IS*6110* have been done, these have focused on communities with a low-incidence of tuberculosis or those having a high proportion immigrant population. Chapter 8, therefore, compares the molecular epidemiological analyses using IS*6110* and various combinations of MIRU marker sets in data from a high-incidence community to evaluate the usefulness of the latter in this context.

This series of studies has substantially enhanced our understanding of tuberculosis, successfully challenged pre-existing dogma and helped to define some of the tools used in the field of molecular epidemiology and, as such, has laid the foundation for future studies in this arena.

1

Molecular epidemiology of *Mycobacterium tuberculosis*

Van der Spuy G. D. and Warren R. M.

Handbook of tuberculosis: clinics, diagnostics, therapy and epidemiology (2008) Ch 3, pp 41-62
Wiley-VCH Verlag GmbH & Co, Weinheim

Table of Contents

Introduction

In order to control the epidemic it is essential to understand the epidemic.

Classical epidemiology, *"the branch of medicine that deals with the study of the causes, distribution, and control of disease in populations"* has, and continues to provide insights into the disease dynamics and the factors driving the TB epidemic. For the first 80 years following the discovery of the causative agent, epidemiologists had no tools to study the genetics of *Mycobacterium tuberculosis* and therefore, with the exception of phenotypic characteristics such as colony morphology, growth rates and drug resistance patterns, investigations largely ignored the bacterial component of the epidemic.

In the absence of genetic information, numerous assumptions were made in order to facilitate epidemiological analysis. Foremost among these was the concept that TB is caused by a single, primary infection. Any recurrence of disease after cure was regarded as a relapse of the same infection [1]. Furthermore, it was understood that the transmission of *M. tuberculosis* occurred under conditions of close contact, pointing to the household as a primary focal point for the spread of the disease [2]. It was also accepted that most cases of drug resistant TB arose as a result of the acquisition of resistance due to non-compliance with the treatment regimen [3]. Many of these assumptions have become entrenched as dogma and have formed the basis of our understanding and management of the disease.

It was only with the discovery of phage typing methods [4] that proof of pathogen diversity was obtained and it was recognised that the epidemic was probably not caused by a single genetic entity. Despite the limited resolution of phage typing, in that it only allows for the differentiation of *M. tuberculosis* into three groups, the method immediately challenged certain dogma and was the first method used to investigate the mechanism leading to recurrent disease. It revealed that the epidemic was a composite of different groups of *M. tuberculosis* and that an individual TB patient could harbour more than one strain.

It took a further 30 years and the development of molecular biological tools before the true extent of genetic heterogeneity was discovered in the species *M. tuberculosis*. This culminated in the birth of molecular epidemiology: *"the application of molecular biology to the answering of epidemiological questions"*. Essentially, molecular epidemiology is a comparative science which aims to identify epidemiological relationships between patients with diagnosed TB through comparison of the genotypes of the disease-causing bacteria.

The Purpose of Molecular Epidemiology

From the onset, molecular epidemiological studies have challenged classical dogma, thereby providing new insights into the true nature of the disease. Such knowledge has been used to inform and develop policy in many countries thereby resulting in the implementation of infrastructure to ensure more effective TB control strategies.

The objective of molecular epidemiology is to complement classical epidemiology by the use of molecular tools, tracking the movement of strains through space and time, thereby enhancing the accuracy and resolution of the epidemiological picture. To a large degree, this is achieved by distinguishing bacterial strains on the basis of genetic differences. The ideal approach would be by whole genome sequence comparison. However, despite major advances in DNA sequencing technology, which may well make this feasible in future, this is currently not a practical solution. Thus, we are compelled to use genetic fingerprinting methods that rely on observing a small subset of *Mycobacterial* genome dynamics. These techniques therefore act as surrogates for the underlying genetic evolution of the bacterial strains. To the extent that they are able to accurately reflect changes within the entire genome, the methods discussed below serve as useful tools to define clonality.

Requirements for the Successful Application of Molecular Epidemiology

The Marker

The inferences drawn from molecular epidemiological data are only as valid as the inherent limitations of the biological and analytical tools used to inform them. These molecular tools present an indirect window onto selected aspects of genomic dynamics and are thus markers of evolutionary change at the DNA level. There is a great deal of diversity between the different molecular markers, which affects their suitability for various applications. The features required of a marker used for the phylogenetic reconstruction of ancient lineages, for example, will differ from those required for the purposes of characterisation and geo-temporal tracking of an ongoing epidemic. In addition, studies attempting to answer specific questions or focussing on particular bacterial sub-populations may make different demands of a molecular marker.

An epidemiological molecular marker should have attributes which permit the discrimination, and thus the tracking, of distinct bacterial sub-populations or strains. This requires that it should evolve at a rate which is reliably predictable and sufficiently rapid so as to distinguish between epidemiologically unrelated cases. At the same time, the rate of change should be slow enough that the marker patterns of bacteria isolated from patients forming part of a chain of transmission will appear identical (or at least, highly similar). This ideal rate will naturally vary between pathogens, but in the case of *M. tuberculosis*, for which the definition of recent transmission allows for a latency interval of up to two years between infection and disease onset, would be slightly longer than this period. A further requirement is that the mechanism of marker evolution should not favour convergence and the number of permutations possible should be great enough to make this unlikely.

The Analytical Tools

To facilitate the meaningful comparison of molecular epidemiological data from different regions, genotype data must be compatible. This requires, firstly, that the laboratory techniques for producing the data be standardised as has been done for a number of currently used markers [5-7]. Secondly, the ability to share data necessitates a standardised classification scheme and, preferably, compatible analytical software tools. International repositories of easily accessible, shared data do much to foster collaboration, facilitate regional or global studies and encourage the implementation and maintenance of standards.

Genotyping Methods

Repetitive Sequences

The earliest methods for genotyping *M. tuberculosis* were based on short, repetitive DNA sequences found scattered throughout the genome. The attribute that makes these elements useful as markers derives from their ability to alter the number of tandem repeats of which they are comprised, causing each element to vary in length. Five types of variable number tandem repetitive elements have been thus far identified and used for *M. tuberculosis* genotyping. Of these, the polymorphic GC-rich repetitive sequence (PGRS) [8], the GTG triplet repeat [9] and the major polymorphic tandem repeat (MPTR) [10] are found in multiple genomic clusters and exist as imperfectly repeated units. Further genomic loci have been identified containing tandem repeats of identical DNA sequence comprising the exact tandem repeat (ETR) elements [11,12]. The last repetitive sequence is a series of 36 bp directly repeated elements which are found at a single locus and are interspersed by unique 35 to 41 bp spacer sequences [7,13].

Polymorphic GC-rich Repetitive Sequence Typing

The PGRS sequences were first identified by de Wit *et al.* in 1990 [14]. These elements are characterised by a 96 bp consensus repeat sequence which is found at a variable number of loci within the *M. tuberculosis* genome (61 loci in H37Rv) associated with the PE gene family [15]. In the PGRS Restriction Fragment Length Polymorphism (RFLP) genotyping method, developed by Ross *et al.* [16], a cloned chromosomal domain containing a PGRS repeat sequence is Southern Hybridized to *Alu*I restricted chromosomal DNA isolated from clinical isolates of *M. tuberculosis*. This method produces a highly complex banding pattern which has been found to be identical (or very similar) in epidemiologically related cases and varies between unrelated cases. Application of this method as a secondary typing tool has shown that the method can be more discriminatory than IS*6110* DNA fingerprinting. This is particularly true for *M. tuberculosis* isolates harbouring less than six IS*6110* elements (low copy-number strains). However, the PGRS typing method has not found favour with molecular epidemiologists, primarily because it has not been standardised. Furthermore, the method is time consuming and labour intensive, requiring culture and DNA extraction. The banding patterns are also extremely complex, with variation in intensity, making reproducible scoring difficult and thus militating against computerised comparison.

Variable Number Tandem Repeats (VNTR)

Frothingham and Meeker-O'Connell originally described the polymerase chain reaction (PCR)-based VNTR typing technique in which unique ETR sequences from five different chromosomal loci are amplified using locus-specific primer sets complementary to the chromosomal domains flanking the respective repeat sequences [12] (see Figure 1). The ETR sequences range in length from 53 to 79 bp. The genotype of *M. tuberculosis* isolates is represented by a five-digit allele profile specifying the number of repeats at each locus, determined from the size of the respective amplification products. The advantage of this method is that crude DNA from *M. tuberculosis* cultures can be rapidly amplified thereby allowing for prospective, high-throughput genotyping. This method has the potential to allow molecular epidemiology to direct TB control in real time. However, it must be acknowledged that the five ETR loci show only limited variability in clinical isolates thereby restricting their use as markers in molecular epidemiological studies.

MIRU-VNTR

Following the release of the whole genome sequence of the *M. tuberculosis* H37Rv strain [17], Supply *et al.* identified 41 variable tandem repeat sequences (consisting of ETR and MPTR elements) which they termed Mycobacterial Interspersed Repetitive Units (MIRU's) for use as possible markers for genotyping [6]. These repeat sequences range in length from 40 to 100 bp and bear many similarities to eukaryotic minisatellites. Initial studies identified twelve MIRU loci, containing between two and eight repeat elements, which were shown to be polymorphic in clinical isolates. Extensive analysis of the allelic diversity of these twelve MIRU loci has been done in many different settings and the results have been compared to the "gold standard" IS*6110* genotyping method. It is generally accepted that the discriminatory power of MIRU typing, using the twelve-allele format, is lower than that of IS*6110* genotyping. This may in part be explained by the slow evolutionary rate of the different loci in comparison to IS*6110* transposition. This is particularly true for strains with more than six IS*6110* insertions, however, MIRU typing shows greater differentiation of strains with six or fewer IS*6110* insertions. To date, the epidemiological significance of strain genotypes defined by MIRU typing remains largely unknown. Furthermore, concern has been raised about possible convergence, and thus the usefulness of these markers in phylogenetic studies [18].

More recently, investigators have included additional VNTR sequences and have shown that the discriminatory power is proportional to the number of alleles analysed. However, the different nomenclatures used to describe the various VNTR loci has lead to a certain amount of confusion. Current recommendations suggest the use of 15 MIRU-VNTR loci for molecular epidemiological studies and two loci for phylogenetic analysis [19]. Before these recommendations can be adopted it will be necessary to fully evaluate their performance in different settings, including high and low incidence communities. The large number of alleles recommended make the genotyping method rather cumbersome and time consuming. To streamline the method a multiplex PCR system using fluorescently labelled primers in combination with capillary fractionation has been described [20].

Spoligotyping

Hermans *et al.* [13] first described the Direct Repeat (DR) region based on genetic analysis of *M. bovis* BCG and it was later suggested that this locus may be informative for epidemiological studies of the *M. tuberculosis* complex [7]. The DR region consists of multiple conserved 36 bp directly-repeated sequences (DRs) interspersed by non-repetitive, unique spacer sequences ranging from 35 to 41 bp in length. One DR and its neighbouring spacer sequence are termed a direct variable repeat (DVR). Polymorphisms in this region arise from homologous recombination between neighbouring or distant DRs or adjacent IS*6110* elements, IS*6110* insertions and single nucleotide polymorphisms (SNPs) in the DR's spacer sequences [21,22].

The internationally standardised method for the analysis of the DR region has been termed Spoligotyping [7]. This is a PCR-based technique designed to determine the presence or absence of a set of 43 spacer sequences. The primers (of which one is biotinylated) are complementary to the DR sequence and allow the amplification of the spacers between the target DRs. The amplification products are hybridised to a set of immobilised, complementary oligonucleotides and the presence of each spacer sequence is subsequently detected by chemiluminesence (see Figure 1). Strains can be differentiated according to the observed hybridisation pattern which indicates the presence or absence of the individual DVRs. An octal coding system has been adopted to facilitate the recording and collaborative exchange of strain types [23].

Spoligotyping is a simple technique which is highly reproducible. The discriminatory power is, however, significantly lower than that of IS*6110* RFLP, except when the insertion element is present in fewer than six copies. Despite these limitations, this method allows for the rapid genotyping of clinical isolates using relatively crude DNA isolated from culture or Ziehl-Neelsen-positive slides.

Insertion Sequences

IS6110 Restriction Fragment Length Polymorphism (RFLP)

The most epidemiologically informative repeated sequence in general use is the transposable element IS*986* (more commonly known as IS*6110*). An internationally standardised method using this sequence led to the birth of modern molecular epidemiology of TB which has allowed for the analysis of the epidemic on both community and global scales [24].

IS*6110* genotyping relies on the ability of the element both to replicate itself randomly into the genome at different positions and to excise itself. These mechanisms allow for a theoretically infinite number of combinations of IS*6110* elements inserted at different loci around the chromosome. In practice, the number of IS*6110* elements found in clinical isolates appears to be limited to about 26 and their informativeness is restricted by the resolution of the RFLP technology.

The IS*6110* element is characterised by an imperfect 28 bp terminal repeat sequence and a single *Pvu*II restriction site. On *Pvu*II digestion of the chromosomal DNA, the IS*6110* element is split into two domains, each attached to their respective adjacent genomic segments. The *Pvu*II restricted DNA is then electrophoretically fractionated on an agarose gel, Southern transferred to a nylon membrane and hybridised with a labelled probe complementary to the 3'-domain of IS*6110* (see Figure 1). The resulting banding pattern is visualized by autoradiography and is a measure of the number of IS*6110* elements in the genome and their distance from their adjacent chromosomal 3' *Pvu*II restriction sites.

Over a broad range of geographic settings and strains, IS*6110* genotyping shows the greatest discriminatory power of all the markers currently in general use and is regarded as the "gold standard" for *M. tuberculosis* genotyping. The method is internationally standardised which allows for inter-laboratory comparisons using specialised software. However, it is a cumbersome and time consuming technique, requiring prior culture of clinical isolates and purification of large quantities of good quality DNA. The result is that molecular epidemiological data can only be analysed retrospectively.

The information contained within the banding pattern, and thus the discriminatory power of the technique, is dependent on the number of IS*6110* elements present in the genome of the *M. tuberculosis* isolate. Epidemiological relationships can only be confidently inferred when more than six IS*6110* elements are present and therefore, classification of strains having fewer copies must depend on secondary typing methods.

As with all genotyping markers, evolution may occur during transmission, thereby complicating the interpretation of the data. Estimates of stability of the IS*6110* banding pattern suggest that it is sufficiently stable to infer epidemiological contact [25]. However, the observation of IS*6110* banding pattern changes over the course of ongoing transmission has led to the suggestion that epidemiological calculations based on this marker should account for banding pattern evolution [26].

Mixed Linker

This is a rapid, PCR-based technique which measures the position of IS*6110* elements relative to adjacent restriction sites [27]. The method entails restricting the bacterial genome with *Hha*I followed by ligating a linker oligonucleotide, in which the thymidine residues have been replaced by uracil in one strand, to the ends of the restriction fragments. The DNA is then treated with N-glycosylase to eliminate the uracil-containing oligonucleotides. The remaining fragments are PCR-amplified using primers complementary to the linker and to IS*6110*, thereby generating fragments corresponding to the size of the adjacent chromosomal domains. The primary advantage of this technique is the ability to obtain DNA fingerprints without the requirement of first culturing the organism. In a comparative methodological study, mixed-linker PCR was shown to have only slightly less discriminatory power than that of IS*6110* genotyping. However, the method has not been internationally standardised and is seldom used for molecular epidemiological studies.

Single nucleotide polymorphisms (SNPs)

Comparative genomics based on whole genome sequencing has demonstrated a remarkable degree of conservation between the genome sequences of various strains of *M. tuberculosis*. This sequencing data has identified three groups of single nucleotide polymorphisms (SNPs). Non-synonymous SNPs (nsSNP) are often associated with amino acid changes which may be subject to various selection pressures. Intergenic SNPs may be subject to selective pressure as they may affect gene expression. Synonymous SNPs (sSNP) do not alter the amino acid sequence and are therefore generally considered neutral to selective pressure. As such, they provide a powerful tool for general molecular epidemiology. A limitation in the use of SNPs is that studies based on this technique suffer from ascertainment bias (*i.e.* analyses based on such a dataset are skewed by the non-random nature of the selection of the discriminating features). This problem can only be overcome by using many SNPs identified from a wide diversity of strain sequences. However, even with large numbers of sSNPs it is uncertain whether the information generated would be sufficient to differentiate closely related strains for epidemiological calculations.

FAFLP

Fluorescence Amplified-Length Polymorphism (FAFLP) is an extension of the AFLP technique [28] which shows much promise. It was first applied to the question of *M. tuberculosis* strain typing by Goulding *et al* [29]. Briefly, the technique involves restricting *M. tuberculosis* genomic DNA, usually with two restriction enzymes, ligation of linkers onto the restricted DNA, followed by PCR amplification of the fragments with fluorescently labelled primers. Discrimination may be further enhanced by the use of four primers which differ by one base in the position adjacent to the restriction site, each of which is labelled with a different fluorescent marker. The fluorescently labelled amplified fragments are then sized using an automated sequencer. FAFLP is based on the detection of random, rather than selected SNPs, as well as a variety of other genomic events, thereby avoiding the problem of ascertainment bias associated with SNP typing. This also minimises the adverse affects of the often complex behaviour of genomic elements such as insertion sequences and minisatellites. The method appears to have discriminatory power at least equal to that of IS*6110* RFLP, depending on how it is applied and has been successfully used in a number of studies. However, the technique still requires further development, characterisation and standardisation before it is more generally accepted.

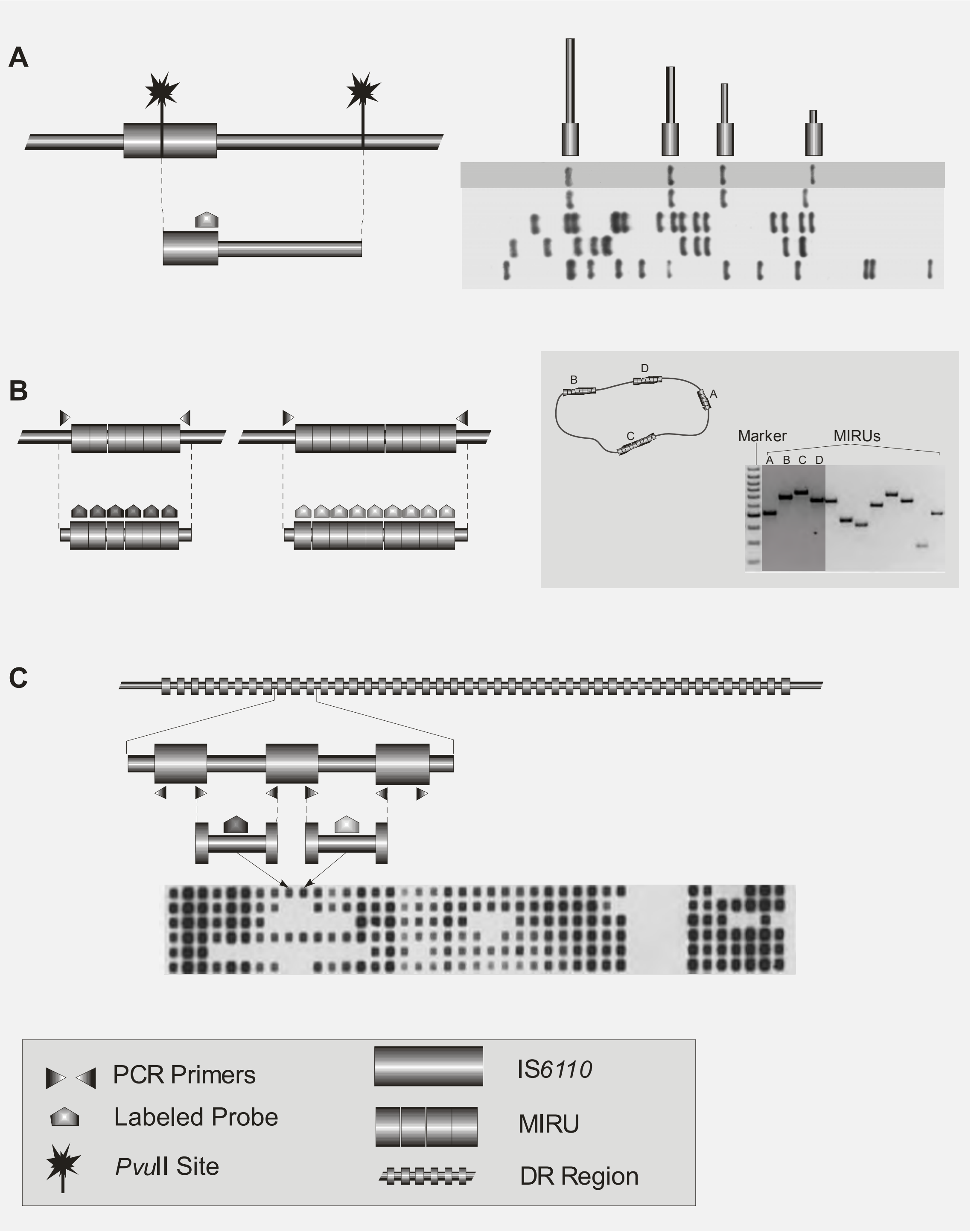

Figure 1. (A) IS*6110* genotyping showing, on the left, a schematic representation of the PvuII restriction of the genome and subsequent hybridization of the probe to the 30 end of the IS*6110* domain. The resulting banding patterns, after gel electrophoresis, of four strains are shown on the right with the corresponding labeled DNA fragments for one*M. tuberculosis* strain (for details see Section 3.4.2.1). (B) MIRU genotyping showing PCR products of two repeat units of differing sizes. The right-hand panel shows a MIRU-sizing gel, inwhich each PCR product is run in a separate lane. The lanes containing the two MIRUs as illustrated in the cartoon are highlighted (for details see Section 3.4.1.3). (C) Spoligotyping showing the DR chromosomal region with an expanded view of two DVRs and their PCR products hybridized to their specific labeled probes. The corresponding spots on the spoligoblot of six strains, indicating the presence of each of the two DVRs, are indicated by the arrows (for details see Section 3.4.1.4).

Epidemiological Interpretation

All molecular epidemiological definitions are based on the understanding that bacterial genomes are in a constant state of flux. According to this assumption, patients whose bacterial populations show identical (or nearly identical) genotypes (termed clusters) can be inferred to have been in contact. Such cases are though to reflect recent transmission, where contact with a source case was followed by infection and relatively rapid progression to disease (*i.e.* within 2 years of infection). The remaining group of cases are those whose bacterial populations have genotypes which do not match those of any other case. Such cases are though to reflect reactivation of a latent infection, which may have been acquired many years prior to the onset of disease. The differences between the genotypes of reactivation cases and those currently circulating may be due either to the absence of genetic change during the latent phase or, alternatively, to rapid change during this period.

By measuring the relative proportion of TB cases falling into clusters, it is possible to estimate the proportion of either recent or ongoing transmission. The former is calculated using the "n–1" formula [(the number of cases in clusters minus the number of clusters)/(the total number of cases in the study)] which assumes that the first case in each cluster represents a reactivation event [30]. Ongoing transmission is calculated using the "n" formula [(the number of cases in clusters)/(the total number of cases in the study)] [31,32]. Both of these calculations are used as public health tools to measure the efficacy of TB treatment programs.

It should be noted that a number of confounding factors could influence estimates of recent transmission. Firstly, a major, common source of error in molecular epidemiological studies arises as a result of under-sampling the infected population [33]. This has the effect of reducing the apparent proportion of strain clustering which gives rise to an underestimate of the degree of recent transmission and is particularly relevant in settings where the average number of cases in clusters is low.

Secondly, most studies fail to address the issue of migration of patients into or out of the study community. This is particularly relevant to mobile, high incidence communities and will tend to result in underestimating the degree of recent transmission. The reason for this is that patients entering the community may be infected with strains not present in the community and will therefore appear to be reactivation cases. Likewise, patients leaving the community will have a similar effect to that of under-sampling.

Thirdly, the accuracy of these epidemiological calculations is also dependent on the duration of the study. Because there may be a delay of up to two years between infection and progression to disease, all molecular epidemiological studies of TB are subject to "edge effects" where it is difficult to predict with any degree of accuracy the events that occurred before the initiation of the study or after the study was terminated. Studies conducted over very short intervals are particularly prone to errors caused by this phenomenon. In order to overcome this problem, it is advisable to allow two-year lead-in and lag phases during which, strains already present in the community and newly detected strains respectively, are disregarded.

Fourthly, not all genotyping methods are appropriate to every situation and the choice of technique will depend on the study population, the nature of the epidemic and the particular questions being addressed. Cost and turnaround time may play a significant role in this decision. A method with lower discriminatory resolution may be appropriate in a low-incidence community where strains are highly diversified and a large proportion of cases are due to reactivation or immigration. In a high-incidence community, on the other hand, where the epidemic is largely driven by transmission of endemic strains, greater discriminatory power will be needed to identify strain sub-populations and transmission patterns, which will often be complicated by evolutionary changes. The choice of method will also be influenced by the possible need for comparison of data with that of other research teams, in which case it would be necessary to choose a standardised technique.

Lastly, the stability of the genotype defines the accuracy of epidemiological inferences. Given that all bacterial genotypes are evolving, strict definitions of clustering based on genotypic identity (inferring transmission) may lead to an under-estimate of transmission as closely related variants may be excluded. To accommodate evolution, the definition of clustering may be relaxed to encompass closely related genotypes.

Evolution of Genetic Markers and Nearest Genetic Distance

The feature of genetic markers such as IS*6110* that makes them useful as indicators of transmission, namely their propensity to evolve at a sufficiently high rate, is, ironically, also a complicating factor in the interpretation of data thus derived. The difficulty arises when a mutational event alters the genetic fingerprint of a transmitted strain. The standard interpretation of this data would regard the altered bacteria as a new strain and thus not a part of the original transmission chain. This has significant consequences for calculations of ongoing transmission. Salamon *et al.* proposed the concept of nearest genetic distance as a means to overcome this problem where strains that are most closely related, and whose differences fall below a certain threshold, can be said to be derived from one another and therefore form part of the same chain of transmission [34]. Application of this modified interpretation of molecular fingerprinting data has been shown to dramatically alter the results of calculations of recent transmission [26]. This has profound implications for understanding the mechanisms driving an epidemic and, consequently, what measures need to be taken to improve control strategies.

Application of Molecular Epidemiology

Molecular epidemiology has revealed and clarified a number of phenomena not accurately understood by classical interpretation.

M. tuberculosis Population Structure

In contrast to classical understanding, molecular based genotyping methods have identified genotypic heterogeneity among clinical isolates. Accordingly, it has become possible to quantify the epidemiological factors contributing to the incidence of TB in different settings. Initial studies done in high incidence settings suggested genotypic homogeneity, thereby questioning the value of molecular approaches to investigating epidemiology in these communities. However, many high incidence settings have subsequently shown an unexpectedly high level of genotypic heterogeneity, probably reflecting a past history of immigration/colonisation.

Genotypic comparisons have revealed that *M. tuberculosis* strains can be grouped according to a similarity index. These groupings have been termed strain families, representing clonal expansion from a common progenitor. Thus, the epidemic in different settings can not be viewed as a single entity, but must rather be seen as a combination of sub-epidemics, each represented by different strains with differing characteristics and in various phases of progression. The success of each strain may also be determined by host-pathogen compatibility.

Transmission vs. reactivation

Comparison of genotypes of clinical isolates using cluster analysis has enabled epidemiologists to accurately differentiate cases arising from transmission or reactivation. Contrasting with previous assumptions, such studies have also revealed the significant role of recent transmission in low-incidence communities [30,32]. This knowledge has directed public health policy to implement strategies aimed at limiting the spread of disease. While it has long been accepted that recent transmission plays a significant role in high-incidence communities, population-based molecular studies have shown that the role of transmission may be as high as 70% [35-37]. By relaxing the definition of a cluster to account for evolution of the marker, the role of transmission may well be considerably higher [26]. This suggests that infectious source cases are not being promptly diagnosed and appropriately treated thereby perpetuating the transmission cycle.

Casual Contact

An unexpected insight resulting from molecular investigations is that transmission of *M. tuberculosis* is not dependent on repeated, close contact, but may often occur as a result of casual contact. Studies in San Francisco, CA and Baltimore, MD were able to identify only 10% and 25% respectively of epidemiological contacts between molecularly related cases [30,38]. These, and other studies, have shown that the majority of transmission arises from complex, casual social interactions which are untraceable by classical methods.

Where transmission occurs

Classical epidemiology shows that close contacts of TB patients have a higher risk of infection and disease than do casual contacts. However, these conclusions are based primarily on household contact studies where the number of casual contacts may be limited. Using genotyping methods it has been possible to investigate this question in a high incidence setting. Contrary to popular belief, it has been shown that most transmission events occurred outside of the household [39]. Similar results were observed for children with TB, raising concern as to where such infections occurred, the role of prophylaxis and the validity of contact tracing [40].

Mapping of Outbreaks

Outbreaks of TB were classically identified by observing an usually high incidence of disease in a community over a defined period. The mechanisms leading to sudden changes in incidence were revealed only with the introduction of genotyping methods which allowed for the tracking of specific strains over space and time. Through these methods it has been possible to document transmission in health care facilities, prisons and in communities, thereby highlighting the inadequacies of infection control measures [41]. The ability of molecular epidemiology to identify new outbreaks of *M. tuberculosis* strains, even in the context of endemic disease, has proved extremely useful. Amongst other things, this has facilitated the identification of possible contributory host- or strain-specific risk factors for transmission. Such micro-epidemic strains may be characterised in terms of genotypically-derived features such as virulence, transmissibility and tropism [42-44].

Risk factors for Transmission

When combined with patient demographic data, molecular epidemiology provides a powerful tool for identifying factors associated with the transmission of *M. tuberculosis*. These risk-factors may be common to all patients or may be specific to communities or geographical areas. Ethnicity, age and HIV infection have all been associated with transmission in some, but not all populations, whereas belonging to immigrant communities has generally not been associated with clustering [30].

Transmission from Smear negative cases

As a rapid, inexpensive procedure, the Ziehl-Neelsen sputum smear has long been regarded as the primary, and in many instances, the conclusive diagnostic test for active TB. Until recently, it has been generally accepted that smear-negative patients are less infectious, if at all [45]. In a study of culture-confirmed TB patients in San Francisco, CA, Behr *et al.* [46] found that at least 21% of IS*6110* strain clusters, for which adequate smear data was available for the presumed index case, were initiated by smear-negative patients. While classical epidemiological reports had suggested the possibility of smear-negative transmission of *M. tuberculosis* [47], this could only be confirmed and quantified using molecular strain typing techniques and this result has subsequently influenced official TB management policy decisions [48].

Recurrent tuberculosis

The mechanism leading to recurrent TB has long been disputed. Until recently, two hypotheses coexisted. The unitary concept of pathogenesis, as expounded by Stead in 1967 [1], stated that TB always began with a primary infection and that recurrent episodes were due to reactivation of dormant bacteria. An alternative hypothesis, proposing exogenous re-infection, first received solid support when it was demonstrated by Raleigh in 1975 by means of phage typing [49]. This finding has subsequently been repeatedly confirmed using molecular strain typing techniques [50]. In such studies re-infection is defined by the isolates from each episode having significantly different genotypes while isolates from reactivated infections have the same genotype. These definitions are only valid in settings where genotypic diversity is high. Not only is re-infection possible, it has been shown to be common in high-incidence communities [51] and a recent study has suggested that patients who have had an episode of disease are at a higher risk of having a subsequent episode [52]. This paradigm shift in epidemiological understanding has important ramifications for classical epidemiological calculations, and for vaccine and drug trials.

Mixed infection

The epidemiological significance of re-infection extends beyond recurrence and may play an important role in secondary TB. Patients simultaneously infected with more than one strain of *M. tuberculosis* have been identified in both high and low incidence settings [53,54]. The high prevalence of multiple- and re-infection suggests that infection with *M. tuberculosis* provides little or no immuno-protection which has implications both for disease control programs and vaccine development. Most significantly, mixed infection has been shown to be a novel mechanism whereby drug resistance may develop in a patient [54]. This phenomenon influences the diagnosis of drug resistance which is a new concern to the TB control programme.

Laboratory Error

Laboratory error resulting in false-positive diagnoses or the misidentification of drug-resistance may have considerable implications both for the welfare of the patient and the resources of the health system. Such errors may occur due to the mislabelling of clinical samples or the cross-contamination of samples during handling and analysis. DNA genotyping methods have both highlighted the extent of this problem and helped to identify specific instances, thereby facilitating the implementation of appropriate corrective measures [55,56]. Detection of possible laboratory error relies on the identification of disparate DNA fingerprints of serial isolates from the same patient or detection of identical strains from different patients whose clinical samples were processed within a predefined period of one-another. The quality, and thus the interpretation of this information will depend on the incidence of disease in the relevant community and the diversity of infecting strains found there. A confounding factor is the existence of patients simultaneously infected with more than one strain, one or both of which may be detected in any sputum sample [54,57].

Drug Resistance

It has long been assumed that drug resistance in TB cases is largely acquired. While poor adherence, and inappropriate treatment regimes are certainly important contributory factors, molecular studies have revealed the significant contribution of primary resistance (*i.e.* transmission of resistant bacteria) [58]. The outbreak of the notorious, multi-drug-resistant 'strain W' in New York, NY challenged the dogma that drug resistance necessarily incurs a fitness cost which lowers transmissibility [59]. From a public health perspective, classical and molecular epidemiological surveys may often yield diametrically opposed conclusions and the resulting TB control strategy adopted will depend largely on the method used.

Insights into the global TB epidemic at the level of the pathogen

Collation of genotype data in large international databases promises to provide insight into the global epidemic. The largest data set is SpolDB representing, to date, in excess of 39000 spoligotypes of clinical isolates cultured from patients in 122 countries [60]. This data has demonstrated the prevalence of different strains families in different geographical settings. It has been suggested that the global distribution of strain families may reflect host-pathogen compatibility [61].

Genotype – phenotype

Different strains of *M. tuberculosis* may have different *in vivo* growth characteristics and may induce a diverse range of host immune responses which may, in turn, lead to variable pathologies and have consequences for transmissibility. Strains of the Beijing genotype, in communities where this strain family is emergent, are associated with younger age groups and are more likely to be drug-resistant than the rest of the strain population [62]. Experiments in mice have shown that the highly transmissible strain, CDC1551, is associated with a vigorous cytokine response and longer survival times [63]. In contrast, the HN878 strain, which belongs to the Beijing grouping, is associated with a reduced cytokine response, high pulmonary bacterial load and shorter host survival times. It appears that this strain selectively induces a predominantly TH2-mediated response which is less protective against *M. tuberculosis* [64,65]. Zhang *et al.* showed that Beijing strain 210 was able to replicate in human macrophages four to eight times faster than unrelated strains [66]. The four major global strain families have demonstrated a diversity of immunopathologies in a mouse intratracheal infection model [67]. As with HN878, the Beijing representative in this study elicited a weak immune response and demonstrated the greatest virulence. Using 19 different strains of *M. tuberculosis* in a murine model, Dormans *et al.* showed a wide range of responses in terms of virulence, lung pathology, bacterial load and delayed hypersensitivity responses [68]. The observation that certain strains or strain families tend to predominate, either globally or locally, suggests that they may well possess characteristics which, at least in particular contexts, enhance their fitness. In the case of locally dominant strains, this fitness advantage may be related to aspects of the TB control program, HIV prevalence or the host population genetic makeup. To the extent that differences between strains of *M. tuberculosis* affect epidemiological parameters such as transmissibility, progression to active disease, and ability to reactivate after latency, they have relevance to molecular epidemiology which may be used to identify and track the spread of specific strains having particular characteristics and inform the treatment and management of patients infected with them.

Vaccines and Clinical Trials

As an intracellular pathogen, host immunity is expected to have provided the most significant selection force during the evolutionary history of *M. tuberculosis*. More recently, introduced evolutionary pressures include the BCG vaccine and anti-TB drugs. The response of *M. tuberculosis* to these new selective parameters can be observed by the numerous outbreaks of drug resistant TB and the implication that mass BCG vaccination in Eastern Asia may have been a selective force in the emergence of the Beijing family phenotype [69]. The recent emergence of HIV has resulted in a new selection parameter. The evolutionary consequences of *M. tuberculosis* infection in the context of HIV and the antiretroviral therapy used to treat it remains to be determined. Molecular epidemiological studies running concurrently with ant-TB vaccine or drug trials will provide essential insights into how the implementation of these novel therapies influence the strain population structure and in particular will aid in the identification of "vaccine escape" or drug resistant mutants.

Summary

Since molecular technology was first applied to the epidemiology of TB it has had a significant impact on the field. As we have shown, it has repeatedly challenged assumptions, altered perceptions and made possible the answering of numerous important questions as well as raising many more. A range of molecular tools are now available which continue to be added to and refined. Our clearer understanding of the epidemic, in terms of the contribution of transmission, particularly in the case of drug-resistant TB, has placed us in a position to devise better-informed control programs. New diagnostic technologies as well as better treatments and an effective vaccine are three essential elements required for combating the epidemic. However, the demonstration of the prevalence of re-infection and multiple infections is a cause for concern both for TB control programs and the development of anti-TB vaccines and novel drugs. The inference drawn from the findings of molecular studies is that *M. tuberculosis* is highly successful pathogen, skilled at evading the host defence systems. The battle to understand, control, and possibly eradicate it will be neither quick nor easy.

Reference List

1. **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**:729-745.

2. **Hussain, S.F., Watura, R., Cashman, B., Campbell, I.A., and Evans, M.R.** Tuberculosis contact tracing: are the British Thoracic Society guidelines still appropriate? *Thorax* 1992, **47**:984-985.

3. **Macdonald, J.B.** *Mycobacterium tuberculosis* resistance to rifampicin and ethambutol: a clinical survey. *Thorax* 1977, **32**:1-4.

4. **Bates, J.H. and Fitzhugh, J.K.** Subdivision of the species *M. tuberculosis* by mycobacteriophage typing. *Am.Rev.Respir.Dis.* 1967, **96**:7-10.

5. **van Soolingen, D. and Hermans, P.W.** Epidemiology of tuberculosis by DNA fingerprinting. *Eur.Respir.J.Suppl* 1995, **20**:649s-656s.

6. **Supply, P., Mazars, E., Lesjean, S., Vincent, V., Gicquel, B., and Locht, C.** Variable human minisatellite-like regions in the *Mycobacterium tuberculosis* genome. *Mol.Microbiol.* 2000, **36**:762-771.

7. **Kamerbeek, J., Schouls, L., Kolk, A., van Agterveld, M., van Soolingen, D., Kuijper, S., Bunschoten, A., Molhuizen, H., Shaw, R., Goyal, M., and van Embden, J.** Simultaneous detection and strain differentiation of *Mycobacterium tuberculosis* for diagnosis and epidemiology. *J.Clin.Microbiol.* 1997, 35:907-914.

8. **Ross, B.C., Raios, K., Jackson, K., and Dwyer, B.** Molecular cloning of a highly repeated DNA element from *Mycobacterium tuberculosis* and its use as an epidemiological tool. *J.Clin.Microbiol.* 1992, **30**:942-946.

9. **Wiid, I.J., Werely, C., Beyers, N., Donald, P., and van Helden, P.D.** Oligonucleotide (GTG)5 as a marker for *Mycobacterium tuberculosis* strain identification. *J.Clin.Microbiol.* 1994, **32**:1318-1321.

10. **Hermans, P.W., van Soolingen, D., and van Embden, J.D.** Characterization of a major polymorphic tandem repeat in *Mycobacterium tuberculosis* and its potential use in the epidemiology of Mycobacterium kansasii and Mycobacterium gordonae.*J.Bacteriol.* 1992, **174**:4157-4165.

11. **Goyal, M., Young, D., Zhang, Y., Jenkins, P.A., and Shaw, R.J.** PCR amplification of variable sequence upstream of katG gene to subdivide strains of *Mycobacterium tuberculosis* complex.*J.Clin.Microbiol.* 1994, **32**:3070-3071.

12. **Frothingham, R. and Meeker-O'Connell, W.A.** Genetic diversity in the *Mycobacterium tuberculosis* complex based on variable numbers of tandem DNA repeats. *Microbiology* 1998, 144 (**Pt** 5):1189-1196.

13. **Hermans, P.W., van Soolingen, D., Bik, E.M., de Haas, P.E., Dale, J.W., and van Embden, J.D.** 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, 59:2695-2705.

14. **de Wit, D., Steyn, L., Shoemaker, S., and Sogin, M.** Direct detection of *Mycobacterium tuberculosis* in clinical specimens by DNA amplification.*J.Clin.Microbiol.* 1990, 28:2437-2441.

15. **Brennan, M.J., Gey van Pittius, N.C., and Espitia, C.** The PE and PPE multigene families of the Mycobacteria. *In* **Cole, S.T., Eisenach, K.D., McMurray, D.N., and Jacobs, W.R., Jr.** (eds.), Part X: Host Immune Responses and Antigenic Variation. 2004, ASM Press, Washington, USA.

16. **Ross, B.C., Raios, K., Jackson, K., Sievers, A., and Dwyer, B.** Differentiation of *Mycobacterium tuberculosis* strains by use of a nonradioactive Southern blot hybridization method. *J.Infect.Dis.* 1991, 163:904-907.

17. **Cole, S.T., Brosch, R., Parkhill, J., Garnier, T., Churcher, C., Harris, D., Gordon, S.V., Eiglmeier, K., Gas, S., Barry, C.E., III, Tekaia, F., Badcock, K., Basham, D., Brown, D., Chillingworth, T., Connor, R., Davies, R., Devlin, K., Feltwell, T., Gentles, S., Hamlin, N., Holroyd, S., Hornsby, T., Jagels, K., Krogh, A., McLean, J., Moule, S., Murphy, L., Oliver, K., Osborne, J., Quail, M.A., Rajandream, M.A., Rogers, J., Rutter, S., Seeger, K., Skelton, J., Squares, R., Squares, S., Sulston, J.E., Taylor, K., Whitehead, S., and Barrell, B.G.** Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* 1998, 393:537-544.

18. **Filliol, I., Motiwala, A.S., Cavatore, M., Qi, W., Hazbon, M.H., Bobadilla, d., V, Fyfe, J., Garcia-Garcia, L., Rastogi, N., Sola, C., Zozio, T., Guerrero, M.I., Leon, C.I., Crabtree, J., Angiuoli, S., Eisenach, K.D., Durmaz, R., Joloba, M.L., Rendon, A., Sifuentes-Osornio, J., Ponce, d.L., Cave, M.D., Fleischmann, R., Whittam, T.S., and Alland, D.** 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, 188:759-772.

19. **Sola, C., Filliol, I., Legrand, E., Lesjean, S., Locht, C., Supply, P., and Rastogi, N.** Genotyping of the *Mycobacterium tuberculosis* complex using MIRUs: association with VNTR and

spoligotyping for molecular epidemiology and evolutionary genetics. *Infect.Genet.Evol.* 2003, 3:125-133.

20. **Supply, P., Lesjean, S., Savine, E., Kremer, K., van Soolingen, D., and Locht, C.** Automated high-throughput genotyping for study of global epidemiology of *Mycobacterium tuberculosis* based on mycobacterial interspersed repetitive units.*J.Clin.Microbiol.* 2001, **39**:3563-3571.

21. **van Embden, J.D., van Gorkom, T., Kremer, K., Jansen, R., Der Zeijst, B.A., and Schouls, L.M.** Genetic variation and evolutionary origin of the direct repeat locus of *Mycobacterium tuberculosis* complex bacteria.*J.Bacteriol.* 2000, **182**:2393-2401.

22. **Warren, R.M., Streicher, E.M., Sampson, S.L., van der Spuy, G.D., Richardson, M., Nguyen, D., Behr, M.A., Victor, T.C., and van Helden, P.D.** Microevolution of the direct repeat region of *Mycobacterium tuberculosis*: implications for interpretation of spoligotyping data. *J.Clin.Microbiol.* 2002, 40:4457-4465.

23. **Dale, J.W., Brittain, D., Cataldi, A.A., Cousins, D., Crawford, J.T., Driscoll, J., Heersma, H., Lillebaek, T., Quitugua, T., Rastogi, N., Skuce, R.A., Sola, C., van Soolingen, D., and Vincent, V.** Spacer oligonucleotide typing of bacteria of the *Mycobacterium tuberculosis* complex: recommendations for standardised nomenclature. *Int.J.Tuberc.Lung Dis.* 2001, 5:216-219.

24. **van Embden, J.D., Cave, M.D., Crawford, J.T., Dale, J.W., Eisenach, K.D., Gicquel, B., Hermans, P., Martin, C., McAdam, R., and Shinnick, T.M.** Strain identification of *Mycobacterium tuberculosis* by DNA fingerprinting: recommendations for a standardized methodology.*J.Clin.Microbiol.* 1993, 31:406-409.

25. **de Boer, A.S., Borgdorff, M.W., de Haas, P.E., Nagelkerke, N.J., van Embden, J.D., and van Soolingen, D.** Analysis of rate of change of IS*6110* RFLP patterns of *Mycobacterium tuberculosis* based on serial patient isolates.*J.Infect.Dis.* 1999, **180**:1238-1244.

26. **van der Spuy, G.D., Warren, R.M., Richardson, M., Beyers, N., Behr, M.A., and van Helden, P.D.** Use of genetic distance as a measure of ongoing transmission of *Mycobacterium tuberculosis.J.Clin.Microbiol.* 2003, 41:5640-5644.

27. **Haas, W.H., Butler, W.R., Woodley, C.L., and Crawford, J.T.** Mixed-linker polymerase chain reaction: a new method for rapid fingerprinting of isolates of the *Mycobacterium tuberculosis* complex.*J.Clin.Microbiol.* 1993, 31:1293-1298.

28. **Vos, P., Hogers, R., Bleeker, M., Reijans, M., van de, L.T., Hornes, M., Frijters, A., Pot, J., Peleman, J., and Kuiper, M.** AFLP: a new technique for DNA fingerprinting. *Nucleic Acids Res.* 1995, 23:4407-4414.

29. **Goulding, J.N., Stanley, J., Saunders, N., and Arnold, C.** Genome-sequence-based fluorescent amplified-fragment length polymorphism analysis of *Mycobacterium tuberculosis*. *J.Clin.Microbiol.* 2000, **38**:1121-1126.

30. **Small, P.M., Hopewell, P.C., Singh, S.P., Paz, A., Parsonnet, J., Ruston, D.C., Schecter, G.F., Daley, C.L., and Schoolnik, G.K.** The epidemiology of tuberculosis in San Francisco. A population-based study using conventional and molecular methods. *N.Engl.J.Med.* 1994, 330:1703-1709.

31. **Murray, M. and Alland, D.** Methodological problems in the molecular epidemiology of tuberculosis. *Am.J.Epidemiol.* 2002, **155**:565-571.

32. **Alland, D., Kalkut, G.E., Moss, A.R., McAdam, R.A., Hahn, J.A., Bosworth, W., Drucker, E., and Bloom, B.R.** Transmission of tuberculosis in New York City. An analysis by DNA fingerprinting and conventional epidemiologic methods. *N.Engl.J.Med.* 1994, **330**:1710-1716.

33. **Glynn, J.R., Vynnycky, E., and Fine, P.E.** Influence of sampling on estimates of clustering and recent transmission of *Mycobacterium tuberculosis* derived from DNA fingerprinting techniques. *Am.J.Epidemiol.* 1999, **149**:366-371.

34. **Salamon, H., Behr, M.A., Rhee, J.T., and Small, P.M.** Genetic distances for the study of infectious disease epidemiology. *Am.J.Epidemiol.* 2000, **151**:324-334.

35. **Glynn, J.R., Crampin, A.C., Yates, M.D., Traore, H., Mwaungulu, F.D., Ngwira, B.M., Ndlovu, R., Drobniewski, F., and Fine, P.E.** 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**:480-487.

36. **Godfrey-Faussett, P., Sonnenberg, P., Shearer, S.C., Bruce, M.C., Mee, C., Morris, L., and Murray, J.** Tuberculosis control and molecular epidemiology in a South African gold-mining community. *Lancet* 2000, **356**:1066-1071.

37. **Verver, S., Warren, R.M., Munch, Z., Vynnycky, E., van Helden, P.D., Richardson, M., van der Spuy, G.D., Enarson, D.A., Borgdorff, M.W., Behr, M.A., and Beyers, N.** Transmission of tuberculosis in a high incidence urban community in South Africa. *Int.J.Epidemiol.* 2004, 33:351-357.

38. **Bishai, W.R., Graham, N.M., Harrington, S., Pope, D.S., Hooper, N., Astemborski, J., Sheely, L., Vlahov, D., Glass, G.E., and Chaisson, R.E.** Molecular and geographic patterns of tuberculosis transmission after 15 years of directly observed therapy. *JAMA* 1998, 280:1679-1684.

39. **Verver, S., Warren, R.M., Munch, Z., Richardson, M., van der Spuy, G.D., Borgdorff, M.W., Behr, M.A., Beyers, N., and van Helden, P.D.** Proportion of tuberculosis transmission that takes place in households in a high-incidence area. *Lancet* 2004, 363:212-214.

40. **Schaaf, H.S., Michaelis, I.A., Richardson, M., Booysen, C.N., Gie, R.P., Warren, R., van Helden, P.D., and Beyers, N.** Adult-to-child transmission of tuberculosis: household or community contact? *Int.J.Tuberc.Lung Dis.* 2003, 7:426-431.

41. **Hutton, M.D., Cauthen, G.M., and Bloch, A.B.** Results of a 29-state survey of tuberculosis in nursing homes and correctional facilities. *Public Health Rep.* 1993, 108:305-314.

42. **Garcia, d.V., Lorenzo, G., Cardona, P.J., Rodriguez, N.A., Gordillo, S., Serrano, M.J., and Bouza, E.** Association between the infectivity of *Mycobacterium tuberculosis* strains and their efficiency for extrarespiratory infection.*J.Infect.Dis.* 2005, 192:2059-2065.

43. **Reed, M.B., Domenech, P., Manca, C., Su, H., Barczak, A.K., Kreiswirth, B.N., Kaplan, G., and Barry, C.E., III**. A glycolipid of hypervirulent tuberculosis strains that inhibits the innate immune response. *Nature* 2004, 431:84-87.

44. **Valway, S.E., Sanchez, M.P., Shinnick, T.F., Orme, I., Agerton, T., Hoy, D., Jones, J.S., Westmoreland, H., and Onorato, I.M.** An outbreak involving extensive transmission of a virulent strain of *Mycobacterium tuberculosis*. *N.Engl.J.Med.* 1998, 338:633-639.

45. **Shaw, J.B. and Wynn-Williams, N.** Infectivity of pulmonary tuberculosis in relation to sputum status. *Am.Rev.Tuberc.* 1954, 69:724-732.

46. **Behr, M.A., Warren, S.A., Salamon, H., Hopewell, P.C., Ponce, d.L., Daley, C.L., and Small, P.M.** Transmission of *Mycobacterium tuberculosis* from patients smear-negative for acid-fast bacilli. *Lancet* 1999, 353:444-449.

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

48. Prevention and control of tuberculosis in correctional and detention facilities: recommendations from CDC. Endorsed by the Advisory Council for the Elimination of Tuberculosis, the National Commission on Correctional Health Care, and the American Correctional Association. *MMWR Recomm.Rep.* 2006, 55:1-44.

49. **Raleigh, J.W., Wichelhausen, R.H., Rado, T.A., and Bates, J.H.** Evidence for infection by two distinct strains of *Mycobacterium tuberculosis* in pulmonary tuberculosis: report of 9 cases. *Am.Rev.Respir.Dis.* 1975, 112:497-503.

50. **Dwyer, B., Jackson, K., Raios, K., Sievers, A., Wilshire, E., and Ross, B.** DNA restriction fragment analysis to define an extended cluster of tuberculosis in homeless men and their associates.*J.Infect.Dis.* 1993, 167:490-494.

51. **van Rie, A., Warren, R., Richardson, M., Victor, T.C., Gie, R.P., Enarson, D.A., Beyers, N., and van Helden, P.D.** Exogenous reinfection as a cause of recurrent tuberculosis after curative treatment. *N.Engl.J.Med.* 1999, 341:1174-1179.

52. **Verver, S., Warren, R.M., Beyers, N., Richardson, M., van der Spuy, G.D., Borgdorff, M.W., Enarson, D.A., Behr, M.A., and van Helden, P.D.** Rate of reinfection tuberculosis after successful treatment is higher than rate of new tuberculosis. *Am.J.Respir.Crit Care Med.* 2005, 171:1430-1435.

53. **Braden, C.R., Morlock, G.P., Woodley, C.L., Johnson, K.R., Colombel, A.C., Cave, M.D., Yang, Z., Valway, S.E., Onorato, I.M., and Crawford, J.T.** Simultaneous infection with multiple strains of *Mycobacterium tuberculosis*. *Clin.Infect.Dis.* 2001, 33:e42-e47.

54. **Warren, R.M., Victor, T.C., Streicher, E.M., Richardson, M., Beyers, N., Gey van Pittius, N.C., and van Helden, P.D.** Patients with active tuberculosis often have different strains in the same sputum specimen. *Am.J.Respir.Crit Care Med.* 2004, 169:610-614.

55. **Burman, W.J. and Reves, R.R.** Review of false-positive cultures for *Mycobacterium tuberculosis* and recommendations for avoiding unnecessary treatment. *Clin.Infect.Dis.* 2000, 31:1390-1395.

56. **Small, P.M., McClenny, N.B., Singh, S.P., Schoolnik, G.K., Tompkins, L.S., and Mickelsen, P.A.** Molecular strain typing of *Mycobacterium tuberculosis* to confirm cross-contamination in the mycobacteriology laboratory and modification of procedures to minimize occurrence of false-positive cultures.*J.Clin.Microbiol.* 1993, 31:1677-1682.

57. **Richardson, M., Carroll, N.M., Engelke, E., van der Spuy, G.D., Salker, F., Munch, Z., Gie, R.P., Warren, R.M., Beyers, N., and van Helden, P.D.** Multiple *Mycobacterium tuberculosis* strains in early cultures from patients in a high-incidence community setting. *J.Clin.Microbiol.* 2002, 40:2750-2754.

58. **van Rie, A., Warren, R.M., Beyers, N., Gie, R.P., Classen, C.N., Richardson, M., Sampson, S.L., Victor, T.C., and van Helden, P.D.** Transmission of a multidrug-resistant *Mycobacterium tuberculosis* strain resembling "strain W" among noninstitutionalized, human immunodeficiency virus-seronegative patients.*J.Infect.Dis.* 1999, 180:1608-1615.

59. **Bifani, P.J., Plikaytis, B.B., Kapur, V., Stockbauer, K., Pan, X., Lutfey, M.L., Moghazeh, S.L., Eisner, W., Daniel, T.M., Kaplan, M.H., Crawford, J.T., Musser, J.M., and**

Kreiswirth, B.N. Origin and interstate spread of a New York City multidrug-resistant *Mycobacterium tuberculosis* clone family.*JAMA* 1996, 275:452-457.

60. **Brudey, K., Driscoll, J.R., Rigouts, L., Prodinger, W.M., Gori, A., Al Hajoj, S.A., Allix, C., Aristimuno, L., Arora, J., Baumanis, V., Binder, L., Cafrune, P., Cataldi, A., Cheong, S., Diel, R., Ellermeier, C., Evans, J.T., Fauville-Dufaux, M., Ferdinand, S., Garcia, d., V, Garzelli, C., Gazzola, L., Gomes, H.M., Guttierez, M.C., Hawkey, P.M., van Helden, P.D., Kadival, G.V., Kreiswirth, B.N., Kremer, K., Kubin, M., Kulkarni, S.P., Liens, B., Lillebaek, T., Ho, M.L., Martin, C., Martin, C., Mokrousov, I., Narvskaia, O., Ngeow, Y.F., Naumann, L., Niemann, S., Parwati, I., Rahim, Z., Rasolofo-Razanamparany, V., Rasolonavalona, T., Rossetti, M.L., Rusch-Gerdes, S., Sajduda, A., Samper, S., Shemyakin, I.G., Singh, U.B., Somoskovi, A., Skuce, R.A., van Soolingen, D., Streicher, E.M., Suffys, P.N., Tortoli, E., Tracevska, T., Vincent, V., Victor, T.C., Warren, R.M., Yap, S.F., Zaman, K., Portaels, F., Rastogi, N., and Sola, C.** *Mycobacterium tuberculosis* complex genetic diversity: mining the fourth international spoligotyping database (SpolDB4) for classification, population genetics and epidemiology. *BMC.Microbiol.* 2006, 6:23.

61. **Gagneux, S., DeRiemer, K., Van, T., Kato-Maeda, M., de Jong, B.C., Narayanan, S., Nicol, M., Niemann, S., Kremer, K., Gutierrez, M.C., Hilty, M., Hopewell, P.C., and Small, P.M.** Variable host-pathogen compatibility in *Mycobacterium tuberculosis*. *Proc.Natl.Acad.Sci.U.S.A* 2006, 103:2869-2873.

62. **Glynn, J.R., Kremer, K., Borgdorff, M.W., Rodriguez, M.P., and van Sooligen, D.** Beijing/W genotype *Mycobacterium tuberculosis* and drug resistance. *Emerg.Infect.Dis.* 2006, 12:736-743.

63. **Manca, C., Tsenova, L., Barry, C.E., III, Bergtold, A., Freeman, S., Haslett, P.A., Musser, J.M., Freedman, V.H., and Kaplan, G.** *Mycobacterium tuberculosis* CDC1551 induces a more vigorous host response in vivo and in vitro, but is not more virulent than other clinical isolates. *J.Immunol.* 1999, 162:6740-6746.

64. **Manca, C., Tsenova, L., Bergtold, A., Freeman, S., Tovey, M., Musser, J.M., Barry, C.E., III, Freedman, V.H., and Kaplan, G.** Virulence of a *Mycobacterium tuberculosis* clinical isolate in mice is determined by failure to induce Th1 type immunity and is associated with induction of IFN-alpha /beta. *Proc.Natl.Acad.Sci.U.S.A* 2001, 98:5752-5757.

65. **Manca, C., Reed, M.B., Freeman, S., Mathema, B., Kreiswirth, B., Barry, C.E., III, and Kaplan, G.** Differential monocyte activation underlies strain-specific *Mycobacterium tuberculosis* pathogenesis. *Infect.Immun.* 2004, 72:5511-5514.

66. **Zhang, M., Gong, J., Yang, Z., Samten, B., Cave, M.D., and Barnes, P.F.** Enhanced capacity of a widespread strain of *Mycobacterium tuberculosis* to grow in human macrophages. *J.Infect.Dis.* 1999, 179:1213-1217.

67. **Lopez, B., Aguilar, D., Orozco, H., Burger, M., Espitia, C., Ritacco, V., Barrera, L., Kremer, K., Hernandez-Pando, R., Huygen, K., and van Soolingen, D.** A marked difference in pathogenesis and immune response induced by different *Mycobacterium tuberculosis* genotypes. *Clin.Exp.Immunol.* 2003, 133:30-37.

68. **Dormans, J., Burger, M., Aguilar, D., Hernandez-Pando, R., Kremer, K., Roholl, P., Arend, S.M., and van Soolingen, D.** 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, 137:460-468.

69. **Abebe, F. and Bjune, G.** The emergence of Beijing family genotypes of *Mycobacterium tuberculosis* and low-level protection by bacille Calmette-Guerin (BCG) vaccines: is there a link? *Clin.Exp.Immunol.* 2006, 145:389-397.

2

Calculation of the stability of the IS*6110* banding pattern in patients with persistent *M. tuberculosis* disease

R. M. Warren, G. D. van der Spuy, M. Richardson, N. Beyers, M. W. Borgdorff,
M.A. Behr, P. D. van Helden

J. Clin. Microbiol. (2002) 40(5):1705-1708

Abstract

The interpretation of molecular epidemiologic data of *M. tuberculosis* infection is dependent on the understanding of the stability and evolutionary characteristics of the DNA fingerprinting marker used to classify clinical isolates. This study investigated the stability of the IS*6110* banding pattern in serial tuberculosis isolates collected from patients resident in a high incidence area. Evolutionary changes were observed in 4 % of the strains and a half-life of 8.74 years was calculated, assuming a constant rate of change over time. This rate may be composed of a fast rate of change seen during the early disease phase (t½ = 0.57 years) and a slow rate of change seen in late disease phase (t½ = 10.69 years). The early rate probably reflects change occurring during active growth prior to therapy, while the slow, late rate may reflect change occurring during or after treatment. We demonstrate that the calculation of these rates will be strongly influenced by the time interval between onset of disease and sputum sampling. These calculations are further complicated by partial replacement of the original strain population resulting in the sporadic appearance of clonal variants in sputum specimens. Therefore, the true extent of genetic diversity may be underestimated within each host, thereby influencing molecular epidemiological data used to establish transmission chains.

Introduction

Molecular epidemiology needs to differentiate clinical isolates of pathogenic organisms as the same or different. This differentiation relies on the evolutionary process generating genotypic diversity among isolates. In general, the observation of a number of isolates sharing identical patterns is used to infer infection from a common source. In contrast, isolates with unrelated genotypes are thought to represent infection from different organisms, and therefore from independent sources.

The molecular epidemiology of *M. tuberculosis* is most frequently based on the RFLP (restriction fragment length polymorphism) banding patterns generated by Southern hybridization with the probe IS*6110* (10,13). In community settings isolates with identical IS*6110* genotypes are grouped together into clusters and are thought to represent recent epidemiological events (5), while isolates with IS*6110* banding patterns unrelated to any others in the database (isolates with unique banding patterns) are thought to reflect reactivation of latent infection (1,9). This methodology has been used in numerous settings to quantify the relative proportion of recent epidemiological events and thereby estimate the extent of recently transmitted disease (1,9,14).

In order to estimate the stability of the IS*6110* RFLP patterns, studies have examined serial isolates collected from patients with persistent disease (3,6,15) and have demonstrated that the IS*6110* banding pattern may change over time in a subset of these patients. When survival analysis was applied to the RFLP data collected from patients with persistent disease in The Netherlands, De Boer *et al.* (1999), calculated that the half-life of the IS*6110* banding pattern was of the order of 3.2 years. A similar rate was calculated for isolates originating from San Francisco, although this was restricted to isolates obtained at least 90 days apart, to exclude the possibility of variable patterns representing co-infection rather than change over time (3). In both datasets, the estimated rate of IS*6110* pattern change was felt to be of the right magnitude to allow the use of IS*6110* as a marker for molecular epidemiological studies (3).

However, the calculated rate of change may not be uniform in two groups of patients: those with multiple isolates at the time of diagnosis and those in whom a new pattern develops during or after treatment. In this study we have investigated the stability of IS*6110* banding pattern in a group of patients for whom more than one isolate of *M. tuberculosis* was collected during the course of disease or during re-treatment. The results are discussed in the context of the applicability of calculating a rate of change in these patient groups and how such change may influence molecular epidemiological calculations.

Methods

Study setting

Between January 1992 and December 1998 *M. tuberculosis* isolates were collected from patients attending the healthcare clinics within adjacent suburbs in Cape Town, South Africa (2). This community is experiencing an extremely high incidence of tuberculosis, with approximately 250 new bacteriologically confirmed adult cases per 100 000 population, per year (11). Before 1995 the National TB Program focused on sputum-cultures for the diagnosis of tuberculosis with follow-up sputum samples requested by the attending physician on clinical grounds. Both Ziehl Neelsen positive and negative samples collected during this period were included in the study. After 1995 the South African National TB Program has been conducted according to the WHO DOTS strategy with sputum taken for Ziehl Neelsen smear at presentation for diagnosis, at 2 months for sputum conversion and again at six months after initiation of therapy to monitor response to therapy (additional sputum specimens may be requested by the attending physician on clinical grounds). As part of the research project, all Ziehl Neelsen positive and negative samples are sent for culture. Clinical information for each patient was recorded at the time of diagnosis and this data was stored in a Microsoft Access database. The restricted sampling of sputum specimens during the course of disease in the different patients may influence rate calculations.

RFLP generation and Gelcompar Analysis.

The entire culture representing all the possible clonal variants present in each isolate was used for the DNA isolation and this DNA was genotypically classified according to the internationally standardized DNA fingerprinting protocol, using the IS-3' probe (10,13). The Southern blot autoradiographs were normalized and the IS-3' bands were assigned using GelCompar 4.1 software. The assignments were visually checked by two independent persons, and only bands with an intensity of > 20 % of the other bands were scored as representing IS*6110* mediated evolutionary events (4). Replicative transposition was identified when the evolved strain showed an additional IS*6110* hybridizing band, while deletion of an IS*6110* was identified when an IS*6110* hybridizing band was absent from the evolved strain. Variation in the electrophoretic mobility of an IS*6110* hybridizing band was classified as a band shift representing a mutational event in the chromosomal domain flanking the IS*6110* element. Banding pattern changes suspected of being the result of partial digestion of methylated restriction sites (12) were excluded.

Study population

Serial isolates from patients with persistent disease

Patients from whom DNA fingerprints from more than one isolate were available, were selected from the complete data-set. This excluded isolates where laboratory error (11), contamination or loss of viability had been noted. Patients who had been assigned as being re-infected with a genotypically unrelated strain were treated as two separate cases of disease if serial isolates were available for both infections. On average 3 isolates were collected from each case, and this was independent of the tuberculosis control program at the time of sample collection. Serial isolates collected from patients who presented with a genotypically identical or related strain after therapy were treated as a single case. Selected patients were divided into two categories; 1) patients infected with a strain which remained genotypically identical during the course of sampling were termed invariant, and 2) patients infected with a strain which changed during the course of sampling were termed variant. Change in the IS*6110* banding pattern was identified by the appearance, disappearance or electrophoretic shift of

IS-3' hybridising elements. To avoid a possible bias in the invariant patient group, variant strains with more than four changes in the IS*6110* banding pattern were considered to reflect reinfection rather than evolution and were excluded from all calculations. In addition, if any variant strain (independent of the number of changes in the IS*6110* banding pattern) was found to be present in the community prior to appearance in the patient, the patient was excluded from the study. In such a case, the patient may have been reinfected from a community contact. This stringent inclusion criteria excludes the possibility that the IS*6110* banding pattern may revert to a previous evolution state or that identical banding patterns could have evolved convergently. It is envisaged that the above exclusion criteria will lead to an overestimate of the stability of the IS*6110* banding pattern.

The sampling interval (days) for patients infected with invariant strains was calculated from the date when the first isolate was collected to the date when the last isolate was collected in each patient. The sampling interval (days) for patients infected with variant strains was calculated from the date when the first isolate was collected to the date when the first variant isolate was identified (3). The overall rate of change was calculated using survival analysis according to the method described by de Boer *et al.* (1999). Similarly, the early rate of change was calculated using cases with inter-isolate intervals of ≤ 90 days, while the late rate of change was calculated using cases with inter-isolate intervals of > 90 days.

Statistical analysis

Fisher's exact test was used to test whether changes in the IS*6110* banding pattern were associated with the number of IS*6110* insertions present in the evolving strain. This test was also used to identify associations between previous therapy and the appearance of a variant strain. Chi-squared analysis was used to identify differences between the study sample set and the sample set from The Netherlands (3).

Results

Study population

M. tuberculosis isolates were cultured from 954 patients (representing a recovery of approximately 70 % of all culture positive patients) attending the healthcare clinics in adjacent suburbs of Cape Town, South Africa (2,14). Fifty of these patients (5.2 %) were excluded from the study as their cultures were either contaminated or lost viability. Of the remaining patients, 901 (94.4 %) had an *M. tuberculosis* isolate genotypically classified by IS-3' DNA fingerprinting (10,13).

Patients with serial isolates of *M. tuberculosis*

Three-hundred and forty-six patients had ≥ 2 *M. tuberculosis* isolates, representing serial isolates from 351 active cases of tuberculosis. Analysis of the serial isolates by IS-3' DNA fingerprinting identified 335 cases (95.4 %) where the serial *M. tuberculosis* isolates remained genotypically constant over the sampling period (range 0 to 2203 days), termed invariant strains (Table 1). Serial isolates from 16 cases showed an IS*6110* banding pattern change over the sampling period, termed variant strains

(Table 1). Two of the patients infected with a variant strain were excluded, as the variant strain was found to have been present in the community prior to appearing in the patient, thereby suggesting possible exogenous reinfection from a community source case. The remaining 14 variant cases were sampled over a period of 0 to 1141 days (Table 1). Comparing patients with invariant strains to those with variant strains did not reveal any differences in the demographic variables. However, a previous history of disease was weakly associated with the appearance of variant strains (Fisher's exact test, P = 0.045) (Table 1), which may suggest that the evolution of clonal variants occurred either during the latent disease phase after the previous infection or during the reactivation process.

Table 1. Demographics of patients with serial isolates, where the infecting *M. tuberculosis* strain was either genotypically variant or invariant over time.

Demographics	Variant	Invariant
Total	14	335
Male	7 (50%)	187 (55.8%)
Mean Age (years)	35.6	34.1
Previous episode of TB	9 (64%)*	120 (35.8%)
Pulmonary TB	14 (93.3%)	326 (97.3%)
Sampling range (days)	0 - 1141	0 - 2203
Low IS*6110* copy number strains	0**	61
High IS*6110* copy number strains	14**	274

*(Fisher's exact, P = 0.045)

**(Fisher's exact, P = 0.142)

The proportion of variant strains identified n = 14 (4 %) is similar to that reported for isolates collected in The Netherlands (4.6 %) (3), although this is lower than reported in Germany (9 %) (6) and in the restricted analysis performed in San Francisco (24.5 %) (15). Change in the IS*6110* banding pattern was specifically associated with strains with > 5 hybridizing elements, however, because of the low number of observations, this was not statistically significant at the 95% confidence interval (Fisher's exact test, P = 0.142) (Table 1).

Analysis of the RFLP data of the variant isolates showed that in 10 of 14 cases (71.4 %) the IS*6110* banding patterns changed by replicative transposition, while in 2 of 14 cases (14.3 %) a hybridizing band was deleted. In 2 of 14 cases (14.3 %), an electrophoretic shift was observed suggestive of a chromosomal mutation. In 4 patients the variant strain disappeared from subsequent serial isolates demonstrating incomplete clonal replacement. This suggests that two clonal variant populations may be present in the patient and that a single sputum specimen may not reflect the true genetic diversity of the bacterial population in the host.

Survival analysis (3) of this data estimated the overall half-life of the IS*6110* banding pattern in our serial isolates to be 8.74 years (95 % CI, 7.51 – 10.45) (Figure 1). However, the survival plot suggests the presence of two distinct rates of change (Figure 1). In 8 of the 14 patients (57 %) both the initial and variant isolate were sampled within the first 20 days (median of 10 days), suggesting that the banding pattern change occurred early in the disease process or during the latent phase of a previous infection, or during reactivation. This was followed by only partial clonal replacement at the time of diagnosis or shortly thereafter. In the remaining 6 patients (43 %), variant isolates were only sampled after > 250 days, suggesting that these isolates evolved late in the disease phase.

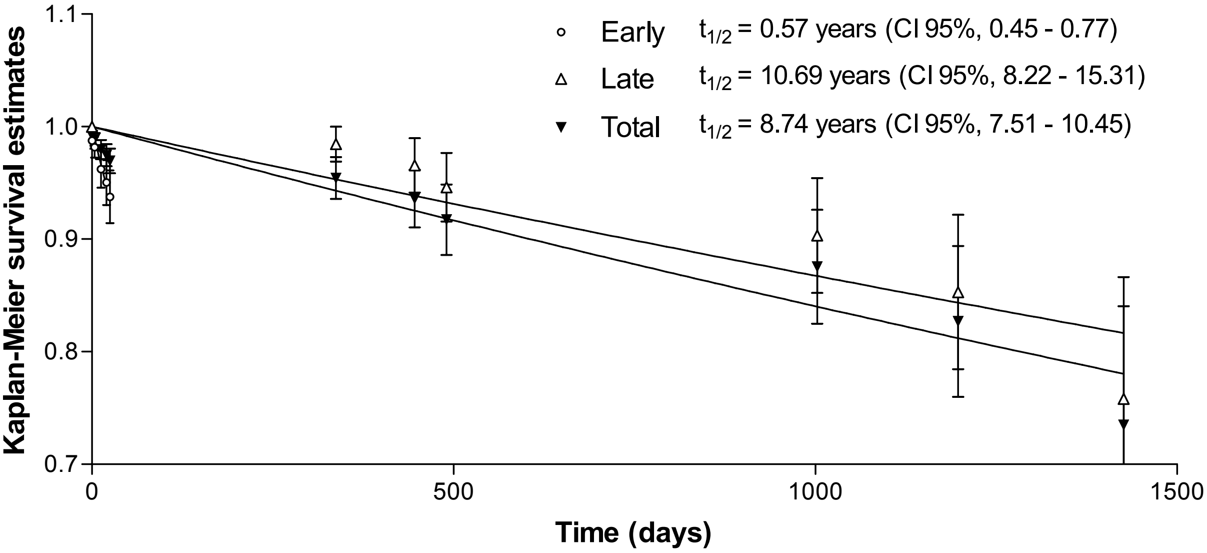

Figure 1. Survival analysis of IS*6110* fingerprint patterns using Kaplan-Meier estimates. The survival function is $\exp^{(-K*Time)}$ and the values of K for the early, late and total variant groups are -0.003333,-0.0001776 and -0.0002173, respectively.

Since all strains sampled between 20 days and 250 days were invariant, this provides further evidence for these late variant strains evolving over the course of time, rather than these being due to under-sampling at an earlier point. Taking into account these two populations of observations, survival analysis estimated a half-life of 0.57 years (95 % CI, 0.45 – 0.77) for the early group and 10.69 years (95 % CI, 8.22 – 15.31) for the late group (Figure 1).

Discussion

The use of RFLP data to infer epidemiologic associations is dependent on an understanding of the marker system, including the sources and frequency of variability in that marker. To date, studies to determine the frequency of detection of variant strains have used different inclusion criteria and assumptions in order to provide an estimate on the 'molecular clock' of RFLP for *M. tuberculosis*. The rate of change needs to be fast enough to ensure that most cases not linked through recent transmission have different fingerprints, but slow enough to ensure that most cases linked through recent transmission have identical fingerprints. A half-life in the order of 2-5 years, as observed in San Francisco and The Netherlands suggests that the IS*6110* RFLP pattern changes rapidly enough but not too rapidly for studying ongoing transmission (3). A half-life in the order of 8 years, as observed in the present study, suggests that recent transmission may be overestimated by RFLP typing using IS*6110*.

In this study, the proportion of IS*6110* banding pattern changes identified in serial isolates collected from patients with persistent disease was comparable to those reported in The Netherlands (3) but lower than reported in Germany (6) and San Francisco (15). The comparable proportion of variant isolates seen in this study and The Netherlands study suggests that the rate of change is independent of the incidence of disease. However, in keeping with the observations of other groups, the ability to observe change appears to depend on the number of IS*6110* elements present in the precursor strain as strains with less than five IS*6110* elements were genotypically invariant.

On further analysis, the emergence of these variant strains does not appear to be uniform, suggesting the presence of two distinct populations. The first type of change would occur early in the disease process, prior to anti-tuberculosis treatment. This change might then represent evolution influenced by active growth or adaptation to the new host environment. It is also possible that change was induced in certain previously treated patients during either the latent growth phase or during the subsequent reactivation process. The second type of change would occur in the late disease phase or after treatment. This slower rate of change may be a consequence of a slower growth rate of the bacilli. Similar results were seen in data from the Netherlands (3). Although we have calculated an early rate, it is highly probable that this rate will be an under-estimate of the stability of the banding pattern, as most of the IS*6110* band changes will have occurred prior to the sampling of the first isolate. This is supported by the slow *in vivo* growth rate of *M. tuberculosis*, requiring extended periods before strain population replacement would be observed. Furthermore, if there are long delays in the period between onset of early active disease and the first collection of sputa, then it is possible that newly evolved strains may have been missed and thereby excluded from the rate calculations.

Due to the uncertainty of when the early evolutionary events occurred, we attempted to calculate a rate of change for variant isolates appearing after 90 days (15). The rate estimate of 10.69 years for variant isolates appearing after 90 days was significantly slower than the 2 years calculated from the San Francisco data (3). Review of the different rate calculation methods (3,8) shows that these methods assume that the rate of appearance of variant strains is constant over time. If this assumption is incorrect, as is suggested by this study, then the rate calculated by these methods (3,8) will be sensitive to the proportion of (invariant) cases with extended inter-isolate time intervals arising due to either non-compliance treatment failure or treatment interruption. Comparison of our data with that from The Netherlands shows that there is a significant difference in the proportion of patients with invariant isolates after 210 days (this study (21.2 %), The Netherlands (6.6 %); P = <0.0001). This implies that if treatment efficacy is higher, the proportion of serial isolates with long intervals will be smaller, which may result in a higher rate of change estimate. Consequently, the accuracy of rate calculations are limited by the local epidemiology and therefore rate estimates will be proportional to the effectiveness of the TB control programme.

The epidemiological importance of calculating an evolutionary rate for isolates collected from patients with persistent disease is unclear, given that change is observed only in a very small and unique subset of the patients analyzed (3,6,15). Therefore, this data can not be easily extrapolated to the overall study population. Furthermore, the true extent of genetic diversity within a study population may be masked by incomplete clonal replacement at the time of sputum collection, as extremely low numbers of clonal variants in the sputum specimes will not be detected due to the limited sensitivity of the DNA fingerprinting method. Therefore, strain variants will not be identified in patients where there has been an under-sampling of sputum specimens. Accordingly, it may be difficult to identify chains of transmission based only on genotypic identity (clustering of identical IS*6110* banding patterns).

Although, this study has highlighted the limitations of this sample set, we conclude that this patient set may provide information on how growth rate and/or anti-tuberculosis therapy could influence the IS*6110* mediated evolutionary rate. Furthermore, we do not exclude the possibility that change in the IS*6110* banding pattern may be more frequent in other patient groups. To gain a greater insight into the rate of IS*6110* change it will be essential to determine how transmission influences the evolutionary process (7).

Acknowledgments

This work was made possible grants from the GlaxoSmithKline Action TB Initiative, the Sequella Global Tuberculosis Foundation and the National Research Foundation (THRP). Mrs E. Engelke, Mrs S. Carlini and Mrs M. De Kock are thanked for their technical assistance.

Reference List

1. **Alland, D., G. E. Kalkut, A. R. Moss, R. A. McAdam, J. A. Hahn, W. Bosworth, E. Drucker, and B. R. Bloom**. 1994. Transmission of tuberculosis in New York City. An analysis by DNA fingerprinting and conventional epidemiologic methods. N.Engl.J.Med. **330**:1710-1716.

2. **Beyers, N., R. P. Gie, H. L. Zietsman, M. Kunneke, J. Hauman, M. Tatley, and P. R. Donald**. 1996. The use of a geographical information system (GIS) to evaluate the distribution of tuberculosis in a high-incidence community. S.Afr.Med.J. **86**:40-1, 44.

3. **de Boer, A. S., M. W. Borgdorff, P. E. de Haas, N. J. Nagelkerke, J. D. van Embden, and D. van Soolingen**. 1999. Analysis of rate of change of IS*6110* RFLP patterns of *Mycobacterium tuberculosis* based on serial patient isolates. J.Infect.Dis. **180**:1238-1244.

4. **de Boer, A. S., K. Kremer, M. W. Borgdorff, P. E. de Haas, H. F. Heersma, and D. van Soolingen**. 2000. Genetic heterogeneity in *Mycobacterium tuberculosis* isolates reflected in IS*6110* restriction fragment length polymorphism patterns as low- intensity bands. J.Clin.Microbiol. **38**:4478-4484.

5. **Glynn, J. R., J. Bauer, A. S. de Boer, M. W. Borgdorff, P. E. Fine, P. Godfrey-Faussett, and E. Vynnycky**. 1999. Interpreting DNA fingerprint clusters of *Mycobacterium tuberculosis*. European Concerted Action on Molecular Epidemiology and Control of Tuberculosis. Int.J.Tuberc.Lung Dis. **3**:1055-1060.

6. **Niemann, S., E. Richter, and S. Rusch-Gerdes**. 1999. Stability of *Mycobacterium tuberculosis* IS*6110* restriction fragment length polymorphism patterns and spoligotypes determined by analyzing serial isolates from patients with drug-resistant tuberculosis. J.Clin.Microbiol. **37**:409-412.

7. **Niemann, S., S. Rusch-Gerdes, E. Richter, H. Thielen, H. Heykes-Uden, and R. Diel**. 2000. Stability of IS*6110* Restriction Fragment Length Polymorphism Patterns of *Mycobacterium tuberculosis* Strains in Actual Chains of Transmission. J.Clin.Microbiol. **38**:2563-2567.

8. **Salamon, H., M. A. Behr, J. T. Rhee, and P. M. Small**. 2000. Genetic distances for the study of infectious disease epidemiology. Am.J.Epidemiol. **151**:324-334.

9. **Small, P. M., P. C. Hopewell, S. P. Singh, A. Paz, J. Parsonnet, D. C. Ruston, G. F. Schecter, C. L. Daley, and G. K. Schoolnik**. 1994. The epidemiology of tuberculosis in San Francisco. A population-based study using conventional and molecular methods. N.Engl.J.Med. **330**:1703-1709.

10. **van Embden, J. D., M. D. Cave, J. T. Crawford, J. W. Dale, K. D. Eisenach, B. Gicquel, P. Hermans, C. Martin, R. McAdam, and T. M. Shinnick**. 1993. Strain identification of *Mycobacterium tuberculosis* by DNA fingerprinting: recommendations for a standardized methodology. J.Clin.Microbiol. 31:406-409.

11. **van Rie, A., R. Warren, M. Richardson, T. C. Victor, R. P. Gie, D. A. Enarson, N. Beyers, and P. D. van Helden** . 1999. Exogenous reinfection as a cause of recurrent tuberculosis after curative treatment. N.Engl.J.Med. 341:1174-1179.

12. **van Soolingen, D., P. E. de Haas, R. M. Blumenthal, K. Kremer, M. Sluijter, J. E. Pijnenburg, L. M. Schouls, J. E. Thole, M. W. Dessens-Kroon, J. D. van Embden, and P. W. Hermans**. 1996. Host-mediated modification of PvuII restriction in *Mycobacterium tuberculosis*. J.Bacteriol. 178:78-84.

13. **van Soolingen, D., P. E. de Haas, P. W. Hermans, and J. D. van Embden**. 1994. DNA fingerprinting of *Mycobacterium tuberculosis*. Methods Enzymol. 235:196-205.

14. **Warren, R., J. Hauman, N. Beyers, M. Richardson, H. S. Schaaf, P. Donald, and P. van Helden**. 1996. Unexpectedly high strain diversity of *Mycobacterium tuberculosis* in a high-incidence community. S.Afr.Med.J. 86:45-49.

15. **Yeh, R. W., d. L. Ponce, C. B. Agasino, J. A. Hahn, C. L. Daley, P. C. Hopewell, and P. M. Small**. 1998. Stability of *Mycobacterium tuberculosis* DNA genotypes. J.Infect.Dis. 177:1107-1111.

3

Evolution of the IS*6110* based RFLP pattern during the transmission of *Mycobacterium tuberculosis*

R. M. Warren, G. D. van der Spuy, M. Richardson, N. Beyers, C. Booysen, M.A. Behr, P. D. van Helden

J. Clin. Microbiol. (2002) 40(4):1277-1282

Abstract

Interpretation of the molecular epidemiological data of *M. tuberculosis* is dependent on the validity of the assumptions that have been made. It is assumed that the IS*6110* banding pattern is sufficiently stable to define epidemiological events representing ongoing transmission. However, molecular epidemiological data also supports the observation that the IS*6110* banding pattern may change over time. Factors affecting this rate may include the nature and duration of disease in a host and the opportunity to experience different host environments during the transmission cycle. To estimate the rate of IS*6110* change occurring during the process of transmission, *M. tuberculosis* isolates from epidemiologically linked patients were genotypically characterized by RFLP (restriction fragment length polymorphism) analysis. The identification of IS*6110* banding pattern changes during ongoing transmission suggested that a rate could be estimated. IS*6110* change was significantly associated with strains with > 5 IS*6110* elements (P = 0.013) and was not observed in low-copy isolates. The minimum rate of appearance of variant strains was calculated to be 0.14 variant cases per source-case per year. This data suggests that clustering of isolates based on identical RFLP patterns is expected to underestimate transmission in patients infected with high-copy isolates. A model based on the rate of appearance of both variant and invariant strains demonstrates that the genotypically defined population structure may change by 18.6% during the study period of approximately 6.5 years. The implications for the use of RFLP data for epidemiologic study are discussed.

Introduction

The most extensively researched repeat sequence in the genome *of M. tuberculosis* is the transposon element IS*6110*, a member of the IS3 transposon family (13,17). This element may be repeated up to 25 times per genome, and the distribution and stability of these elements in the chromosome of *M. tuberculosis* has led to the development of an internationally standardized DNA fingerprinting protocol to genotype clinical isolates (18,21). Interpretation of DNA fingerprinting data is based on the assumption that the IS*6110* banding pattern is sufficiently stable to allow the grouping of strains with identical IS*6110* genotypes as recent epidemiological events (termed clusters), while sufficiently variable to allow the classification of strains with unrelated IS*6110* banding patterns (isolates with unique banding patterns) as unrelated epidemiological events (1,16). Furthermore, it is assumed that clustered isolates represent ongoing transmission between epidemiologically linked patients (5). This methodology has been used in numerous settings to quantify the relative proportion of recent epidemiological events and thereby estimate the extent of recently transmitted disease (1,16,23).

While the methodology of generating RFLP (restriction fragment length polymorphism) data is highly standardized, the interpretation varies widely depending on the epidemiological question and assumptions made about the rate of change of RFLP patterns. For instance, in ascribing a case to be a falsely-positive diagnosis due to laboratory cross-contamination, it is usually assumed that the RFLP pattern should be exactly identical to another in the database, as it is unlikely that the organism will evolve a different pattern in the diagnostic laboratory (22). However, in well defined outbreak settings, occasional cases are observed where the RFLP pattern is closely related, but differs by one or two insertion elements (6). The latter data suggest that, as a strain spreads through the community, strain variants will evolve that are progeny of the same epidemic clone but manifest subtle differences in RFLP patterns (4). The rate of RFLP pattern change in a community is largely unknown, yet is an essential element in setting criteria for reporting isolates to be matched or unrelated.

As a first step towards understanding the evolution of RFLP patterns, genotypic analysis of serial isolates collected from patients with persistent disease has been performed to estimate rates of IS*6110* banding pattern change over time (7,14,15,26). The frequency of banding pattern changes reported varied considerably between the different studies, and it was suggested that this may reflect (unspecified) differences in the epidemiology of disease in the different geographical regions (14). Assuming that the IS*6110* banding pattern changes at an equal rate in different strains, de Boer et al. (1999) calculated an evolutionary rate by analyzing serial isolates from 544 patients with persistent disease. Twenty-five of these manifested altered banding patterns over time, permitting the extrapolation of a banding pattern half-life of 3.2 years. When this methodology was applied to data from San Francisco, a half-life of about 2 years was calculated (7).

From these results it was concluded that the evolutionary rate of IS*6110* banding pattern was sufficiently slow to allow molecular epidemiological calculations (7). However, the rate of IS*6110* banding pattern change as a bacterium travels through a community is largely unknown, and is expected to reflect both changes within a host during persistent disease and changes during the transmission cycle, such as when the organism encounters a new host. One study from Germany, looking at cases with epidemiological links, found that pattern alterations in contacts who developed tuberculosis were rare (15). This study however had small numbers of transmission events and did not permit the analysis of chains of transmission, as occurs in a high incidence setting. In this study we investigated the rate of IS*6110* banding pattern change in a high incidence community of the Western Cape Province of South Africa by studying RFLP patterns in patients who reside in the same or neighbouring households. These results are discussed in the context of the rate of appearance of variant strains and the influence of variant strains on the genotypic bacterial population structure as a function of time.

Methods

Study setting

During the period from mid 1992 to December 1998, *M. tuberculosis* isolates were collected from patients attending healthcare clinics within two adjacent suburbs in Cape Town, South Africa (3). This community experiences an extremely high incidence of tuberculosis, with approximately 251 new bacteriologically confirmed adult cases per 100 000 population, per year (19). Before 1995, the National TB Program depended on sputum cultures for the diagnosis of tuberculosis with follow-up sputum samples requested by the attending physician on clinical grounds. Since 1995, the South African National TB Program has been conducted according to the WHO DOTS strategy with sputum taken for Ziehl-Neelsen smear at presentation for diagnosis, at 2 months for sputum conversion and again at six months after initiation of therapy to monitor response to therapy. As part of the research project, all Ziehl-Neelsen positive samples are sent for culture. Clinical information, including the residential address, of each patient was recorded at the time of diagnosis and this data was stored in a Microsoft Access database.

RFLP generation and GelCompar Analysis

All isolates were classified according to the internationally standardized DNA fingerprinting protocol, using the IS-3' probe (18,21). The Southern blot autoradiographs were normalized and the IS-3' bands assigned using GelCompar 4.1 software. The assignments were visually checked by two independent persons, and only bands with an intensity of > 20 % of the average band intensity were scored as representing IS*6110* mediated evolutionary events (8). Replicative transposition was identified when the evolved strain showed an additional IS*6110* hybridizing band, while deletion of an IS*6110* element was identified when an IS*6110* hybridizing band was absent from the evolved strain. Variation in the electrophoretic mobility of an IS*6110* hybridizing band was classified as a band shift representing a mutational event in the chromosomal domain flanking the IS*6110* element. Banding pattern changes suspected of being the result of partial digestion of methylated restriction sites (20) were excluded. All band changes were confirmed by repeating the digestion at least once. Cluster analysis was done using the UPGMA (unweighted pair group method with arithmetic mean) and Dice coefficient (12). Each

IS-3' banding pattern was assigned a cluster number and tabulated in a Microsoft Access table to enable linking of the DNA fingerprinting data to the clinical data.

Study population

Isolates collected from epidemiologically related patients: All patients residing in the same household (including the next door households (n ± 2) on the same side of the street) and infected with either identical strains or genotypically related strains (IS*6110* banding pattern differing by up to 5 hybridizing bands) were identified from the databases. It is assumed that patients residing in these households are epidemiologically linked and that their genotypically related or identical *M. tuberculosis* isolates represent ongoing transmission. These households were divided into two groups; 1) households where the strain genotype was invariant during the course of sampling and 2) households where the strain genotypes included a variant identified during the course of sampling. To avoid a possible bias in the number of patients included in the invariant patient group, strains with more than four changes in the IS*6110* banding pattern were considered to reflect reinfection rather than evolution and were excluded from the study. In addition, if the variant strain was found to be present in the community prior to appearance in the patient, these patients were also excluded. In such cases, it was suspected that the patient may have been reinfected by a community source case. This stringent exclusion criteria excludes the possibility that the IS*6110* banding pattern may revert to a previous evolutionary state, thereby conceivably underestimating the extent of IS*6110* change.

The appearance time interval was calculated as the time (days) from the date when the first isolate was collected in each household (source-case) to the date when the first isolate was collected from each secondary case.

Rate calculation

Assuming that the first case in each household was the source-case (2) and that this case infected all other cases in the household who subsequently developed disease, the rate at which variant isolates appeared (**R_V**) (variant cases per source-case per year) as function of ongoing transmission (in the different households) can be described as follows:

$$R_V = C_V/N_{t0}/t_{av}$$

Where **C_V** is the number of cases with variant isolates (excluding subsequent transmission of the variant strain), **N_{t0}** is the total number of source-cases, and **t_{av}** is the average appearance time interval (years) where **$t_{av} = t_i/T$**, and **t_i** is the sum of the appearance time intervals (years) and **T** is the number of transmission events.

The rate at which invariant isolates appeared (**R_{IV}**) (invariant cases per source-case per year) as a function of ongoing transmission is calculated as:

$$R_{IV} = C_{IV} - N_{t0}/N_{t0}/t_{av}$$

Where **C_{IV}** is the total number of cases with invariant isolates.

Statistical analysis

Fisher's exact test was used to test whether changes in the IS*6110* banding pattern were associated with the number of IS*6110* insertions present in the evolving strain. In addition, the Fisher's test was used to establish whether the appearance of variant isolates was related to a previous history of disease. The Mann-Whitney test was used to determine whether the appearance of variant isolates was related to the median sample collection period thereby demonstrating a correlation between appearance and time.

Results

Study population

During the period from mid 1992 to the end of 1998, *M. tuberculosis* isolates were cultured from 865 patients (representing a recovery of approximately 70 % of all culture positive patients) resident and attending the healthcare clinics in a suburb of Cape Town, South Africa (Figure 1) (3,23). Forty-six of these patients (5 %) were excluded from the study as their cultures were either contaminated or lost viability. Of the remaining patients, eight-hundred and seventeen patients (99.7 %) had an *M. tuberculosis* isolate genotypically classified by IS-3' DNA fingerprinting (Figure 1) (18,21).

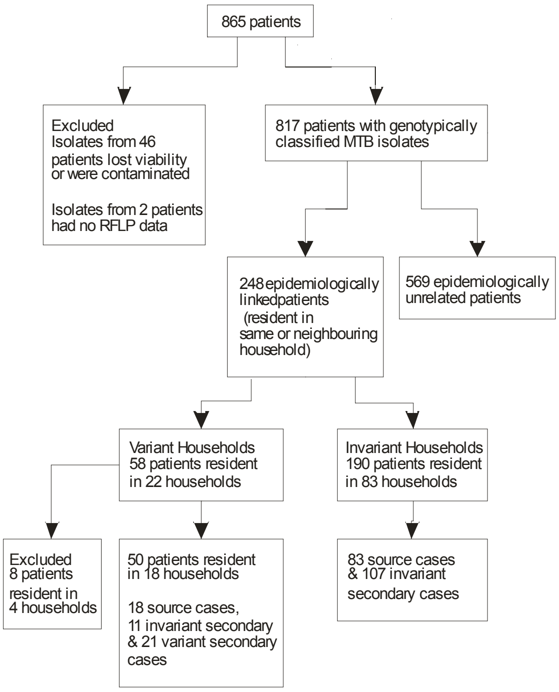

Figure 1. Flow chart showing patients included into the study.

M. tuberculosis isolates collected from epidemiologically related patients

To investigate the relationship between transmission and the appearance of strain variants, all patients who were epidemiologically linked (as they reside in the same household or adjacent households) were identified. A total of 248 patients from 105 households met the inclusion criteria. In eighty-three households, transmission of an invariant strain from 83 source-cases to 107 secondary cases was observed (Figure 1). In the remaining 22 households the appearance of a variant strain/s was observed during ongoing transmission. Four of the households representing patients infected with a variant strain were excluded (4 source-cases and 4 secondary cases), as it was suspected that these patients may have been reinfected by a community contact. In the remaining 18 households, 21 patients were infected with a variant *M. tuberculosis* isolate originating from 18 source-cases (11 patients infected with an invariant isolate were also identified in these 18 households) (Figure 1). Analysis of the IS*6110* banding pattern changes (Figure 2) showed that 13 (62 %) of the variant strains evolved by replicative transposition, while 6 (29 %) evolved by deletion of one or more IS*6110* elements. Only 1 (4.5 %) strain evolved by band shifts, while a further 1 strain (4.5 %) evolved by a combination of replicative transposition and band shifts. This suggests that replicative transposition is the predominant evolutionary mechanism.

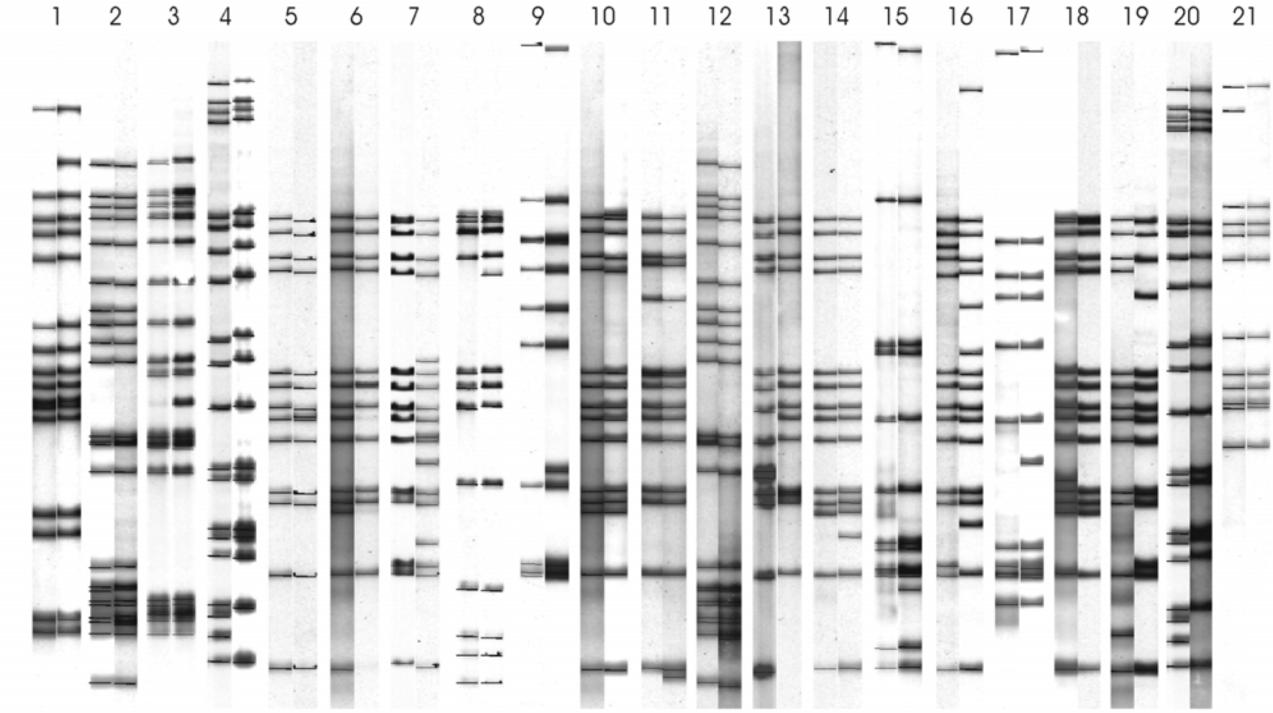

Figure 2. IS*6110* restriction fragment length polymorphisms of *M. tuberculosis* isolates collected form epidemiologically linked patients. Each pair of lanes shows the IS*6110* banding pattern of the source case (left) and the variant secondary case (right) resident in the same or neighbouring household.

Figure 3 shows a plot of the frequency of observed banding pattern changes as a function of IS*6110* copy number. Two major strain groups representing either low IS*6110* copy number or high IS*6110* copy number strains were analyzed, as it has previously been suggested that these groups represent different evolutionary lineages which evolve independently (10,25). The absence of IS*6110* banding pattern changes in strains with fewer than 6 IS*6110* insertions is significant (Fisher's exact test, P = 0.013), demonstrating that the low IS*6110* copy number strains evolve at a different rate from that of the high IS*6110* copy strains. For this reason, the low copy number strains have been excluded from any rate calculations (see below).

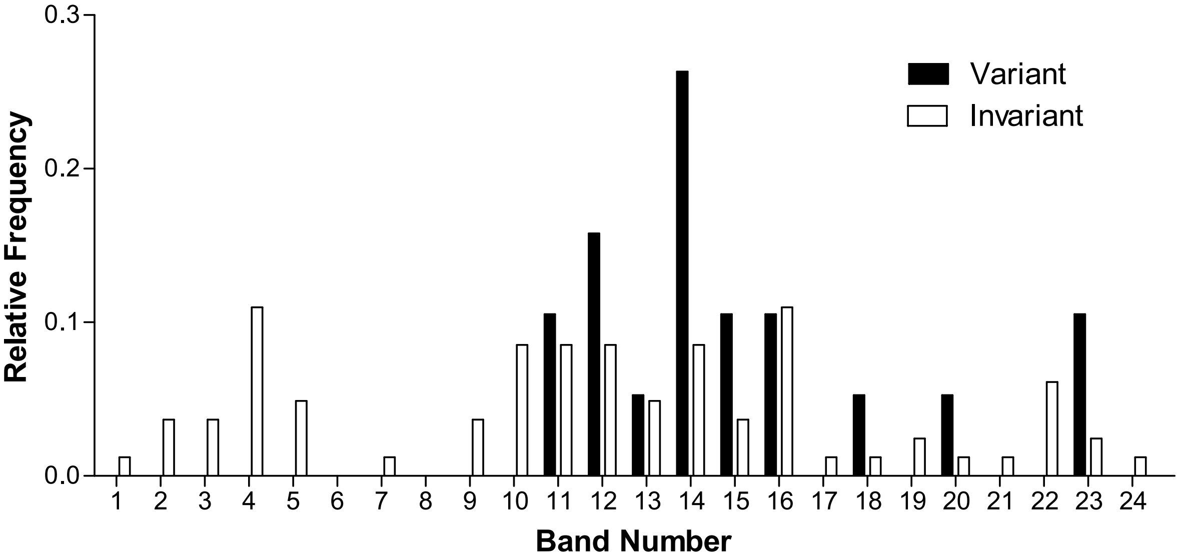

Figure 3. Relative frequency of source-case strains generating either variant or invariant secondary cases in epidemiologically related episodes, as a function of the number of IS*6110* insertions.

Comparison of patients who were infected by a variant strain and those infected by an invariant strain did not identify significant demographic differences between the two patient groups (Table 1). Therefore, it is unlikely that the variant strains arose due to patient specific macro factors which could stimulate IS*6110* mediated genome variation. Review of the clinical records of variant secondary cases showed that in 19 cases (90.5 %) the observed banding pattern changes occurred prior to the initiation of anti-tuberculosis therapy. Therefore, it is unlikely that tuberculosis therapy influences genomic evolution. The appearance of variant isolates was not associated with previous episode of disease (Fisher's exact test , P = 0.62). However, there is a weakly significant difference (Mann-Whitney, P = 0.043) in the median time interval (between the first isolate of the source-case and the first isolate of the secondary case) for the patients with variant and invariant strains. This could suggest that the longer the time interval between infection and the development of disease the greater the chance of observing change.

Table 1. Demographics of patients recently infected within a household (excluding the source-case), where the infecting *M. tuberculosis* strain has > 5 IS*6110* hybridizing elements and is either variant or invariant from the source-case strain.

Demographic	Variant	Invariant
Cases	21	92
Male	14 (67 %)	50 (54 %)
Mean Age (years)	34 ± 11	33 ± 12
Previous episode of TB	10 (48 %)*	36 (39 %)
Pulmonary TB	19 (90 %)	91 (99 %)
Median time (days)	739**	392

* (Fisher's exact, P = 0.62)

** (Mann-Whitney, P = 0.043)

Given the epidemiological links for patients residing in the households studied, it is likely that any changes in the IS*6110* banding pattern must have occurred during the interval between the collection of the first isolate (source-case) and the first variant isolate. Based on the assumption that the first patient to present with tuberculosis in each household infected all of the patients who subsequently presented with disease with a similar genotype (2), a rate of appearance of IS*6110* variants can be expressed as a function of both time and number of infectious sources. Using the rate calculation $R_V = C_V/N_{t0}/t_{av}$, the minimum average rate for which variant high copy number strains appear as tuberculosis cases is $R_V = 0.140$ variant cases/source-case/year. The rate of appearance of invariant isolates is $R_{IV} = 0.614$ invariant cases/source-case/year.

To determine the influence of genotypic variation on the bacterial population structure as defined by IS-3' RFLP analysis, it is assumed that each household represents a subset of the study population. Within each household, strains can be broadly categorized as those that transmit and cause disease in a secondary case, and those that do not cause disease in a secondary case within the study period. As genotypic variation could not be identified in the latter, calculations have focused on determining a rate of change for strains which were transmitted and had the propensity to change (only high copy number strains with an IS*6110* copy number of >5). We propose a simple model to describe the total number of variant and invariant strains (N_{t1}) appearing in the household population as a function of both transmission and evolution within a defined time interval (Figure 4). From this model the rate of genotypic change is calculated as:

$$N_{t1} - [R_V \times N_{t0} + T_V(R_V \times N_{t0}) + R_{IV} \times N_{t0}] \times t_1 + (1 - C)N_{t0}$$

At the onset of the study time interval (t_0) the number of source-cases will be N_{t0}. During the following time interval (t_1) these cases will transmit disease to contacts within their respective households. The rate at which the contacts develop disease with a variant strain will be R_V x N_{t0}. The rate of appearance of an invariant strain will be R_{IV} x N_{t0}, while the rate of transmission of the variant strains will be $T_V(R_V \text{ x } N_{t0})$, where T_V is a transmission rate. It is assumed that a proportion (C) of the source-cases will be cured. The number of infected patients at time t_1 will be N_{t1}.

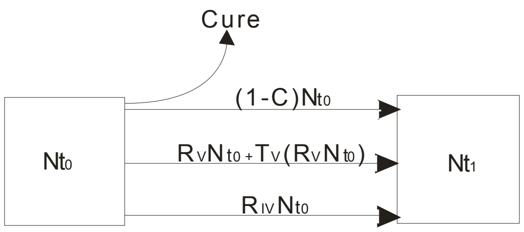

Figure 4. Model describing genotypic change in the bacterial population.

Applying this formula to the data (Table 2), 82 source-cases would yield 113 secondary cases (assuming that all the source-cases were cured during this time interval). Twenty-one of these cases will be expected to be newly evolved strain variants. This suggests that the genotypic population structure in this patient group will change by 18.6 % within the study period of approximately 6.5 years.

Table 2. Summary of the characteristics of the *M. tuberculosis* isolates with > 5 IS*6110* hybridizing elements which were transmitted within the context of the different households

Characteristics	
Sum of appearance intervals (ti) (years)	206.5
Total number of transmission events (T)	113
Average appearance interval (t_{av} = ti/T) (years)	1.83
Number of source-cases (N_{t0})	82
Number of variants cases (C_V)	21
Number of invariants cases (C_{IV} - N_{t0})	92

Cluster analysis of the DNA fingerprint database showed that the variant source-cases subsequently infected 5 patients. This result confirms that variant strains are amplified by secondary transmission and thereby will alter the genotypic bacterial population structure as a function of time.

Discussion

The advent of molecular epidemiology has permitted great advances in the study of tuberculosis and other infectious diseases. While the techniques needed for molecular epidemiological studies have become largely standardized, the interpretation of molecular epidemiological data remains dependent on a number of assumptions. One such assumption is that genotypic identity is a measure of recent epidemiological events, allowing an estimation of the degree to which disease is due to recent or ongoing transmission (1,11,16). However, in the molecular epidemiological analysis of *M. tuberculosis,* this criterion excludes any change in the IS*6110* banding patterns associated with transmission and therefore such isolates may be incorrectly classified as representing remote epidemiological events. The data presented in this study suggests that pattern changes will occur in strains with high IS*6110* copy number and that the impact of these changes on epidemiological analysis should be considered.

A number of practical and theoretical reasons exist for excluding strains differing by one or two bands from RFLP-defined clusters. Foremost among these are the simplicity of dichotomously classifying isolates as either having the exact match, or not. While the imposition of identical matching may not always be appropriate, the alternative of accepting subtle alterations in patterns is problematic because isolates from chains of transmission will develop increasingly divergent patterns. Determining at what point the pattern is no longer similar enough then becomes an arbitrary process. In theory, this process may be facilitated by understanding the evolution of pattern change, however, previously published observations demonstrate that a single DNA fingerprinting probe is not able to accurately define evolutionary direction (9). Thus, a database may contain an RFLP pattern, A, for which there are two similar patterns (B and C), but it is usually not possible to determine if both B and C derive from A (at which point all three patterns should belong to the same cluster) or whether B derived from A which itself derived at some prior time from C (in which case the epidemiologically relevant cluster may include only A and B). Further complicating the interpretation is that a molecular clock of the IS*6110* banding pattern during ongoing transmission is required to suggest rules by which variant strains should be included in or excluded from the clusters. As DNA fingerprinting databases become larger, representing longer time intervals and isolates from wider geographic regions, these questions will become more and more important. This is particularly the case in geographical regions where there is a high incidence of disease and large populations of genotypically related strains are found (24).

In an effort to address some of these questions, this study analyzed the stability of the IS*6110* banding pattern during the transmission of a strain from a source-case to a close contact. As transmission involves the passage of a strain from a diseased person to a contact (who subsequently developed disease), it is conceivable that evolutionary change will occur at any point within this time interval. Preliminary analysis of serial isolates of *M. tuberculosis* showed that evolutionary events are more likely to have occurred prior to sampling (data not shown), therefore to study change it is more appropriate to identify the original source-case strain for comparison. In this study, a significant number of IS*6110* band pattern changes were observed during transmission of *M. tuberculosis* isolates with > 5 copies of IS*6110*. This ability to change was not found to be associated with either the patient demographics or the exposure to anti-tuberculosis therapy. This implies that anti-tuberculosis drugs, in their the present form, do not significantly influence molecular strain identity and therefore the observed evolutionary changes reflect a natural genetic flux. The identification of chromosomal alterations permitted the calculation of a rate of appearance of 0.14 variant cases per source-case per year. The calculated rate implies that while the IS*6110* banding pattern remains stable in the majority of epidemiologically linked events, changes in a subset of transmission events are expected to alter the genotypes observed by 18.6% over the course of a 6.5 year study (2.9% per year). As the patients included in this study are a subset of the tuberculosis patients in the community, this rate of change applies to the broader tuberculosis patient population and suggests that RFLP variants will also occur during transmission events that do not occur within households. Using absolute genotypic identity of strains with > 5 IS*6110* copies to estimate transmission in a community is therefore expected to result in an underestimate of transmission.

The strong association between the ability to change and the number of IS*6110* insertions demonstrates that mutation mediated by IS*6110* in low copy strains is significantly slower than that detected in the high copy strains. Furthermore, no correlation between the number of IS*6110* elements per genome and evolutionary rate could be identified. Therefore, it would appear more prudent for molecular epidemiological calculations to treat these strain groupings separately rather than assuming that characteristics identified in one group of strains can be extrapolated to another group of strains. For this reason a rate of change was calculated using only the transmission data from patients infected with high copy number strains.

The calculations presented herein also highlight the importance of the appearance rate of invariant strains. The absence of evolution within this group suggests that while genotypes are expected to evolve during an epidemiological study, certain genotypes will also persist for extended periods within a study population, representing endemic strains. The observation of such a genotype more than once during a study period may therefore represent either a new transmission event or independent reactivation events. The interpretation of such observations may be aided by further epidemiological information, including the pre-test likelihood of ongoing transmission suggested by additional knowledge about tuberculosis transmission in that community. For these numerous reasons, the use of identical genotypes to estimate recent epidemiological events requires due consideration.

Acknowledgements

This work was made possible grants from the GlaxoSmithKline Action TB Initiative, the Sequella Global Tuberculosis Foundation and the National Research Foundation (THRP). Mrs E. Engelke, Mrs S. Carlini and Mrs M. De Kock are thanked for their technical assistance.

Reference List

1. **Alland, D., G. E. Kalkut, A. R. Moss, R. A. McAdam, J. A. Hahn, W. Bosworth, E. Drucker, and B. R. Bloom**. 1994. Transmission of tuberculosis in New York City. An analysis by DNA fingerprinting and conventional epidemiologic methods. N.Engl.J.Med. **330**:1710-1716.

2. **Behr, M. A., S. A. Warren, H. Salamon, P. C. Hopewell, d. L. Ponce, C. L. Daley, and P. M. Small**. 1999. Transmission of *Mycobacterium tuberculosis* from patients smear-negative for acid-fast bacilli. Lancet **353**:444-449.

3. **Beyers, N., R. P. Gie, H. L. Zietsman, M. Kunneke, J. Hauman, M. Tatley, and P. R. Donald**. 1996. The use of a geographical information system (GIS) to evaluate the distribution of tuberculosis in a high-incidence community. S.Afr.Med.J. **86**:40-1, 44.

4. **Bifani, P. J., B. B. Plikaytis, V. Kapur, K. Stockbauer, X. Pan, M. L. Lutfey, S. L. Moghazeh, W. Eisner, T. M. Daniel, M. H. Kaplan, J. T. Crawford, J. M. Musser, and B. N. Kreiswirth**. 1996. Origin and interstate spread of a New York City multidrug-resistant *Mycobacterium tuberculosis* clone family. JAMA **275**:452-457.

5. **Classen, C. N., R. Warren, M. Richardson, J. H. Hauman, R. P. Gie, J. H. Ellis, P. D. van Helden, and N. Beyers**. 1999. Impact of social interactions in the community on the transmission of tuberculosis in a high incidence area. Thorax **54**:136-140.

6. **Daley, C. L., P. M. Small, G. F. Schecter, G. K. Schoolnik, R. A. McAdam, W. R. Jacobs, Jr., and P. C. Hopewell**. 1992. An outbreak of tuberculosis with accelerated progression among persons infected with the human immunodeficiency virus. An analysis using restriction-fragment-length polymorphisms. N.Engl.J.Med. **326**:231-235.

7. **de Boer, A. S., M. W. Borgdorff, P. E. de Haas, N. J. Nagelkerke, J. D. van Embden, and D. van Soolingen**. 1999. Analysis of rate of change of IS*6110* RFLP patterns of *Mycobacterium tuberculosis* based on serial patient isolates. J.Infect.Dis. **180**:1238-1244.

8. **de Boer, A. S., K. Kremer, M. W. Borgdorff, P. E. de Haas, H. F. Heersma, and D. van Soolingen**. 2000. Genetic heterogeneity in *Mycobacterium tuberculosis* isolates reflected in IS*6110* restriction fragment length polymorphism patterns as low- intensity bands. J.Clin.Microbiol. **38**:4478-4484.

9. **Fang, Z., N. Morrison, B. Watt, C. Doig, and K. J. Forbes**. 1998. IS*6110* transposition and evolutionary scenario of the direct repeat locus in a group of closely related *Mycobacterium tuberculosis* strains. J.Bacteriol. **180**:2102-2109.

10. **Fomukong, N., M. Beggs, H. el Hajj, G. Templeton, K. Eisenach, and M. D. Cave**. 1997. Differences in the prevalence of IS*6110* insertion sites in *Mycobacterium tuberculosis* strains: low and high copy number of IS*6110*. Tuber.Lung Dis. 78:109-116.

11. **Glynn, J. R., J. Bauer, A. S. de Boer, M. W. Borgdorff, P. E. Fine, P. Godfrey-Faussett, and E. Vynnycky**. 1999. Interpreting DNA fingerprint clusters of *Mycobacterium tuberculosis*. European Concerted Action on Molecular Epidemiology and Control of Tuberculosis. Int.J.Tuberc.Lung Dis. 3:1055-1060.

12. **Hermans, P. W., F. Messadi, H. Guebrexabher, D. van Soolingen, P. E. de Haas, H. Heersma, H. de Neeling, A. Ayoub, F. Portaels, and D. Frommel**. 1995. Analysis of the population structure of *Mycobacterium tuberculosis* in Ethiopia, Tunisia, and The Netherlands: usefulness of DNA typing for global tuberculosis epidemiology. J.Infect.Dis. 171:1504-1513.

13. **McAdam, R. A., P. W. Hermans, D. van Soolingen, Z. F. Zainuddin, D. Catty, J. D. van Embden, and J. W. Dale**. 1990. Characterization of a *Mycobacterium tuberculosis* insertion sequence belonging to the IS3 family. Mol.Microbiol. 4:1607-1613.

14. **Niemann, S., E. Richter, and S. Rusch-Gerdes**. 1999. Stability of *Mycobacterium tuberculosis* IS*6110* restriction fragment length polymorphism patterns and spoligotypes determined by analyzing serial isolates from patients with drug-resistant tuberculosis. J.Clin.Microbiol. 37:409-412.

15. **Niemann, S., S. Rusch-Gerdes, E. Richter, H. Thielen, H. Heykes-Uden, and R. Diel**. 2000. Stability of IS*6110* Restriction Fragment Length Polymorphism Patterns of *Mycobacterium tuberculosis* Strains in Actual Chains of Transmission. J.Clin.Microbiol. 38:2563-2567.

16. **Small, P. M., P. C. Hopewell, S. P. Singh, A. Paz, J. Parsonnet, D. C. Ruston, G. F. Schecter, C. L. Daley, and G. K. Schoolnik**. 1994. The epidemiology of tuberculosis in San Francisco. A population-based study using conventional and molecular methods. N.Engl.J.Med. 330:1703-1709.

17. **Thierry, D., M. D. Cave, K. D. Eisenach, J. T. Crawford, J. H. Bates, B. Gicquel, and J. L. Guesdon**. 1990. IS*6110*, an IS-like element of *Mycobacterium tuberculosis* complex. Nucleic Acids Res. 18:188.

18. **van Embden, J. D., M. D. Cave, J. T. Crawford, J. W. Dale, K. D. Eisenach, B. Gicquel, P. Hermans, C. Martin, R. McAdam, and T. M. Shinnick**. 1993. Strain identification of *Mycobacterium tuberculosis* by DNA fingerprinting: recommendations for a standardized methodology. J.Clin.Microbiol. 31:406-409.

19. **van Rie, A., R. Warren, M. Richardson, T. C. Victor, R. P. Gie, D. A. Enarson, N. Beyers, and P. D. van Helden** . 1999. Exogenous reinfection as a cause of recurrent tuberculosis after curative treatment. N.Engl.J.Med. **341**:1174-1179.

20. **van Soolingen, D., P. E. de Haas, R. M. Blumenthal, K. Kremer, M. Sluijter, J. E. Pijnenburg, L. M. Schouls, J. E. Thole, M. W. Dessens-Kroon, J. D. van Embden, and P. W. Hermans**. 1996. Host-mediated modification of PvuII restriction in *Mycobacterium tuberculosis*. J.Bacteriol. **178**:78-84.

21. **van Soolingen, D., P. E. de Haas, P. W. Hermans, and J. D. van Embden**. 1994. DNA fingerprinting of *Mycobacterium tuberculosis*. Methods Enzymol. **235**:196-205.

22. **van Soolingen, D., P. W. Hermans, P. E. de Haas, D. R. Soll, and J. D. van Embden**. 1991. Occurrence and stability of insertion sequences in *Mycobacterium tuberculosis* complex strains: evaluation of an insertion sequence-dependent DNA polymorphism as a tool in the epidemiology of tuberculosis. J.Clin.Microbiol. **29**:2578-2586.

23. **Warren, R., J. Hauman, N. Beyers, M. Richardson, H. S. Schaaf, P. Donald, and P. van Helden**. 1996. Unexpectedly high strain diversity of *Mycobacterium tuberculosis* in a high-incidence community. S.Afr.Med.J. **86**:45-49.

24. **Warren, R., M. Richardson, S. G. D. van der Spuy, T. Victor, S. Sampson, N. Beyers, and P. van Helden**. 1999. DNA fingerprinting and molecular epidemiology of tuberculosis: use and interpretation in an epidemic setting. Electrophoresis **20**:1807-1812.

25. **Warren, R. M., S. L. Sampson, M. Richardson, G. D. van der Spuy, C. J. Lombard, T. C. Victor, and P. D. van Helden**. 2000. Mapping of IS*6110* flanking regions in clinical isolates of *M. tuberculosis* demonstrates genome plasticity. Mol.Microbiol. **37**:1405-1416.

26. **Yeh, R. W., d. L. Ponce, C. B. Agasino, J. A. Hahn, C. L. Daley, P. C. Hopewell, and P. M. Small**. 1998. Stability of *Mycobacterium tuberculosis* DNA genotypes. J.Infect.Dis. **177**:1107-1111.

4

Genetic Distance: A measure of ongoing transmission of *Mycobacterium tuberculosis*

GD van der Spuy, RM Warren, M Richardson, N Beyers, MA Behr, PD van Helden

J. Clin. Microbiol. (2003) 41(12):5640-5644

Abstract

The stability of the genotypic marker IS*6110*, used to define the epidemiology of *M. tuberculosis,* is one of the most important factors influencing the interpretation of DNA fingerprint data. We propose that evolved strains should be considered together with clustered strains to represent chains of ongoing transmission. We used a large fingerprint dataset, collected between 1992 and 1998 from a high-incidence community resident in Cape Town, South Africa, for this study. Inter-strain genetic distances were calculated by counting the banding pattern mismatches in the IS*6110* DNA fingerprints from different isolates. This data demonstrates that the propensity to change by 1 or 2 bands is independent of IS*6110* copy number. Hence, the genetic distance between pairs of isolates can be simply expressed as the number of differences in the banding pattern. From this foundation, a data set has been generated which identifies newly evolved strains. Inclusion of these evolved strains into various molecular epidemiological calculations significantly increased the estimate of ongoing transmission in this study setting. The indication is that nearly all TB in this community is due to ongoing transmission. This has important implications for TB control as it indicates that current control measures are unable to reduce the level of transmission. This technique may also be applicable to the study of low-incidence TB outbreaks as well as the analysis of epidemiological data from other disease epidemics.

Introduction

The discovery of repeated sequences, such as the transposable element *IS6110,* in the genome of *M. tuberculosis* has facilitated the accurate genotypic classification of the disease causing organism infecting an individual within the context of a community (16). Restriction fragment length polymorphism (RFLP) data has become the primary, standardized tool for the molecular epidemiology of *M. tuberculosis*, greatly enhancing the understanding of the tuberculosis epidemic in different settings (2,7,9,14,19). The interpretation of this data remains complex, however, and is dependent on a thorough understanding of the stability of IS*6110* as a marker of relatedness (5,6,12,21,23).

It is generally accepted that, in order for a marker to be useful for epidemiological tracking, its rate of evolution must be low enough so that epidemiologically linked cases will have identical RFLP patterns (recent transmission) (2,7,14); while being sufficiently rapid so as to enable the discrimination of closely related cases from those more distantly related (*i.e.* transmission *vs.* reactivation). Unique RFLP patterns, therefore, are regarded as reflecting reactivation of dormant infections, influx of strains from an outside community or cases transmitted from unidentified sources (2,7,14). This scenario, while convenient, is an oversimplification as it ignores the possibility that, as an organism multiplies within its host, it may give rise to clonal variants of itself, (6,11,22,23) characterized by minor changes in the IS*6110* RFLP banding pattern. These evolved strains may, in turn, be transmitted to new hosts (12,21), leading, over time, to increasing genetic diversity in the bacterial population.

Assuming a constant rate of mutation, genetic distance (GD), defined as the number of mutational events separating two strains, may be regarded as an indicator of the evolutionary time since their divergence from a common ancestor. Thus, a high degree of similarity between two strains implies close temporal coupling. Conversely, the greater the evolutionary time elapsed since divergence, the higher the probability of accumulating mutations and therefore the greater the GD. Studies of the stability of IS*6110* fingerprints have demonstrated a half-life in the order of 2 to 3.2 years (6,11,23). These authors have concluded that this rate of change is sufficiently low so as to facilitate epidemiological tracking. However, the calculated rate is also high enough to significantly impact the interpretation of relatedness in molecular epidemiological studies, especially those where the study duration is similar to, or longer than the half-life of the marker system. Given this non-zero rate of IS*6110* fingerprint variation, clonal variants, appearing within a relatively short time-frame, and differing by a few bands, may represent recent evolution and therefore be regarded as constituting an ongoing transmission chain (12,21,23). In a recent study we found evidence for this and reported that the manifestation of evolution was associated with transmission (21). We suggested that these events probably reflect the overall evolutionary dynamics of the bacterial population in the study setting. Failure to account for recent evolution in assessing epidemiological transmission may, therefore, hinder our understanding of the factors driving the epidemic. This is the case for most currently employed algorithms which regard clonal variants as belonging to independent transmission chains or reactivation events (2,7,14,19).

In this study we have examined the effect of evolution within transmission chains on the interpretation of *M. tuberculosis* molecular epidemiological data. We used a systematic approach based on inter-strain GD to group strains into molecular 'superclusters' representing chains of ongoing transmission. Analysis of this data indicates that the rate of evolution of *M. tuberculosis* strains remains constant and is largely independent of IS*6110* copy number. We estimated the amount of ongoing transmission to be at least 20 percent higher than predicted by more established methods (14). Our results show that the incorporation of evolution in the algorithms quantifying the extent of ongoing transmission may have a profound influence on our understanding of the disease dynamics with consequent implications for epidemiological control strategies.

Methods

This study forms part of a larger, long-term molecular epidemiological project which was approved by the ethics committee of Stellenbosch University.

Study population: During the period: mid 1992 to December 1998, *M. tuberculosis* isolates were collected from patients resident and attending healthcare clinics in two adjacent suburbs in Cape Town, South Africa (3). In this community, approximately 350 new bacteriologically confirmed adult cases per 100 000 population are reported each year. Prior to 1996, all patients were treated at one of the primary care clinics by directly observed therapy, although there was no systematic surveillance for cure rates. In 1996 the WHO DOTS strategy was implemented with all its attendant requirements resulting in the availability of cure rates.

DNA fingerprinting

Each isolate was classified by DNA fingerprinting using the internationally standardized protocol (16,20). The resulting autoradiographs were scanned and analysed using the program, GelCompar II (Applied Maths, Belgium). *M. tuberculosis* isolates with fewer than six IS*6110* bands were excluded from the study as it has previously been shown that the IS*6110* banding patterns in these strains show very little diversity (21), precluding their use in epidemiological tracking (14). A total of 168 isolates suspected of being cross-contaminated (17) or identified as non-tuberculosis mycobacteria were excluded from the study.

Genetic distance

The RFLP fingerprints were aligned, using GelCompar, to maximize the number of matching bands between each fingerprint pair, with tolerance parameters allowing for a 6 percent shift in each pattern as a whole and a 0.4 percent variance in individual band positions. This yielded 332 strains (as defined by distinct IS*6110* patterns), having >five IS*6110* elements, from 708 disease cases. Cases from the same patient were distinguished on the basis of *IS6110* strains differing by more than 4 bands. An exhaustive, pair-wise comparison between each IS*6110* banding pattern was performed using a band matching algorithm (GelCompar II) to generate a GD matrix. This consisted of an N by N table of the number of mismatched IS*6110* bands between each pair of strains. The results were imported into a Microsoft Access database as a table of strain pairs with their corresponding GD's. Based on the assumption that recent evolutionary events are represented by a maximum of four banding pattern differences (6,21), strain pairs with a GD of ≤ four were assigned a putative transmission status of source or secondary based on order of appearance in the community. These assignments were made subsequent to the application of the following filters: 1) Strains occurring only in patients who were <12 years of age or who did not present with pulmonary tuberculosis were excluded as possible sources as they were considered unlikely to transmit the bacillus (1). 2) Strain pairs where the time interval between the last case of the source strain and the first case of the secondary strain was greater than a defined interval (2 or 5 years) were excluded. These intervals were arbitrarily chosen to represent the minimum and maximum period within which progression to active disease would be considered recent infection. For each remaining secondary strain, the source with the Nearest Genetic Distance (NGD) (*i.e.* that one being the most similar) was selected as being the most probable candidate. Where two or more possible source strains had the same NGD, candidacy was equally apportioned between them.

To assess the propensity of the IS*6110* RFLP to change, as a function of the number of insertion elements present in the genome, we determined the number of variant strains produced by source strains, categorized by their number of IS*6110* bands, as a proportion of all new cases attributable to those source strains.

Estimation of recent transmission

To calculate the extent of ongoing transmission, strain pairs were linked together into transmission chains on the basis of common source or secondary strain type using a custom written Perl script. (Source code available at http://www.sun.ac.za/med_biochem/) This process was performed five times for each maximum NGD in the range 0 to 4. Isolates belonging to strain pairs having NGD's less than, or equal to the chosen limit were grouped into 'superclusters'. Ongoing transmission was determined using the formula: $(N\text{-}S)/T$ where N = the number of cases grouped in superclusters, S = the number of superclusters and T = the total number of cases in the study. Because the probability of detecting a primary index case diminishes with increasing temporal proximity to the study commencement date, we formulated an alternative estimate of ongoing transmission where the calculation of S was limited to those superclusters initiated after the first 18 months of the study. Thus, ongoing transmission is defined as $(N_E+N_L–S_L)/T$, where N_E is the number of cases in superclusters initiated within the first 18 months of the study and N_L and S_L are the number of cases in superclusters initiated after the first 18 months of the study and the number of superclusters into which they are grouped, respectively.

Results

During the period from mid 1992 to the end of 1998, 1630 *M. tuberculosis* isolates were cultured from 866 patients resident and attending the healthcare clinics in two adjacent suburbs of Cape Town, South Africa. Of these, 164 isolates were excluded as their cultures were either contaminated, lost viability, or were subsequently found to be non-tuberculosis mycobacteria. The remaining 807 patients corresponded to 849 disease cases as some patients reported with multiple, consecutive infections. This represented approximately 70 percent of all bacteriologically confirmed cases resident in the community. The isolates representing these cases were subjected to RFLP analysis using the internationally standardized method in combination with the IS*6110* probe. A total of 342 different IS*6110* banding patterns were identified. Ten of these banding patterns, representing 140 cases, possessed <6 IS*6110* insertions and were excluded from further analysis as it has previously been shown that these strain genotypes are extremely stable and are therefore unsuitable for epidemiological tracking (14).

Pair-wise analysis of the 332 high-copy-number IS*6110* banding patterns was used to quantify the number of differences between all possible strain pairs. From this data set a total of 3019 strain pairs with ≤ 4 differences were identified, of which 1168 fulfilled the case inclusion criteria and reflected NGD pairs. In this study, we have assumed that strains with an NGD of one to four reflect recent evolutionary events, as previous studies have shown that up to four IS*6110* banding pattern changes may occur during recent transmission (12,21) or persistent disease (11,22,23).

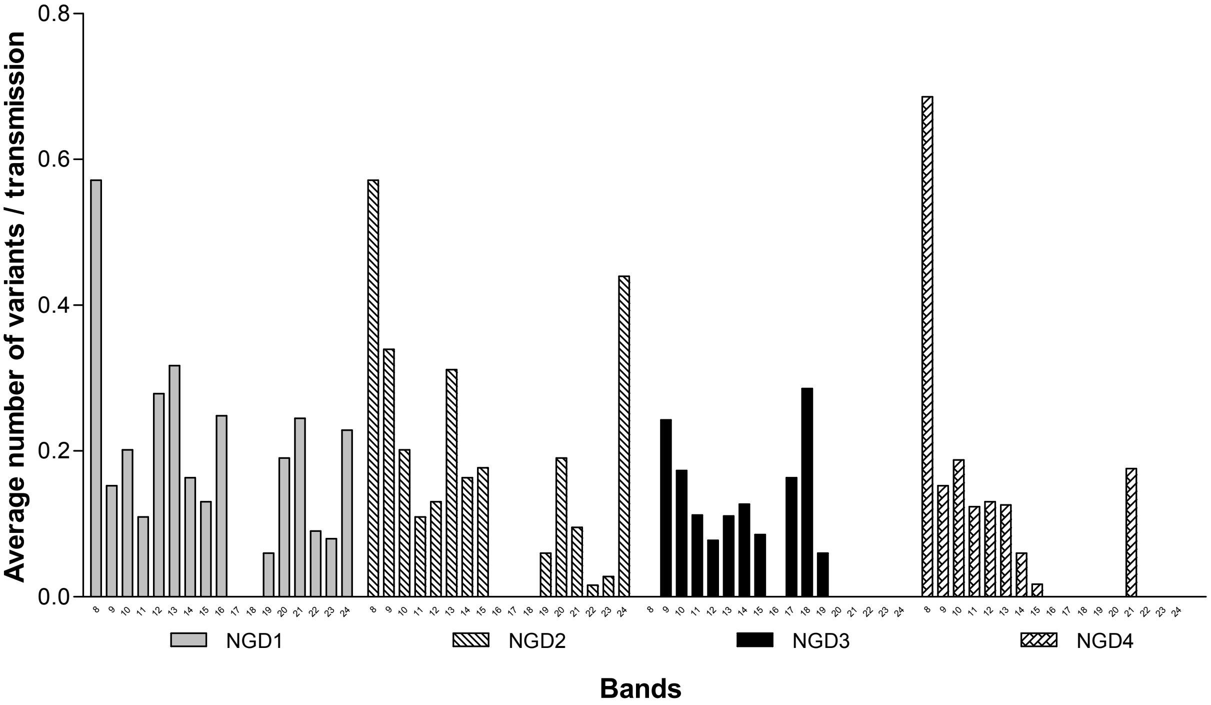

Figure 1. The number of variant strains produced as a proportion of observed transmission events from sources cases with a defined number of IS*6110* bands. Variant strains were produced by the loss or gain of IS*6110* hybridizing bands. Values are for strain pairs with Nearest Genetic Distance (NGD) = 1, 2, 3 & 4. The data shown is for a maximum inter-strain interval of 2 years.

Analysis of these strain pairs showed that the propensity to evolve by either one or two mutational events (NGD = 1 or 2) appears to be independent of the number of IS*6110* hybridizing bands present in the source strain (see Figure 1). This result differs from previous assumptions which have suggested that the rate of change was proportional to the number of IS*6110* elements in the source strain (13,15). IS*6110* mediated mutational events generating three or four banding pattern changes (NGD = 3 or 4) were also found to occur at a constant frequency in strains with 8 to 16 IS*6110* insertions (Figure 1). However, such events were largely absent from strains with 17 to 25 IS*6110* insertions (Figure 1), suggesting that multiple transposition events do not occur or are selected against in very high-copy number strains.

The absence of a clear correlation between propensity to change and the IS*6110* copy number implies a simple relationship between mutational events (up to an NGD = 4) and time, which is independent of the IS*6110* copy number. This is demonstrated in Figure 2, where it is shown that the production of variant strains as a proportion of all new cases is constant for the different NGD values. However, it is also clear that there was a considerable amount of instability in the first 18 months, suggesting that this early phase reflects a lead-in period in which the data is incomplete. For this reason, data from this period was excluded from rate calculations.

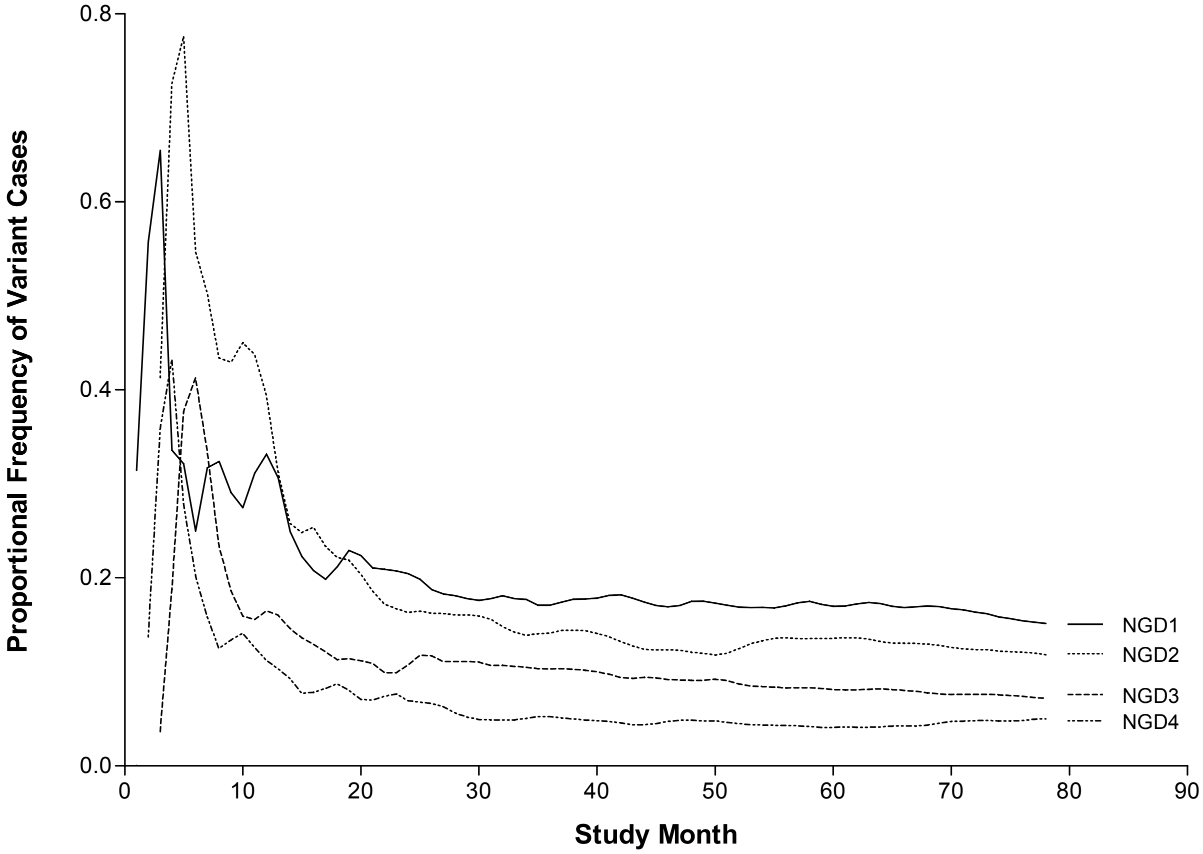

Figure 2. The frequency of new variant strains appearing in the community, as a proportion of observed transmissions over time elapsed since the study epoch. The data shown is for a maximum inter-strain interval of 2 years and is plotted as a 5-month moving average.

The number of variant strains appearing in the community, which could be linked by NGD to an identifiable source strain, divided by the total number of observed transmissions, from month 19 to the end of the study period, is a measure of the rate of variant strain production. The estimates of these rates are given in Table 1. The rate at which variant strains were produced was 13.8 and 10.0 percent of transmissions for NGD = 1 & 2 respectively, using a 2-year inter-strain interval.

Table 1. Estimates of the rate of variant strain production as a proportion of new cases due to transmission appearing per month.

	Nearest Genetic Distance			
	NGD1	NGD2	NGD3	NGD4
Rate (variants/transmission)	0.1384	0.0998	0.0644	0.0445
R^2	0.9997	0.9983	0.9930	0.9988

Using NGD as a measure of recent evolutionary events, occurring during the process of transmission, it was possible to define genotypically related groups of strains (superclusters) representing ongoing transmission chains. From these calculations, assuming a 2-year maximum inter-strain interval, between 54 and 140 isolates, previously classified as unique, were included into superclusters for NGD = 1 and NGD = 4, respectively. Using the formula (N-S)/T (14), the extent of recent transmission ranged from 66 to 89 percent. However, these values may be an under-estimate due to the possible erroneous assignment of primary index status to cases in the early phase of the study. To circumvent this problem we propose the formula $(N_E+N_L–S_L)/T$ in which it is assumed that the source cases can only be defined in superclusters initiated after the first 18 months of the study. Using this formula we estimate the extent of recent transmission to be between 73 percent and 94 percent

Table 2. The degree of superclustering and the standard and alternative calculations of recent transmission for various allowable ranges of NGD. Values are presented for both 2- and 5-year maximum inter-strain intervals.

	Nearest Genetic Distance								
	NGD 0	NGD 0→1		NGD 0→2		NGD 0→3		NGD 0→4	
Inter-strain interval		2 yr	5 yr	2 yr	5 yr	2 yr	5 yr	2 yr	5 yr
Superclustered Cases (N)	497	552	836	589	957	616	1096	638	1087
Superclusters	109	85	92	71	81	59	77	55	64
% Superclustering (N/T)	70.2	78.0	84.9	83.1	90.1	87.0	93.9	90.1	94.9
% Recent transmission ((N-S)/T)	54.8	66.0	75.5	73.2	82.5	78.7	87.3	82.3	89.3
% Recent transmission (NE+NL–SL)/T	63.0	72.9	81.4	79.4	87.5	84.5	92.4	87.9	93.8

Discussion

The simplistic interpretation of DNA fingerprinting data disregards the fact that the genome is in a state of flux (23), evolving at a rate that will influence the interpretation of molecular epidemiological data (6,11,12,21,23).

In this study we have used an algorithmic method to explore the implications of IS*6110* RFLP pattern evolution on the understanding of an epidemic within a high-incidence community. Closely related *M. tuberculosis* strains were linked according to nearest genetic distance. In contrast to previous studies (13), this method of analysis shows that IS*6110* evolution is independent of the number of IS*6110* hybridizing bands present in the source strain. Consequently, genetic distance is purely a measure of the number of band mismatches. However, a number of assumptions have been made in the interests of simplification in the calculation of genetic distance. Firstly, the loss or gain of a band have been assumed to occur at the same rate and therefore, assigned an equal GD. Since >60 percent of IS*6110* fingerprint changes occur by replicative transposition, a more refined method might assign a higher GD to band loss. Secondly, a band shift has been counted as two events, *i.e.* a combination of a loss and a gain. The true frequency of this type of event is obscured by multiple events, but is probably sufficiently rare to validate this assumption.

Analysis of the NGD data shows that closely related variant strains are appearing in the community at a constant annual rate. Thus, for a maximum NGD of 1 or 2 (using a 2-year inter-strain interval), we found 14 to 24 percent of transmission events produced variant strains. This figure is similar to a previous estimate where we found that approximately 18.6 percent of transmission events within households generated a variant strain (21). The high proportion of newly evolved variant strains confirms that the *M. tuberculosis* strain population is diversifying at a constant rate in the study setting and that the process of its evolution is linked to transmission, significantly influencing molecular epidemiologic calculations.

Consequently, studies depending on the stratification of cases according to genetic identity will under-estimate the extent of transmission or incorrectly group cases for risk factor analysis. The factoring of NGD into clustering calculations suggests that transmission estimates may be 20 percent higher than predicted by previously accepted formulae (14). However, the accuracy of this calculation is influenced by a number of factors: 1) The number of source cases initiating transmission chains. To minimize the over-estimation of the number of primary index cases, we proposed an alternative formula in which it is assumed that clusters identified in the first 18 months of the study were initiated by source cases occurring prior to the onset of the study. 2) The estimate assumes a 100 percent sample recovery. In this study only 70 percent of cases were included and therefore, possible source or secondary cases may have been missed, leading to an under-estimate of transmission (10). 3) Currently, there is little data on the extent of *M. tuberculosis* transmission in areas surrounding the study community. As this region experiences an extremely high incidence of disease, it is probable that patients may have been infected by sources outside of the community. 4) While we have demonstrated the propensity to change, this study also shows that most of the transmitted strains remain identical to their source, persisting in the community for extended periods. Using the current methodology it is not possible to differentiate transmission from reactivation of such strains, possibly leading to an over-estimate of transmission. Considering these limitations, we propose that the calculations presented here probably represent a conservative estimate of the true extent of disease due to transmission. Mathematical modelling predicts that ± 95 percent of cases should correspond to transmission given the high infection pressure in this community (18)(PB Fourie, J Lancaster, K Weyer and N Beyers, Medical Research Council of South Africa, Unpublished Manuscript; E Vynnycky, London School of Hygiene & Tropical Medicine, Personal communication, 2002).

Depending on the parameters chosen, this study estimates the proportion of the local epidemic due to ongoing transmission to be between 66 and 94 percent. This is considerably higher than the 55 percent estimated using genetic identity as a measure of transmission. From these results, we conclude that the epidemic is predominantly driven by transmission and not, as indicated by conventional calculations, by reactivation of dormant infections. A positive implication of this conclusion is that interventions which target transmission have the potential to dramatically impact the epidemic. In a setting of passive case-finding, largely based on positive smear-microscopy results, it is hypothesized that the majority of transmission events occur prior to diagnosis and treatment. This is a component of the epidemic which is not targeted by the present WHO DOTS strategy. Based on these results, it is envisaged that interventions which interrupt transmission should coincide with a decrease in GD-based superclustering (8). We suggest that GD should prove be a useful tool in the analysis of longitudinal molecular epidemiological data which will aid in determining the efficacy of *M. tuberculosis* control strategies with a focus on reducing transmission.

While the current study focused on a community with a high prevalence of TB, we believe that the approach presented here may also be relevant to lower-incidence communities. Given the sizeable evolutionary rates reported in studies conducted in such areas (6,12,23), a GD-based analysis of the data may well be expected to produce an estimate of recent transmission significantly different from that obtained by conventional calculations. The rapid diversification of *M. tuberculosis* isolates in the New York outbreak of strain W provides further weight to this argument (4). We feel that the above evidence indicates the need for a similar study to be done on a low TB-incidence community. In addition, we suggest that this technique is not limited to the study of *M. tuberculosis*, but may also prove useful in the analysis of epidemiological data from other disease epidemics.

Acknowledgements

This work was made possible by grants from the GlaxoSmithKline Action TB Initiative, the Sequella Global Tuberculosis Foundation and the National Research Foundation (THRP). Mrs E. Engelke, Mrs S. Carlini and Mrs M. De Kock are thanked for their technical assistance.

Reference List

1. **Agrons, G. A., R. I. Markowitz, and S. S. Kramer.** 1993. Pulmonary tuberculosis in children. Semin. Roentgenol. 28:158-172.

2. **Alland, D., G. E. Kalkut, A. R. Moss, R. A. McAdam, J. A. Hahn, W. Bosworth, E. Drucker, and B. R. Bloom.** 1994. Transmission of tuberculosis in New York City. An analysis by DNA fingerprinting and conventional epidemiologic methods. N. Engl. J. Med. 330:1710-1716.

3. **Beyers, N., R. P. Gie, H. L. Zietsman, M. Kunneke, J. Hauman, M. Tatley, and P. R. Donald.** 1996. The use of a geographical information system (GIS) to evaluate the distribution of tuberculosis in a high-incidence community. S. Afr. Med. J. 86:40-1, 44.

4. **Bifani, P. J., B. Mathema, Z. Liu, S. L. Moghazeh, B. Shopsin, B. Tempalski, J. Driscol, R. Frothingham, J. M. Musser, P. Alcabes, and B. N. Kreiswirth.** 1999. Identification of a W variant outbreak of *Mycobacterium tuberculosis* via population-based molecular epidemiology. JAMA 282:2321-2327.

5. **Cave, M. D., K. D. Eisenach, G. Templeton, M. Salfinger, G. Mazurek, J. H. Bates, and J. T. Crawford.** 1994. Stability of DNA fingerprint pattern produced with IS*6110* in strains of *Mycobacterium tuberculosis*. J. Clin. Microbiol. 32:262-266.

6. **de Boer, A. S., M. W. Borgdorff, P. E. de Haas, N. J. Nagelkerke, J. D. van Embden, and D. van Soolingen.** 1999. Analysis of rate of change of IS*6110* RFLP patterns of *Mycobacterium tuberculosis* based on serial patient isolates. J. Infect. Dis. 180:1238-1244.

7. **Glynn, J. R., J. Bauer, A. S. de Boer, M. W. Borgdorff, P. E. Fine, P. Godfrey-Faussett, and E. Vynnycky.** 1999. Interpreting DNA fingerprint clusters of *Mycobacterium tuberculosis*. European Concerted Action on Molecular Epidemiology and Control of Tuberculosis. Int. J. Tuberc. Lung Dis. 3:1055-1060.

8. **Jasmer, R. M., J. A. Hahn, P. M. Small, C. L. Daley, M. A. Behr, A. R. Moss, J. M. Creasman, G. F. Schecter, E. A. Paz, and P. C. Hopewell.** 1999. A molecular epidemiologic analysis of tuberculosis trends in San Francisco, 1991-1997. Ann. Intern. Med. 130:971-978.

9. **Maguire, H., J. W. Dale, T. D. McHugh, P. D. Butcher, S. H. Gillespie, A. Costetsos, H. Al Ghusein, R. Holland, A. Dickens, L. Marston, P. Wilson, R. Pitman, D. Strachan, F. A. Drobniewski, and D. K. Banerjee.** 2002. Molecular epidemiology of tuberculosis in London 1995-7 showing low rate of active transmission. Thorax 57:617-622.

10. **Murray, M.** 2002. Sampling bias in the molecular epidemiology of tuberculosis. Emerg. Infect. Dis. 8:363-369.

11. **Niemann, S., E. Richter, and S. Rusch-Gerdes.** 1999. Stability of *Mycobacterium tuberculosis* IS*6110* restriction fragment length polymorphism patterns and spoligotypes determined by analyzing serial isolates from patients with drug-resistant tuberculosis. J. Clin. Microbiol. 37:409-412.

12. **Niemann, S., S. Rusch-Gerdes, E. Richter, H. Thielen, H. Heykes-Uden, and R. Diel.** 2000. Stability of IS*6110* Restriction Fragment Length Polymorphism Patterns of *Mycobacterium tuberculosis* Strains in Actual Chains of Transmission. J. Clin. Microbiol. 38:2563-2567.

13. **Salamon, H., M. A. Behr, J. T. Rhee, and P. M. Small.** 2000. Genetic distances for the study of infectious disease epidemiology. Am. J. Epidemiol. 151:324-334.

14. **Small, P. M., P. C. Hopewell, S. P. Singh, A. Paz, J. Parsonnet, D. C. Ruston, G. F. Schecter, C. L. Daley, and G. K. Schoolnik.** 1994. The epidemiology of tuberculosis in San Francisco. A population-based study using conventional and molecular methods. N. Engl. J. Med. 330:1703-1709.

15. **Tanaka, M. M. and N. A. Rosenberg.** 2001. Optimal estimation of transposition rates of insertion sequences for molecular epidemiology. Stat. Med. 20:2409-2420.

16. **van Embden, J. D., M. D. Cave, J. T. Crawford, J. W. Dale, K. D. Eisenach, B. Gicquel, P. Hermans, C. Martin, R. McAdam, and T. M. Shinnick.** 1993. Strain identification of *Mycobacterium tuberculosis* by DNA fingerprinting: recommendations for a standardized methodology. J. Clin. Microbiol. 31:406-409.

17. **van Rie, A., R. Warren, M. Richardson, T. C. Victor, R. P. Gie, D. A. Enarson, N. Beyers, and P. D. van Helden.** 1999. Exogenous reinfection as a cause of recurrent tuberculosis after curative treatment. N. Engl. J. Med. 341:1174-1179.

18. **Vynnycky, E., M. W. Borgdorff, D. van Sooligen, and P. E. Fine.** 2003. Annual *Mycobacterium tuberculosis* infection risk and interpretation of clustering statistics. Emerg. Infect. Dis. 9:176-183.

19. **Warren, R., J. Hauman, N. Beyers, M. Richardson, H. S. Schaaf, P. Donald, and P. van Helden.** 1996. Unexpectedly high strain diversity of *Mycobacterium tuberculosis* in a high-incidence community. S. Afr. Med. J. 86:45-49.

20. **Warren, R. M., M. Richardson, S. L. Sampson, G. D. van der Spuy, W. Bourn, J. H. Hauman, H. Heersma, W. Hide, N. Beyers, and P. D. van Helden.** 2001. Molecular

evolution of *Mycobacterium tuberculosis*: phylogenetic reconstruction of clonal expansion. Tuberculosis. (Edinb.) 81:291-302.

21. **Warren, R. M., G. D. van der Spuy, M. Richardson, N. Beyers, C. Booysen, M. A. Behr, and P. D. van Helden.** 2002. Evolution of the IS*6110*-based restriction fragment length polymorphism pattern during the transmission of *Mycobacterium tuberculosis*. J. Clin. Microbiol. 40:1277-1282.

22. **Warren, R. M., G. D. van der Spuy, M. Richardson, N. Beyers, M. W. Borgdorff, M. A. Behr, and P. D. van Helden.** 2002. Calculation of the stability of the IS*6110* banding pattern in patients with persistent *Mycobacterium tuberculosis* disease. J. Clin. Microbiol. 40:1705-1708.

23. **Yeh, R. W., A. Ponce de Leon, C. B. Agasino, J. A. Hahn, C. L. Daley, P. C. Hopewell, and P. M. Small.** 1998. Stability of *Mycobacterium tuberculosis* DNA genotypes. J. Infect. Dis. 177:1107-1111.

Programming

Matrix.pl

```
# This Perl script reformats the distance matrix output from Gelcompar II into a list of
# isolates pairs (identified by their Gelcompar keys) with their associated genetic distance.

open INPUT,"<matrix.txt" or die "Can't open file: matrix.txt\n";
open OUTPUT,">distance.txt" or die "Can't open file: distance.txt\n";

while (<INPUT>) {
     chomp;
     $matrix[$record++] = [split /\s+/];
}
$records=0;
for $line (0..$#matrix) {
     for $position (1..$line+1) {
          $rdistance = int($matrix[$line][$position]+0.5);
          if ($matrix[$line][0] ne $matrix[$position-1][0]){
               print OUTPUT "$matrix[$line][0]\t$matrix[$position-1][0]\t$rdistance\n";
          }
     }
}
```

Transmission_chains.pl

```
# This Perl script links a list of epidemiologically linked strain pairings into transmission
# chains. The input file format is a simple tab-delimited, 2 column table of
# strain types with a single header row. The output is a list of strain
# types with the transmission chain to which they belong.

open INPUT,"<pairs.txt" or die "Can't open file: pairs.txt";

$count = 0;
<INPUT>; # Discard column labels in 1st line
while (<INPUT>) { # Load pairs matrix and initialise chains
     chomp;
     ($src,$sec)=split /\t/;
     $pairs[$src][$sec] = $pairs[$sec][$src] = ++$count;
     $pairs[$src][0] = $pairs[$sec][0] = 1;  # a valid pair exists for this position ->
                                             # saves processing empty positions
}
close INPUT;

$is3 = $#pairs;
$changeFlag = 1;
$itterate = 0;

while ($changeFlag==1){
     $itterate++;
     $changeFlag = 0;
     foreach $row (1..$is3){
          if ($pairs[$row][0]!=1){ # Next row if this one is empty
               next;
          }
          $chain = $count;
          foreach $col (1..$is3){ # Find the lowest chain number for this row
               if ($pairs[$row][$col] > 0){
                    if ($pairs[$row][$col] < $chain){
                         $chain = $pairs[$row][$col];
                    }
               }
          }
          foreach $col (1..$is3){ # Set all chain numbers in this row = the lowest for the row
               if ($pairs[$row][$col] > 0){
                    if ($pairs[$row][$col] > $chain){ # If needed, make change and set flag
                         $pairs[$row][$col] = $chain;
                         $changeFlag = 1;
                    }
               }
          }
     }
     foreach $col (1..$is3){
          if ($pairs[$col][0]!=1){ # Next column if this one is empty
               next;
```

```
        }
        $chain = $count;
        foreach $row (1..$is3){ # Find the lowest chain number for this col
            if ($pairs[$row][$col] > 0){
                if ($pairs[$row][$col] < $chain){
                    $chain = $pairs[$row][$col];
                }
            }
        }
        foreach $row (1..$is3){ # Set all chain numbers in this col = the lowest for the col
            if ($pairs[$row][$col] > 0){
                if ($pairs[$row][$col] > $chain){ # If needed, make change and set flag
                    $pairs[$row][$col] = $chain;
                    $changeFlag = 1;
                }
            }
        }
    }
}

foreach $row (1..$is3){
if ($pairs[$row][0]!=1){
    next;
}
COL: foreach $col (1..$is3){
        if ($pairs[$row][$col] > 0){ # Find the chain for this strain
            $strains{$row} = $pairs[$row][$col]; # and put it in a hash
            last COL;
        }
    }
}
open OUTPUT,">chains.txt" or die print "Can't open file: chains.txt";
print OUTPUT "IS6110\tChain\n";
foreach $strain (keys %strains){ # Print out the strains with their chains
    print OUTPUT "$strain\t$strains{$strain}\n";
}
close OUTPUT;
print "$itterate itterations\n";
```

5

Effect of study duration on the interpretation of tuberculosis molecular epidemiology investigations

GD van der Spuy, PD van Helden and RM Warren

Tuberculosis (submitted, September 2008)

Abstract

Many molecular epidemiological investigations of *M. tuberculosis* are reported using data collected over relatively short timeframes. We postulated that such studies would tend to under-estimate the amount of disease in a community attributable to ongoing transmission. To test this hypothesis we used 12-year datasets of both real and simulated epidemics with the latter being based on two possible models of transmission. We analysed the effect of viewing the datasets through time windows of varying sizes on the measured degree of strain clustering as an indicator of ongoing transmission. We found that shorter windows significantly under-estimated transmission and that this effect was inversely correlated with the size of a cluster. Accordingly, we recommend that molecular epidemiological studies of *M. tuberculosis*, for the purposes of estimating transmission, be conducted over a minimum of three to four years and that the distribution of cluster-size be taken into account in the interpretation of such data.

Introduction

Molecular epidemiology has established its position as an essential tool in the understanding of tuberculosis (TB). However, as with any analytical instrument, the answers it provides are only as good as the data which is analysed. Technical issues aside, there are often logistical constraints imposed on epidemiological studies. Firstly, such studies tend to be expensive which leads to difficulties in obtaining funding for long-term projects, particularly as there may be little return in the short-term. Many published studies, therefore, report results from data collected over relatively short intervals of 1 to 2 years.[6] In addition, even when a long-term study has been initiated, the academic pressure to publish often leads to publication early in the life of a study. While short-term studies may be acceptable in the case of epidemics of acute infections, this is not necessarily the case in a disease as complex as TB. The ability of Mycobacterium tuberculosis to exist as a latent infection for many years[7] and the fact that a case of TB may take up to two years (or more) after infection to present as active disease, greatly complicate the interpretation of epidemiological data.[11] Further exacerbating these difficulties is the prolonged treatment period with the concomitant risk of failure due to drug-resistance or lack of compliance on the part of the patient. Secondly, molecular data is often unobtainable from a significant proportion of the patients as a result of non-viability of bacterial cultures, contamination, laboratory error, failure of patients to report to health-care facilities or failure to obtain a culture sample due to patients lost to the health-care system before commencement of treatment (initial defaulters). Studies in different settings with varying incidences of TB show different degrees of transmission as indicated by the level of clustering (disease cases infected with the same bacterial strain).[1,6] The effect of under-sampling on estimates of transmission, derived from this measurement, can be dramatic and its magnitude is dependent on the relative distribution of cluster-sizes in the M. tuberculosis strain population.[2]

All epidemiological studies are subject to ‘edge effects’ in which the temporal boundaries of a study introduce artefactual anomalies as a result of chains of transmission which extend beyond these points in time. This can cause both a reduction in the apparent size of clusters, which may result in clustered strain being seen as unique, or even the complete loss of strains.

We hypothesise that calculations derived from molecular epidemiological data that describe a TB epidemic, such as the extent of clustering[5] and recent transmission[6], will be affected in a manner similar to that of under-sampling by the duration of the study interval.

In this paper we aimed to investigate the nature and magnitude of the effect of varying study durations on epidemiological parameters by analysing subsets of a 12-year hypothetical dataset to determine a minimum interval which would provide reliable estimates of these values. We compare the results from this model to those from a 12-year longitudinal study in a setting with a high incidence of TB.

Methods

Simulation data: Using a Perl script (available on request from corresponding author), we generated and analysed simulated populations of 20 000 cases, in clusters (cases having identical strains) of a defined size (2 to 15) for each simulation. Briefly, the script generated 20 000/cluster-size index cases separately for each cluster-size, at random time-points over a 30-year period, with subsequent transmission cases being generated from their source cases with likelihoods of occurring within each of the following five 1-year periods of 54.5%, 22.4%, 16.9%, 4.7% and 1.5% respectively[11]. These subsequent cases were assigned randomly to each of the 12 months within the year to which they had been allocated. We created these datasets according to two models of transmission: 1) Each case in a cluster generated only one daughter case (linear transmission model); 2) Any existing case in a cluster had an equal chance of generating the next daughter case (random transmission model). We used data from years 16 to 27 for the analysis, discarding the initial 15 and all subsequent years to eliminate leading-in and tailing-off anomalies caused by the generation process.

Study population: As part of an ongoing molecular epidemiological study, sputum samples were collected for culture at diagnosis from all new and retreatment TB patients who attended primary health care clinics and who were resident in an epidemiological field site in Cape Town, Western Cape, South Africa during the period January 1993 to December 2004. This community has an extremely high notification rate for TB of 761/100 000 per year for all forms of TB.[10] This study forms part of a larger, long-term molecular epidemiological project which was approved by the ethics committee of Stellenbosch University.

DNA fingerprinting: Sputum isolates were collected from all TB patients and cultured in BACTEC 460, MGIT 960 (Becton Dickinson, Franklin Lakes, NJ USA) or on Löwenstein-Jensen medium. These were then sub-cultured on Löwenstein-Jensen medium and DNA was extracted as previously described.[12]

We classified each isolate by IS6110 DNA fingerprinting using the internationally standardized protocol.[9,13] The RFLP fingerprints were aligned, using GelCompar II (Applied Maths, Sint-Martens-Latem, Belgium), to maximize the number of matching bands between each fingerprint pair, with tolerance parameters allowing for a 5 % shift in each pattern as a whole and a 0.6 % variance in individual band positions. Strains were identified according to distinct IS6110 banding patterns using Gelcompar II. Strains having fewer than 6 bands were excluded from this study due to their low rate of evolution which precludes the application of clustering calculations based on IS6110. We defined transmission chains (clusters) as a series of cases having the same strain of M. tuberculosis with inter-case intervals of up to 2 years.

Data analysis: We analysed the real and simulated datasets as sliding window periods of 1 to 12 years. Clustering[6] and recent transmission (n-1 formula)[5,6] were calculated according to the published formulas for each sliding window and reported as an average for each window size. In the case of the simulated data, we further averaged these parameters over repeated analyses of 20 datasets for each cluster-size to reduce the variability introduced by the random case generator.

In the analysis of the real data, we determined the number cases incorrectly classified as have unique genotypes by subjecting a subset of the data which excluded true unique cases to the same analysis. Any cases subsequently observed to be unique would therefore be classified as such erroneously.

The results were analysed using Graphpad Prism 5 (La Jolla, CA USA).

Results

Analysis of Simulated data

To investigate the effect of study duration on estimates of clustering, we required a simulated dataset which modelled the transmission of *M. tuberculosis* in a real epidemic setting. To this end, we considered two transmission models (described above) representing extreme alternatives: a) linear transmission; b) random transmission. In order to evaluate these models, we tested their predictions regarding subsequent serial intervals within a transmission chain by comparing these against the pattern observed for the real data. The linear transmission model predicts that the intervals between each case and the next case of the same strain should, on average, remain constant, independent of the size of the cluster at that point in time. In contrast, the random transmission model, predicts that the serial intervals between clustered cases should decrease at a rate inversely proportional to the current cluster size as the number of possible source cases increases. To determine which model most accurately simulated the reality, we analysed a subset of the real data which excluded strains observed within the first two years to ensure that all recent transmission chains started with the first case. Allowing for a lead-in period of two years, we plotted the average serial interval for newly-emerged strains as a function of cluster-size at that time-point. We found that the interval decays exponentially at a rate of 0.344/case (data not shown) when fitted to the equation: $SI = SI_0 * x^{(-k)}$, where SI = serial interval, SI_0 = the first serial interval, x = the cluster-size-1 and k = the rate constant. As the theoretical rate constant for the linear model is 0, and for the random model is 1, this result suggests that the reality lies somewhere between the two alternatives, tending toward the linear transmission model. Consequently, we included both models in subsequent analysis.

For the analysis of the simulated datasets, we separated the hypothetical epidemic into clusters of specific size in order to demonstrate the how the cluster-size distribution influences the degree of error associated with study duration. As the difference in the effect observed diminishes exponentially with increasing cluster-size, we limited our investigation of the simulated data to clusters of 15 or fewer cases. We found that estimated clustering was dramatically influenced by the duration of the study window, resulting in an under-estimate, in the worst scenario, of 71.4 ± 2.3% and 73.5 ± 2.5% (for the linear and random models respectively) for clusters of two cases using a one-year window (Figure 1). We also noted that this effect is highly dependent on the size of a cluster with smaller clusters being more severely affected so that the corresponding under-estimates for clusters of 15 cases were 52.5 ± 1.8% and 16.5 ± 1.4% for the linear and random models respectively. Estimates of recent transmission are even more dramatically affected as, in this case, each cluster is reduced in size by 1 (accounting for the index case), yielding values under-estimated by 63.6 ± 0.9% and 63.5 ± 0.8% of expected for clusters of two cases, and 72.4 ± 1.1% and 38.9 ± 1.2 of expected for clusters of 15 cases using a one-year window, for the linear and random models respectively (data not shown). As the window-size increases, estimates of clustering approach, but do not reach 100 % for window-size less than infinity as there will always be some loss of cases at the temporal boundaries. Once again, this is more noticeable for smaller clusters. Whereas small clusters (2 or 3 cases) behave similarly, irrespective of the model used, predictions for larger clusters are less affected by the width of the study window in the random transmission model (compare Fig. 1A and 1B).

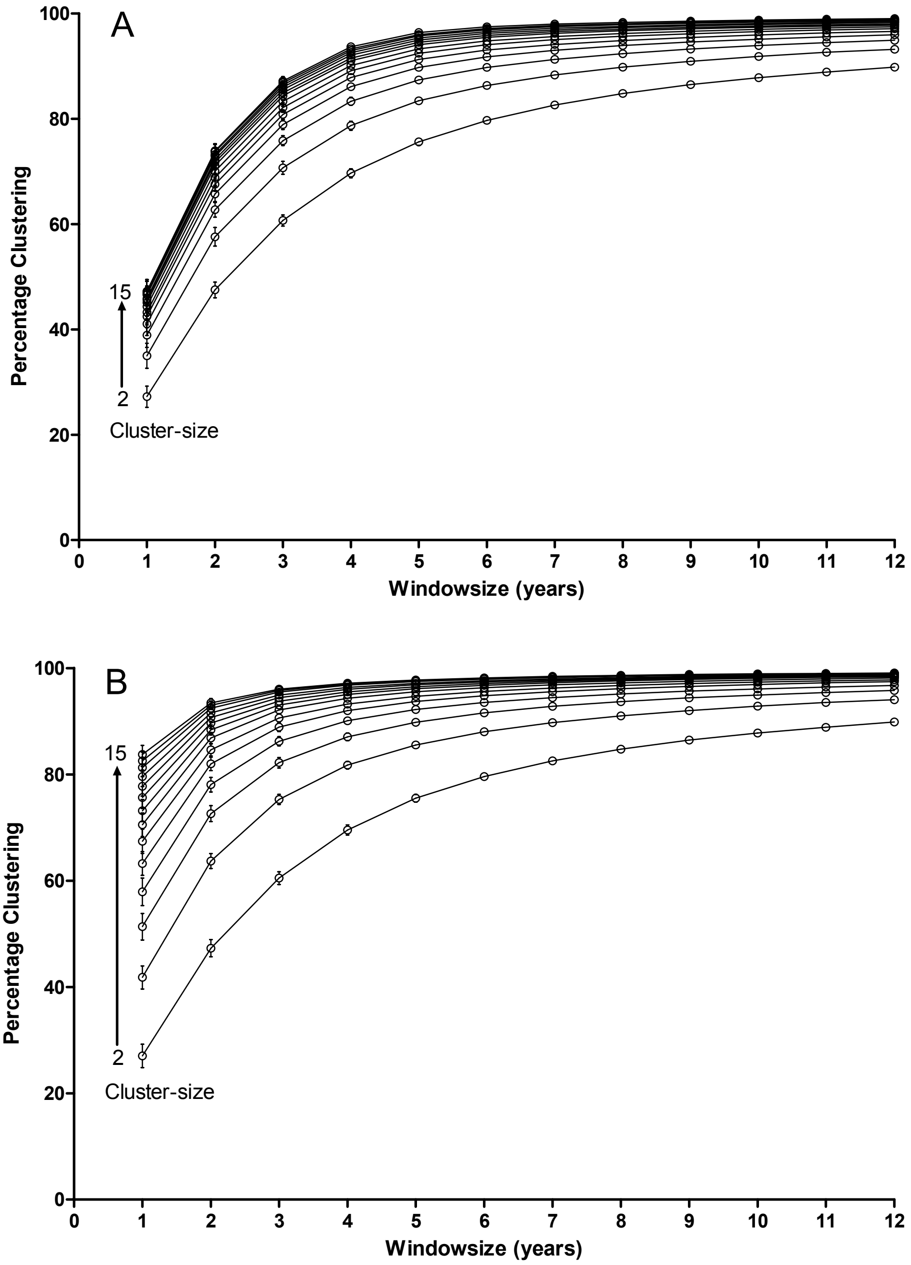

Figure 1. Estimated clustering for simulated clusters of 2 to 15 cases (from bottom to top) as a function of study length using the A) linear and B) random transmission models.

Analysis of real data

Unlike the simulated data, which is divided equally into clusters of 2 to 15, the real-world situation is far more complex. Table 1 shows the distribution of cases within clusters of various sizes for the dataset derived from the study community, as determined from the full 12-year period. It is immediately apparent that the distribution is far from uniform and includes both unique strains and clusters far larger than 15. The latter represent strains which are endemic to the community and span the entire study period. The analysis presented below is based on the total strain population found in the study community and the magnitude of the under-estimates at different window-sizes is dependent on the relative frequency distribution of the cluster-sizes.

Table 1. Distribution of cases by cluster-size

Cluster-size	Cases			
	Cape Town South Africa	San Francisco USA[a]	Zaragoza Spain[b]	Alabama USA[c]
1 (unique)	508	282	215	1038
2	172	40	46	96
3	105	39	18	57
4	116	16	20	40
5	40	10	15	20
6	24	0	6	24
7	7	0	21	63
8	24	8	16	32
9	18	0	0	36
10	40	10	0	10
11	11	0	0	22
12	24	0	12	0
13	26	0	0	0
14	0	0	0	28
15	30	15	0	15
16	0	0	0	16
21	0	0	0	21
23	0	23	0	0
26	0	0	0	26
30	0	30	0	0
35	35	0	0	0
38	0	0	0	38
41	41	0	0	41
44	44	0	0	0
64	64	0	0	0
85	0	0	85	0
136	136	0	0	136

[a] Small PM 1994 2-year study [1]

[b] Lópes-Calleja AI 2007 4-year study [11]

[c] Kempf M-C 2005 (state-wide) 6-year study [12]

We subjected the real dataset to a similar analysis to that of the simulated data by viewing it as a series of rolling time windows of varying sizes. As was the case for the simulated data, we found that the average estimates of clustering and recent-transmission, over all windows of a particular size, decreased as the window-size was reduced (Fig. 2).

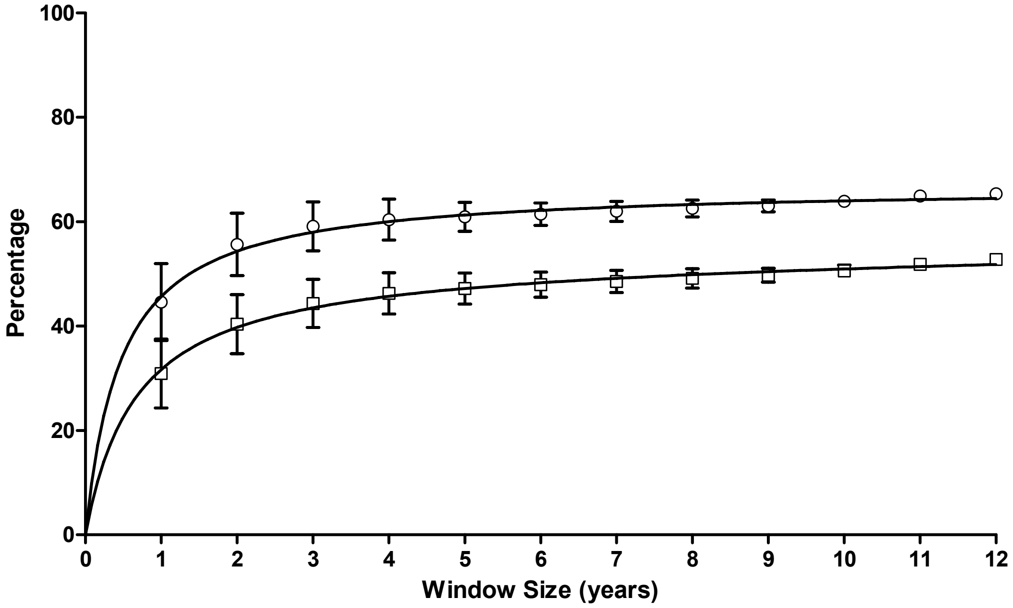

Figure 2. Average percentage clustering (○) and recent-transmission (n-1 formula[1]) (□) with SDs of all cases in the real dataset analysed by sliding windows of varying sizes.

We found values for clustering and recent-transmission for the full 12-years of 65.3% and 52.8% respectively, which were reduced to averages of 44.6 ± 7.4% and 30.9 ± 6.6% for one-year windows. In accordance with these findings, we further found that the number of cases incorrectly classified as unique (not forming part of a transmission chain) increased with decreasing window-size to a maximum average of 23 ± 3.3 per year (representing 19.1% of the dataset) for a one-year study period (data not shown).

We were able to repeat this analysis separately for clusters of 2, 3 and 15 and found that the plots of clustering for these three subsets matched what would be expected from the simulation (Fig. 3).

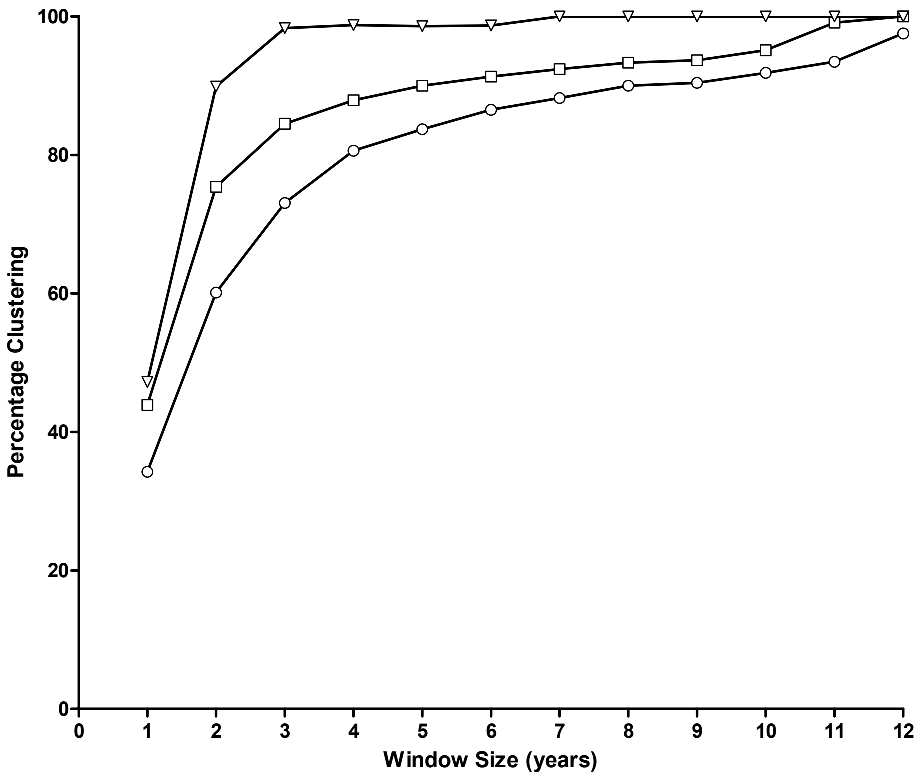

Figure 3. Percentage clustering of real data by window-size for clusters of 2 (○), 3 (□) and 15 (▽) cases. SDs are omitted for clarity, being high as a result of very low numbers of cases at each point.

Discussion

From the analysis of both simulated and real data, it is clear that the duration of an epidemiological study has a significant influence on the calculated parameters describing the degree of transmission in any particular study context. We have demonstrated that, as the time-window through which an epidemic is viewed is shortened, the apparent degree of clustering diminishes at an increasing rate, resulting in an under-estimation of the proportion of disease that is attributable to ongoing transmission. While our theoretical analysis assumes an epidemic is equilibrium over the 12-year duration, the principal still applies in the case of a changing bacterial population as may be seen from our analysis of the real data.

It is also apparent that the magnitude of this effect is strongly influenced by cluster-size, with smaller clusters being more vulnerable to the effects of shorter study periods. Thus, the extent of under-estimating the proportion of disease due to transmission will be dependent on the relative distribution of cluster-sizes found in the population, being an average of the under-estimates for each cluster-size in the proportion in which it is represented. We observe that it is generally the case that real datasets are dominated by clusters of smaller size (Table 1). In this light, it is interesting to note that, if an epidemic contains a large proportion of small clusters, clustering will always be significantly under-estimated for studies of practical duration as a result of the cases which are lost at the temporal boundaries (Fig. 1).

Our analysis of the inter-case intervals for the real data indicated that the transmission pattern lies between those of the two theoretical models, favouring that of the linear model. This is most likely due to a combination of a limited window available for spreading infection before the commencement of treatment and the limited number and overlapping nature of the social contacts of persons forming part of a chain of transmission.[8]

In evaluating the two models of transmission patterns, we noted that larger clusters were less affected by window-size in the random transmission model. This is because, as the cluster-size increases in this model, the inter-case intervals decrease with the consequence that a large cluster will tend to be spread over a smaller time period than in the case of the linear transmission model. As transmission and incubation periods may vary from one community to another, these factors may introduce further variability into the calculation of clustering with more linear transmission patterns and longer inter-case intervals exacerbating the effect of reduced study duration.

We further observe that, as a result of temporal variation, the shorter the study, the higher the variance associated with the estimates of transmission (Fig. 2). This effect will be exacerbated by smaller datasets in settings with a low incidence of TB. We therefore suggest that 'snapshot' studies of an epidemic will tend to be inaccurate and unreliable.

The phenomenon of underestimation of the level of clustering by under-sampling has already been noted by Glynn, et al.[2] The effect of under-sampling will interact with the effects of study-duration, shown here, in a two-fold manner. Not only will it compound the observed decrease in the level of clustering directly, but it will also tend to shift the apparent distribution of cluster-sizes towards smaller clusters, thus, indirectly, further exacerbating the effect of study-duration. The combination, therefore, of a low sampling rate in a study of short duration can be expected lead to a significantly distorted picture of the epidemic under investigation.

As estimates of clustering/recent transmission are surrogates for the efficacy of TB control programs, it is imperative that they be as accurate as possible. Studies which incorrectly attribute a large proportion of disease to reactivation of latent TB or influx from other regions may inform the adoption of suboptimal strategies within the TB control program.

A further cause for concern is the incorrect identification of cases as having unique genotypes within the population. As a result of the loss of cases, both by under-sampling and study window edge-effects, small clusters may be reduced to single cases. Unique strains are generally regarded as reflecting reactivation and as having an impaired ability to transmit. As such, they are often used in studies related to pathogen-based risk-factors for transmission and an error in their identification would invalidate such investigations.

Any study, therefore, which depends on the estimation of transmission using clustering calculations, or on the identification of cases according to their transmissibility will be affected by the issues presented in this paper. In the light of these results, it is our recommendation that molecular epidemiological studies of M. tuberculosis, for the purposes of estimating transmission, be conducted over a minimum of three to four years and that the distribution of cluster-size be taken into account as a possible source of bias in the interpretation of such data.

Acknowledgements

The European Commission 6th framework programme on research technological development demonstration (project no. 037919) are thanked for financially supporting this project. The authors wish to thank Mrs E. Engelke, Mrs S. Carlini, Mrs M. De Kock, and Mr S. Ndabambi for their technical assistance. We also wish to the Department of Health, City of Cape Town for permission to conduct research in the clinics and we are indebted to the residents of the epidemiological field site.

Reference List

1. **Borgdorff MW, Nagelkerke N, van Soolingen D, de Haas PE, Veen J, van Embden JD.** Analysis of tuberculosis transmission between nationalities in the Netherlands in the period 1993-1995 using DNA fingerprinting. *Am J Epidemiol* 1998;**147**:187-195.

2. **Glynn JR, Vynnycky E, Fine PE.** Influence of sampling on estimates of clustering and recent transmission of Mycobacterium tuberculosis derived from DNA fingerprinting techniques. *Am J Epidemiol* 1999;**149**:366-371.

3. **Kempf MC, Dunlap NE, Lok KH, Benjamin WH, Jr., Keenan NB, Kimerling ME.** Long-term molecular analysis of tuberculosis strains in alabama, a state characterized by a largely indigenous, low-risk population.*J Clin Microbiol* 2005;**43**:870-878.

4. **Lopez-Calleja AI, Lezcano MA, Vitoria MA, Iglesias MJ, Cebollada A, Lafoz C, Gavin P, Aristimuno L, Revillo MJ, Martin C, Samper S.** Genotyping of Mycobacterium tuberculosis over two periods: a changing scenario for tuberculosis transmission. *Int J Tuberc Lung Dis* 2007;11:1080-1086.

5. **Murray M, Alland D.** Methodological problems in the molecular epidemiology of tuberculosis. *Am J Epidemiol* 2002;**155**:565-571.

6. **Small PM, Hopewell PC, Singh SP, Paz A, Parsonnet J, Ruston DC, Schecter GF, Daley CL, Schoolnik GK.** The epidemiology of tuberculosis in San Francisco. A population-based study using conventional and molecular methods. *N Engl J Med* 1994;**330**:1703-1709.

7. **Stead WW, Bates JH.** Recurrent tuberculosis due to exogenous reinfection. *N Engl J Med* 2000;**342**:1050.

8. **Uys PW, Warren RM, van Helden PD.** A threshold value for the time delay to TB diagnosis. *PLoS ONE* 2007;**2**:e757.

9. **van Embden JD, Cave MD, Crawford JT, Dale JW, Eisenach KD, Gicquel B, Hermans P, Martin C, McAdam R, Shinnick TM.** Strain identification of Mycobacterium tuberculosis by DNA fingerprinting: recommendations for a standardized methodology. *J Clin Microbiol* 1993;**31**:406-409.

10. **Verver S, Warren RM, Munch Z, Vynnycky E, van Helden PD, Richardson M, van der Spuy GD, Enarson DA, Borgdorff MW, Behr MA, Beyers N.** Transmission of tuberculosis in a high incidence urban community in South Africa. *Int J Epidemiol* 2004;33:351-357.

11. **Vynnycky E, Fine PE.** Lifetime risks, incubation period, and serial interval of tuberculosis. *Am J Epidemiol* 2000;**152**:247-263.

12. **Warren R, de KM, Engelke E, Myburgh R, Gey van PN, Victor T, van Helden P.** Safe Mycobacterium tuberculosis DNA extraction method that does not compromise integrity. *J Clin Microbiol* 2006;44:254-256.

13. **Warren RM, Richardson M, Sampson SL, van der Spuy GD, Bourn W, Hauman JH, Heersma H, Hide W, Beyers N, van Helden PD.** Molecular evolution of Mycobacterium tuberculosis: phylogenetic reconstruction of clonal expansion. *Tuberculosis (Edinb)* 2001;**81**:291-302.

Programming

Window_Clustering.pl

```
# This Perl script calculates the average clustering (or recent transmission) with SD using
# sliding windows with window sizes of 1 to n years for an n year study period.
# The input file has 1 header row (which is not used) followed by the data.
# The first column is the strain with subsequent columns being the case-count for that strain
# for each of the years in the study. Each column is separated by a tab.
# e.g.
# Strain  1991 1992 1993 1994 1995 1996 1997 1998 1999 2000 2001 2002
# 132     0    0    0    0    0    0    0    0    1    4    1    1
# 133     0    0    0    0    0    2    2    1    3    0    0    0
# 134     0    0    1    1    2    1    3    0    0    0    0    0
# 135     1    2    3    0    0    0    0    0    0    0    0    0
# 136     2    1    1    0    0    0    0    0    0    0    0    0
# 138     1    1    4    0    0    0    0    0    0    0    0    0
# 139     0    0    0    0    0    0    0    0    0    1    1    3
# 141     0    0    0    0    0    0    0    1    3    1    1    1

open INFILE, "<pivot.txt" or die "Can't open data file 'pivot.txt'";
open OUTFILE, ">result.txt" or die "Can't open data file 'result.txt'";

$header = <INFILE>;
chomp $header;
($trash,@years) = split /\t/, $header; # Column labels in 1st line
$maxwinsize = scalar(@years);

while (<INFILE>){
     next unless m/.+/;
     ($strain,@cases) = split/\t/;
     $pivot{$strain} = [@cases];
}

print OUTFILE "AvClustering\tSD\tAvRtransmission\tSD\n";

for $winsize (1 .. $maxwinsize){ # Iterate through each window size
     @casecount = ();
     # Iterate through all possible windows for this window size
     for $window (1 .. $maxwinsize - $winsize + 1){
          foreach $strain (keys %pivot){
               foreach $year ($window - 1 .. $window-1 + $winsize - 1){
                    # Sum the cases for each strain over each year in the window
                    $casecount[$window]{$strain} += $pivot{$strain}[$year];
               }
          }
     }
     $sumtotal = $sumuniques = $sumstrains = $sumclustering =  $sumrtrans = 0;
          @total = @strains = @uniques = ();
          for $window (1 .. $maxwinsize - $winsize + 1){
          foreach $strain (keys %pivot){

               if ($casecount[$window]{$strain} > 0){ # If there are any of this strain in this
window
                    # Sum the cases for each  strain over each window of this windowsize
                    $total[$window] += $casecount[$window]{$strain};
                    $strains[$window]++;
                    if ($casecount[$window]{$strain} == 1){
                         $uniques[$window]++;
                    }
               }
          }
          if ($total[$window] > 1){
               $sumclustering += ($total[$window] - $uniques[$window]) / $total[$window] * 100;
               $sumrtrans += ($total[$window] - $strains[$window]) / $total[$window] * 100;
               $sumtotal += $total[$window];
               $sumuniques += $uniques[$window];
               $sumstrains += $strains[$window];
          }
     }

     $windowcount = $maxwinsize - $winsize + 1;
     $avclustering = $sumclustering / $windowcount;
     $avrtrans = $sumrtrans / $windowcount;
     $avtotal = $sumtotal / $winsize;
     $avuniques = $sumuniques / $winsize;
```

```
    $avstrains = $sumstrains / $winsize;

    $sum_of_squares_clust = $sum_of_squares_rt = 0;
    for $window (1 .. $windowcount){
        if ($total[$window] > 1){
            $sum_of_squares_clust += ((($total[$window] - $uniques[$window]) / $total[$window])
* 100 - $avclustering)**2;
            $sum_of_squares_rt += ((($total[$window] - $strains[$window]) / $total[$window]) *
100 - $avrtrans )**2;
        }
    }
    $SD_clust = sqrt( $sum_of_squares_clust / ($windowcount) );
    $SD_rt = sqrt( $sum_of_squares_rt / ($windowcount) );
    print OUTFILE "$avclustering\t$SD_clust\t$avrtrans\t$SD_rt\n";
}

close INFILE;
close OUTFILE;
```

Simulation.pl

```
# This Perl script generates a series of simulated tuberculosis epidemics of specific
# clustersizes from $minclustersize to $maxclustersize. Each epidemic contains 20000 cases
# which are used to calculate % clustering etc. using rolling windows of 1 to $winsizemax.
# Transmission may be linear (one case gives rise to one case) or random (any existing case
# may produce the next case in that cluster) according to the value of $random_model.
#
#
$itterations = 20;    # The number of times to run this analysis
$cases = 20000;   # Constant defining dataset size
$random_model = 1;    # Set to 0 for linear model
$minclustersize = 2;
$maxclustersize = 15;
$yearspan = $maxclustersize * 2;    # The number of years over which cases can be generated
                                    # (create lead-in and lead-out period equal to maxclustersize)
$yearmin = $maxclustersize + 1;# Start of cluster analysis (exclude lead-in period of
                               # maxclustersize)
$winsizemax = 12; # The largest window size to analyse

sub roundup {
    my($number) = shift;
    return int($number + .999);
}

for $itteration (1 .. $itterations){
    print "Itteration: $itteration\n";
    @straincount = ();
    # Generate the dataset
    print "Generating dataset\n";
    for $clustersize ($minclustersize .. $maxclustersize){
        $previous_strain = 0;
        @casedate = ();
        for $case (1 .. $cases){
            # Assign strain type to each case based on the current clustersize
            $strain = roundup($case/$clustersize);
            if ($strain == $previous_strain){
                $generator = 0;   # Set to be the index case;
#           Random source start
                if ($random_model = 1){
# Choose which existing case will generate a new case with each having a 1/(no. of previous
# cases of this strain) chance this time round
                    while (rand() > 1/@casedate){
# If generator number is > the number of existing cases for this strain
                        if ($generator++ == @casedate){
                            $generator = 0;   # Reset to be the index case;
                        }
                    }
                }
#           Random source End
                $yearselector = int(rand()*1000);
                $randmonth = int(rand()*11+1);
# Population month of previous case plus 1 to 11 months or 12 to 23 months etc. up to 5 years
# in ratio of 54.5%, 22.4%, 16.9%, 4.7% and 1.5%
                SWITCH: {
                    $month = $casedate[$generator] + $randmonth, last SWITCH if ($yearselector
<= 545);
                    $month = $casedate[$generator] + 11 + $randmonth, last SWITCH if
($yearselector <= 545+224);
                    $month = $casedate[$generator] + 23 + $randmonth, last SWITCH if
($yearselector <= 545+224+169);
                    $month = $casedate[$generator] + 35 + $randmonth, last SWITCH if
($yearselector <= 545+224+169+47);
                    $month = $casedate[$generator] + 47 + $randmonth;
                }
            } else {
                @casedate = ();
                $month = int(rand()*$yearspan) * 12 + int(rand()*12+1);
            }
            push @casedate, $month;    # Temp var to calculate year for next case of same strain
            $previous_strain = $strain;    # Temp var to compare with the next one
            $year = int($month / 12);
            $straincount[$clustersize][$strain][$year]++;
        }
    }
    # Store cumulative clustering, RTrans and their STDs in 2D arrays (e.g.
$clustering[$clustersize][$winsize])
    # At the end, of all the itterations, divide the totals in each cell by $itterations.
    print "Calculating statistics for clustersize:\n";
    for $clustersize ($minclustersize .. $maxclustersize){
        print "\t$clustersize\n";
        for $winsize (1 .. $winsizemax){    # Iterate through each window size
            @casecount = ();
            # Iterate through all possible windows for this window size
            for $window (1 .. $winsizemax - $winsize + 1){
                for $strain (1 .. (scalar(@{$straincount[$clustersize]}) -1) ){
                    for $year ($yearmin + $window - 1 .. $yearmin + $window-1 + $winsize - 1){
                        $straincount[$clustersize][$strain][$year] += 0; # Assign 0 if null
                        # Sum the cases for each strain over each year in the window
                        $casecount[$window]{$strain} +=
$straincount[$clustersize][$strain][$year];
                    }
                }
            }
#           $sumtotal = $sumuniques = $sumstrains = $sumclustering =  $sumrtrans = 0;
            $sumuniques = $sumclustering =  $sumrtrans = 0;
            @total = @strains = @uniques = ();
            for $window (1 .. $winsizemax - $winsize + 1){
                for $strain (keys %{$casecount[$window]}){
                    # If there are any of this strain in this window
                    if ($casecount[$window]{$strain} > 0){
                        # Sum the cases for each  strain over each window of this windowsize
                        $total[$window] += $casecount[$window]{$strain};
                        $strains[$window]++;
                        if ($casecount[$window]{$strain} == 1){
                            $uniques[$window]++;
                        }
                    }
                }
                $sumclustering += ($total[$window] - $uniques[$window]) / $total[$window] *
100;
                $sumrtrans += ($total[$window] - $strains[$window]) / $total[$window] * 100;
                $sumuniques += $uniques[$window];
#               $sumtotal += $total[$window];
#               $sumstrains += $strains[$window];
            }
            $windowcount = $winsizemax - $winsize + 1;
            $avclustering = $sumclustering / $windowcount;
            $avrtrans = $sumrtrans / $windowcount;
            $avuniques = $sumuniques / $windowcount;
#           $avtotal = $sumtotal / $windowcount;
#           $avstrains = $sumstrains / $windowcount;

            $sum_of_squares_clust = $sum_of_squares_rt = $sum_of_squares_u = 0;
            for $window (1 .. $windowcount){
                $sum_of_squares_clust += ( (($total[$window] - $uniques[$window]) /
$total[$window]) * 100 - $avclustering )**2;
                $sum_of_squares_rt += ( (($total[$window] - $strains[$window]) /
$total[$window]) * 100 - $avrtrans )**2;
                $sum_of_squares_u += ( $uniques[$window] - $avuniques )**2;
            }
            $SD_clust = sqrt( $sum_of_squares_clust / ($windowcount) );
```

```
             $SD_rt = sqrt( $sum_of_squares_rt / ($windowcount) );
             $SD_u = sqrt( $sum_of_squares_u / ($windowcount) );
             $clustering[$clustersize][$winsize] += $avclustering;
             $clusteringSD[$clustersize][$winsize] += $SD_clust;
             $rtrans[$clustersize][$winsize] += $avrtrans;
             $rtransSD[$clustersize][$winsize] += $SD_rt;
             $unique[$clustersize][$winsize] += $avuniques;
             $uniqueSD[$clustersize][$winsize] += $SD_u;
         }
     }
}

open OUTFILE_C, ">Cresult.txt" or die "Can't open data file 'Cresult.txt'";
open OUTFILE_RT, ">RTresult.txt" or die "Can't open data file 'RTresult.txt'";
open OUTFILE_U, ">Uresult.txt" or die "Can't open data file 'Uresult.txt'";
print OUTFILE_C "Windowsize";
print OUTFILE_RT "Windowsize";
print OUTFILE_U "Windowsize";
for $clustersize ($minclustersize .. $maxclustersize){
     print OUTFILE_C "\tAvClust_$clustersize\tSD_$clustersize";
     print OUTFILE_RT "\tAvTrans_$clustersize\tSD_$clustersize";
     print OUTFILE_U "\tAvUniques_$clustersize\tSD_$clustersize";
}
print OUTFILE_C "\n";
print OUTFILE_RT "\n";
print OUTFILE_U "\n";
for $winsize (1 .. $winsizemax){
     print OUTFILE_C "$winsize";
     print OUTFILE_RT "$winsize";
     print OUTFILE_U "$winsize";
     for $clustersize ($minclustersize .. $maxclustersize){
          $clustering[$clustersize][$winsize] = $clustering[$clustersize][$winsize]/$itterations;
          $clusteringSD[$clustersize][$winsize] =
$clusteringSD[$clustersize][$winsize]/$itterations;
          print OUTFILE_C
"\t$clustering[$clustersize][$winsize]\t$clusteringSD[$clustersize][$winsize]";
          $rtrans[$clustersize][$winsize] = $rtrans[$clustersize][$winsize]/$itterations;
          $rtransSD[$clustersize][$winsize] = $rtransSD[$clustersize][$winsize]/$itterations;
          print OUTFILE_RT "\t$rtrans[$clustersize][$winsize]\t$rtransSD[$clustersize][$winsize]";
          $unique[$clustersize][$winsize] = $unique[$clustersize][$winsize]/$itterations;
          $uniqueSD[$clustersize][$winsize] = $uniqueSD[$clustersize][$winsize]/$itterations;
          print OUTFILE_U "\t$unique[$clustersize][$winsize]\t$uniqueSD[$clustersize][$winsize]";
     }
     print OUTFILE_C "\n";
     print OUTFILE_RT "\n";
     print OUTFILE_U "\n";
}
close OUTFILE_C;
close OUTFILE_RT;
close OUTFILE_U;
```

6

Changing *Mycobacterium tuberculosis* population highlights clade-specific pathogenic characteristics.

GD van der Spuy, K Kremer, SL Ndabambi, N Beyers, R Dunbar, BJ Marais, PD van Helden and RM Warren

Tuberculosis (Accepted, September 2008)

Abstract

Mycobacterium tuberculosis strains can be classified into a number of major clades according to defined evolutionary markers. It is hypothesised that strains comprising these clades have evolved different properties which may influence a local strain population structure. To investigate this, we analysed the incidence of tuberculosis caused by the predominant clades (Beijing, Haarlem, LAM, Quebec and the Low-Copy Clade) found in a community within the Cape Town metropole in South Africa over a 12-year period. We found that, while the incidence of cases infected with strains of the Haarlem, LAM, Quebec and the Low-Copy Clades remained relatively stable, that of cases of the Beijing clade increased exponentially over time, with a doubling time of 4.86 years ($P=0.018$). This growth was exclusively attributable to drug-susceptible strains. Although drug-resistant Beijing cases remained constant in number, non-Beijing drug-resistant cases declined over time ($P=0.007$). Drug-susceptible Beijing-infected cases had a greater proportion of smear-positive sputa than their non-Beijing counterparts ($P=0.013$) and were less likely to be successfully treated (retreatment cases) *(*$P=0.026$*)*. Recent evidence suggests that these differences likely reflect enhanced pathogenicity rather than transmissibility. The rapid emergence of Beijing strains demonstrates adaptation to conditions within the study community and poses a grave challenge to future TB control.

Introduction

DNA fingerprinting has enabled the accurate classification of *Mycobacterium tuberculosis* strains, thereby facilitating their geo-temporal tracking. Comparative genotyping has shown that strains can be classified according to genotypic similarity [1]. Such classification schemes are hierarchical in nature, ranging from the clustering of apparently identical clones [2,3] up to the grouping of anciently related, major, global lineages (clades) [4]. Each of these levels in the classification hierarchy is informative in addressing different questions and has provided novel insights into the dynamics of the disease within host populations, both globally and locally. This has included: estimating the extent of recent transmission [2], defining the mechanism of recurrence [5,6], identifying outbreaks [7,8], demonstrating global phylogenies [4], as well as calculating the extent of laboratory cross-contamination [9].

Molecular epidemiological analysis of tuberculosis (TB) epidemics in different settings has shown that they are usually composites of a number of phylogenetically unrelated clades [10], each characterised by the inheritance of unique genomic markers (including insertions, deletions, single nucleotide polymorphisms, and expansion and contraction of repeat sequences)[11-13].

The accumulation of chromosomal mutations may also be associated with the evolution of clade specific phenotypes [11-13].as suggested by the observation that strains representative of three major global clades (Beijing, Haarlem and East-African-Indian) and *M. canetti* demonstrated different immunopathologies in the mouse infection model [14]. Accordingly, we hypothesise that the genetic diversity of the bacterial strains present within an epidemic setting may give rise to an array of pathogenic characteristics. Such a bacterial population would be subject to selective pressures which may give rise to phenotypic variations affecting host-pathogen interactions and, consequently, influencing the structure of the bacterial strain population within a particular host population [4,15]. However, the nature of any such changes in the bacterial strain population over time is currently unknown.

Recent studies have used molecular epidemiological techniques to interrogate longitudinal databases to demonstrate changes in the incidence of tuberculosis cases and rates of clustering, to identify groups and sub-groups at risk of contracting disease and to identify risk factors for both reactivation and the emergence of drug-resistant strains [16-19]. However, these studies failed to report temporal changes in the *M. tuberculosis* population structure. Thus, risk factors which could alter the *M. tuberculosis* strain population structure over time remain largely unknown, with the exception that immigration was the greatest factor defining the population structure in low incidence settings [20].

In this study we describe changes in the population structure of *M. tuberculosis* strains cultured from patients resident in an epidemiological field site in Cape Town [21], South Africa over a twelve-year period. We evaluated the potential influence of clinical, demographical and bacterial factors on the changes observed

Methods

Definition of a case: A case of tuberculosis was defined as a treatment episode having one of the following clinical outcomes: cured or successfully treated, failed, interrupted (provided such interruption exceeded two months), died, initial default, transferred out and unknown [21]. With the exception of 'death', standardised outcomes were not available for drug-resistant cases and could therefore not be reported. For this study, we analysed the sub-set of cases for which an RFLP was available. Each case was associated with a specific strain of *M. tuberculosis*, identified by IS*6110* RFLP as described below. All subsequent retreatment episodes were regarded as separate cases (disease episodes). Cases were defined as smear-positive if they had at least one positive (including scanty) sputum smear.

Study population: According to the National TB Program guidelines, new cases are routinely investigated by sputum smear, while retreatment cases are routinely examined by sputum smear and culture. However, as part of an ongoing molecular epidemiological study, an attempt was made to collect sputum samples for culture at diagnosis from all new and retreatment tuberculosis patients who attended primary health care clinics and who were resident in an epidemiological field site in Cape Town, Western Cape, South Africa [22] during the period January 1993 to December 2004. Census data provided by Statistics South Africa from 1996 and 2001 showed that the population had remained stable both in terms of number (±36 300) and age distribution (data not shown). At least 99% of patients were of indigenous South African origin (non-migrant population) (data not shown). In this community, an average of 320 new bacteriologically-confirmed (having a smear- or culture-positive specimen) adult cases per 100 000 population were reported for the years 1993 to 1998 [23].

Sputum smear microscopy (fluorescent staining) and/or culture in BACTEC 460, MGIT 960 (Becton Dickinson, Franklin Lakes, NJ USA) or on Löwenstein-Jensen media was done by the National Health Laboratory Service (routine laboratory services to the primary health care clinics) or our laboratories at Stellenbosch University, Faculty of Health Sciences, to confirm the presence of *M. tuberculosis*. Clinical and demographic data including previous history of tuberculosis, gender and age, smear positivity and drug-susceptibility test results (if requested), were recorded in a database. Chest radiography was not done routinely by the national program and was therefore not included in this study. HIV testing was not routinely done in the initial years of this study although a recent survey of 366 new adult smear positive tuberculosis cases (2000 to 2002) in this epidemiological field site showed that 10% were HIV positive.

Drug-susceptibility testing was done by the National Health Laboratory Service, using the indirect proportion method on Löwenstein-Jensen medium containing critical concentrations of 0.2 μg/ml Isoniazid and 30 μg/ml Rifampicin. In this study, drug-resistance was defined as resistance to either Isoniazid or Rifampicin or both (MDR-tuberculosis).

All data was captured and stored in a database. To ensure confidentiality, all data was unlinked from the patients' names. This study forms part of a larger, long-term molecular epidemiological project which has been approved by the ethics committee of the Faculty of Health Sciences of Stellenbosch University.

DNA fingerprinting: Sputum isolates were collected from all tuberculosis patients and cultured on MGIT and/or Löwenstein-Jensen media. DNA was extracted as previously described [24]. Each isolate was classified by IS*6110* DNA fingerprinting [25,26] and spoligotyping [27] using internationally standardized protocols. Strains were identified according to distinct IS*6110* banding patterns using Gelcompar II (Applied Maths, Sint-Martens-Latem, Belgium) as previously described [28] and were subsequently grouped into evolutionary clades which were classified based on their spoligotype signatures [29,30]. Within the study community, strains having fewer than 6 IS*6110* bands (low-copy clade) comprise a single lineage as defined by IS*6110* (as previously described [30]) and were therefore regarded as a single clade. Sub-lineages of the Beijing clade were identified as previously described [11]. The proportion of cases arising as a result of on-going transmission was calculated according to the formula (clustered cases – index cases)/total cases [2], where transmission chains were defined as a series of cases having isolates with identical IS*6110* DNA fingerprints (identified using Gelcompar II), with inter-case intervals of less than 2 years and each transmission chain assumed to be initiated by a single index case. A transmission chain unique case was defined as one having no other cases with the identical strain occurring within 2 years either side. Cases classified as retreatment after failure/interruption were regarded as a continuation of the preceding episode and were excluded from this calculation.

Statistical analysis: Cases from each clade were compared in terms of: age, sex, prior history of tuberculosis, treatment outcome, duration of treatment and smear positivity. As the treatment regimen differs between new and re-treatment cases, these groups were analysed separately with respect to treatment duration and outcome. Clades were compared in terms of: number of strains, transmission chain size, proportion of ongoing transmission, frequency of drug-resistance. This data was analysed to identify bacterial, demographic and/or clinical risk factors which may influence temporal changes in each clade's contribution to the epidemic.

Statistical analysis was done using Graphpad Prism 5 (La Jolla, CA USA). A p-value of less than 0.05 was considered as statistically significant.

Non-linear regression of the annual incidence of cases from the various clades was plotted using the exponential equation: $Y = Y_0 * e^{(k*X)}$, where X is the study year, Y_0 is the incidence in year one of the study and k is the rate constant, expressed in inverse years. The doubling time is calculated as *ln2/k*. It was assumed that the growth (either positive or negative) in the incidence of cases infected with a particular strain population could be described by an exponential equation.

The reproductive rates for each clade (divided into drug-susceptible and resistant cases) were calculated as 1 plus the rate constant (*k*), which was derived from the exponential non-linear regression plot of annual incidence.

Results

During the period 1993 to 2004, a total of 2 727 cases of tuberculosis were recorded (notified) from 2 150 patients, of whom 425 had multiple episodes (2 to 5) of disease. A subset of 1 921 (70%) cases had a *M. tuberculosis* culture and an IS*6110* RFLP fingerprint with >98% classified as pulmonary TB. There was no difference in the proportion of new to retreatment cases between those having an RFLP and those that did not. Cases included in the analysis represented 546 strains (distinct IS*6110* banding patterns) which could be grouped into 8 clades of which 5 predominated: Beijing, Latin American – Mediterranean (LAM), Low Copy-number (LC), Haarlem and Quebec (Table S1). Together these 5 clades comprised 84% of the strain population.

Non-linear regression analysis of the annual case numbers of tuberculosis cases according to their respective clades showed that the numbers of cases in the Haarlem, LAM, LC and Quebec clades did not change significantly over the study period (Figure 1). However, the number of cases with a Beijing genotype strain increased significantly as a function of time, with an estimated doubling time of 4.86 years (95% CI, 3.55 to 7.72).

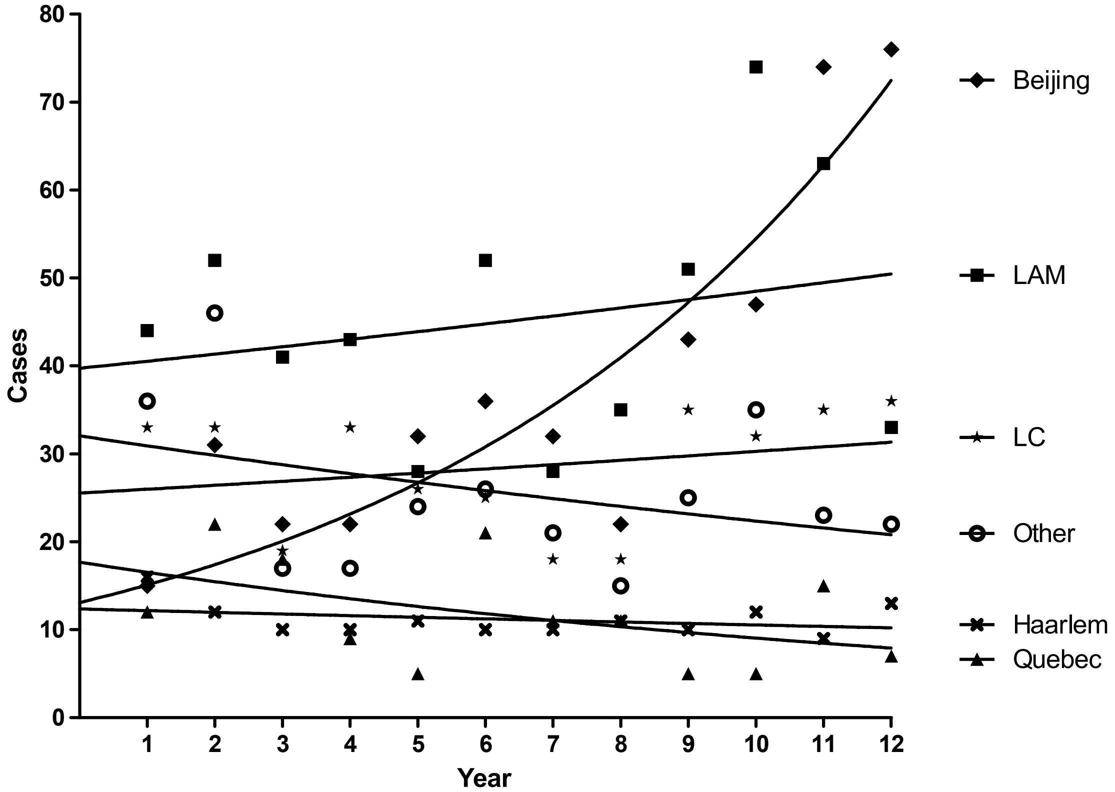

Figure 1. Annual number of cases belonging to the 5 major *M. tuberculosis* clades with the remaining cases grouped as 'Other'. Calculated doubling times in years were: Beijing (♦) = 4.86, LAM (■) = 34.79, LC (★) = 40.71, Haarlem (×) = -43.43, Quebec (▲) = -10.35, Other (○) = -19.25.

To determine whether the increase in the incidence of cases with a Beijing strain was driven by drug-resistance, the non-linear regression analysis was repeated, separating cases classified by routine testing as drug-resistant and drug-susceptible. This analysis showed that the increase in incidence was driven exclusively by drug-susceptible Beijing genotype strains. For this clade, we estimated doubling times of 3.89 years (95% CI, 2.85 to 6.15) and 252.0 years (95% CI, 11.26 to +infinity) for drug-susceptible and drug-resistant cases, respectively (Figure 2 and Table 1). The doubling time for drug-susceptible cases from the remaining clades (non-Beijing) was estimated to be 76.56 years (95% CI, 12.75 to +infinity). Among the drug-resistant non-Beijing-infected cases, the incidence of all the clades declined over time although this was only significant in the case of the LC clade which decreased with a halving time of 4.48 years (95% CI, 2.63 to 14.97) (Table S1). We found that the Beijing clade had a significantly higher proportion of drug-resistant cases than non-Beijing clades (Fisher's exact test OR, 2.24; 95% CI, 1.63 to 3.07; *P* < 0.001) and that, within the drug-resistant subset, Beijing-infected cases were significantly more likely to be due to transmission (Fisher's exact test OR, 8.66; 95% CI, 3.88 to 19.34; *P* < 0.001) (Table 2).

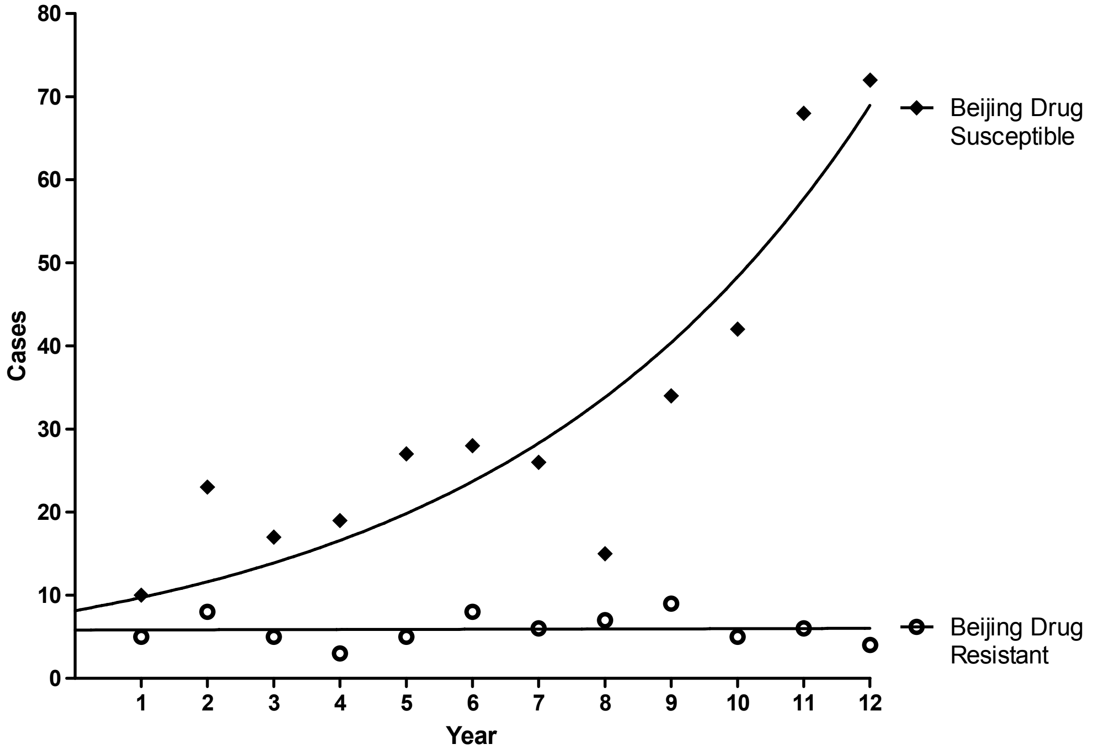

Figure 2. Annual number of drug-susceptible (♦) and drug-resistant (○) cases of the Beijing clade of *M. tuberculosis*. Calculated doubling times in years were: Drug susceptible = 3.89, Drug resistant = 252.0.

We have previously reported that the local Beijing strain population is dominated by sub-lineage 7 (72%) [11]. To determine whether the increase in the incidence of cases with drug-susceptible Beijing genotype strains was dependent on the phylogenetic sub-lineage, the non-linear regression analysis was repeated for cases with strains belonging to sub-lineages 1 to 6 combined and sub-lineage 7. Doubling times for these two groups were found to be 3.69 years (95% CI, 2.45 to 7.45) and 3.94 years (95% CI, 2.80 to 6.66) respectively. These growth rates are not significantly different (extra sum-of-squares F-test) and therefore cases with a drug-susceptible Beijing strain were regarded as a single population for this study.

Table 1. Clinical and demographic analysis of Beijing *vs.* non-Beijing drug-susceptible and drug-resistant cases.

	Drug-susceptible n (%)			Drug-resistant n (%)		
	Beijing	Non-Beijing	Odds Ratio[a] (95% CI) *P* value	Beijing	Non-Beijing	Odds Ratio[a] (95% CI) *P* value
Cases	381	1356		71	113	
Reproductive rate [k + 1] (1/years)	1.178	1.009	*P* = 0.018[b]	1.003	0.873	*P* = 0.007[b]
Doubling Time (years) (95% CI)	3.89 (2.85 to 6.15)	76.56 (12.75 to +∞)		252.00 (11.26 to +∞)	-5.47 (-11.78 to -3.57)	
Male	239 (62.7)	774 (57.1)	1.27 (1.01 to 1.61) *P* = 0.045	35 (49.3)	73 (64.6)	0.53 (0.29 to 0.98) *P* = 0.046
Smear positive	301 (79.0)	981 (72.3)	1.45 (1.08 to 1.96) *P* = 0.013	51 (71.8)	83 (73.7)	NS
Median Age (years)	33.6	35.2	NS[c]	37.0	32.9	NS[c]
New Cases[d]	225 (59.1)	810 (59.7)	NS	23 (32.4)	30 (26.5)	NS
Successfully treated	177 (78.7)	595 (73.5)	NS			
Treatment Interrupted	21 (9.3)	94 (11.6)	NS			
Treatment Failed	2 (0.9)	10 (1.2)	NS			
Died	3 (1.3)	25 (3.1)	NS	2 (8.7)	2 (6.7)	NS
Median Treatment Period (days)	186	190	NS[c]			
Retreatment Cases[d]	145	522		48	82	
Successfully treated	76 (52.4)	330 (62.2)	0.65 (0.45 to 0.94) *P* = 0.026			
Treatment Interrupted	34 (23.4)	97 (18.6)	NS			
Treatment Failed	0 (0.0)	8 (1.5)	NS			
Died	5 (3.4)	18 (3.4)	NS	6 (12.5)	6 (7.3)	NS
Median Treatment Period (days)	222	225	NS[c]			

NS = Not significant

With the exception of 'death', standard outcome definitions were not available for drug-resistant cases and results could therefore not be reported. Results for other, non-listed outcomes are also not reported. Median treatment period could not be reported for drug-resistant cases as they are not necessarily exclusively treated at the community clinic.

[a] Statistical analysis done using Fishers exact test except where otherwise specified

[b] Extra sum-of-squares F-test; [c] Mann-Whitney U test; [d] 11 Beijing and 25 non-Beijing cases could not be classified as either 'New' or 'Retreatment'

Table 2. Molecular epidemiological analysis of Beijing *vs.* non-Beijing drug-susceptible and drug-resistant cases.

	Drug-susceptible			Drug-resistant		
	Beijing	Non-Beijing	Odds Ratio[a] (95% CI) p-value	Beijing	Non-Beijing	Odds Ratio[a] (95% CI) p-value
Cases[b]	352	995		63	63	
Strains[c]	76	429		17	48	
Transmission chains	95	543		17	48	
Maximum transmission chain clustersize	128	42		37	4	
Unique cases	64	399		12	39	
Clustered transmission chains (clusters)	31	144		5	9	
Clustered transmission chain cases (%)	288 (81.8)	596 (59.9)	3.01 (2.23 to 4.06) *P* < 0.001	51 (81.0)	24 (38.1)	6.91 (3.08 to 15.51) *P* < 0.001
Transmitted cases[d]	257	452	3.25 (2.49 to 4.24) *P* < 0.001	46	15	8.66 (3.88 to 19.34) *P* < 0.001
Recent Transmission (%)[e]	73.0	45.4		73.0	23.8	

Clustering data and analysis is derived from transmission chains defined as cases having identical strains with inter-strain intervals of <2 years.

[a] Statistical analysis done using Fishers exact test

[b] Case numbers differ from those in table 1 for two reasons: 1) Cases classified as retreatment after failure/interruption were excluded. 2) Data for the Non-Beijing groups excluded cases from the LC clade (n=302) due to the poor resolution of IS*6110* with fewer than 6 copies. Seven patients acquired resistance and were therefore counted in both the sensitive and resistant categories.

[c] Some strains fall into both drug-susceptible and –resistant groups.

[d] Cases in clusters – Number of transmission chains (corresponding to the index cases)

[e] (Cases in clusters – Number of transmission chains) / Total number of cases (24)

In order to identify factors which may be associated with the observed increase in cases with drug-susceptible Beijing genotype strains, we compared clinical and /or demographic parameters were compared between cases with Beijing and non-Beijing strains. Table 1 shows the comparisons for both drug-susceptible and drug-resistant strains. We found that cases with drug-susceptible Beijing strains had a significantly greater proportion of sputum smear-positive disease than cases with drug susceptible strains of other genotypes (Fisher's exact test OR, 1.45; 95% CI, 1.08 to 1.96; P = 0.013). There was no difference between the drug-resistant groups or between drug-resistant and drug-susceptible cases of the Beijing clade. Among the drug-susceptible cases, we found a weak association between male gender and Beijing genotype (Fisher's exact test OR, 1.27; 95% CI, 1.01 to 1.61; P = 0.045). Conversely, cases infected with drug-resistant Beijing strains were less likely to be male than drug-resistant non-Beijing-infected cases (Fisher's exact test OR, 0.53; 95% CI, 0.29 to 0.98; P = 0.046). Within the Beijing clade, drug-susceptible cases were more likely to be male than drug resistant cases (Fisher's exact test OR, 1.74; 95% CI, 1.05 to 2.90, P = 0.035). With the exception of a slightly lower rate of successful treatment for drug-susceptible re-treatment cases of the Beijing clade (Fisher's exact test OR, 0.65; 95% CI, 0.45 to 0.94, P = 0.026), no differences between Beijing and non-Beijing-infected cases could be found in terms of episode outcome, the proportion of new *vs.* retreatment cases (Fisher's exact test), and median duration of treatment or median age at diagnosis (Mann-Whitney U test). There was also no difference between the rates of growth of new *vs.* retreatment cases of drug-susceptible Beijing-infected cases (Extra sum-of-squares F-test).

To determine whether clade specific pathogenicity factors (ability to spread and cause disease) were associated with the increase in incidence of cases with Beijing genotype strains, the molecular epidemiological characteristics of the Beijing and non-Beijing strains were analysed. Both drug-susceptible and resistant Beijing strains were more often clustered than non-Beijing strains, (Fisher's exact test OR, 3.01; 95% CI, 2.23 to 4.06; $P < 0.001$, and OR, 6.91; 95% CI, 3.08 to 15.51; $P < 0.001$ respectively) (Table 2). This implies a higher rate of transmission for both groups and suggests that Beijing genotype strains are more likely to spread and cause disease that non-Beijing strains.

Discussion

In this 12-year, longitudinal study of an urban community with a high incidence of TB, we have analysed the temporal dynamics of the prevalent *M. tuberculosis* clades. We demonstrate that the bacterial population structure has changed significantly over time. This was largely due to an increase in the incidence of cases with Beijing genotype strains. A previous worldwide survey of the Beijing clade also found that these strains were emerging, both in South Africa and other countries, including Argentina, Cuba, Malawi, Vietnam, the countries of the former Soviet Union and parts of Western Europe [31]. In some of these countries the emergence of this genotype was highly associated with drug resistance. In our longitudinal study, the superior rate of increase of this clade was exclusively due to drug-susceptible strains and was common to all sub-lineages tested. Hence, the acquisition of drug-resistance and the subsequent transmission of resistant strains of *M. tuberculosis* cannot account for the observed rise in the number of Beijing genotype cases. We cannot, however, exclude the possibility of increased drug tolerance in Beijing strains, although this might be expected to lead to an increased rate of treatment failure in this clade, which was not observed (data not shown). Accordingly, we propose that the drug-susceptible Beijing strains represent a highly pathogenic genotype when compared to drug-susceptible strains from the other clades present in the study community. Our analysis has also shown, by a reduction in the rate of generation of new cases, that the evolution of drug-resistance has had a strong influence on the pathogenicity of Beijing genotype strains, thereby confirming that the mutations conferring resistance incur a significant fitness cost [32]. In spite of this, however, these drug-resistant Beijing strains have remained as pathogenic as most drug-susceptible non-Beijing strains. In addition, from the fact that the drug-susceptible and drug-resistant Beijing strain groups show the same degree of clustering, it is apparent that the attenuation of the latter is in the rate and/or likelihood of their progression to active disease rather than their ability to transmit. A similar loss of fitness, following the acquisition of drug-resistance, was seen for strains of the LC and LAM clades, although this was not statistically significant for the latter. Case numbers were too low in the other clades to make any meaningful observations. We found that drug-resistant Beijing strains showed a higher level of transmissibility when compared to drug-resistant non-Beijing strains, supporting the notion of the superior overall fitness of the former despite the resistance-derived fitness loss mentioned above (Table 2). It should be noted that estimates of transmission assume equivalence in the marker's discriminatory power for the different clades. Thus, if the IS*6110* RFLP pattern of the Beijing clade strains were more stable than those of other strains, as has been suggested by van Soolingen, *et al.*[33], it would result in an apparently higher rate of transmission for the former.

Our attempt to identify factors which could explain the increased pathogenicity of the Beijing clade showed that higher levels of bacilli were present in the sputum specimens of patients infected with these strains. However, this difference was small and is therefore unlikely to explain fully the magnitude of the observed growth rate.

The lower rate of successful treatment in drug-susceptible Beijing-infected, retreatment cases may contribute to the success of this clade, but this observation is probably too small to fully account for it's success. Furthermore, this was not observed for new cases. Therefore, we propose that Beijing genotype strains have evolved properties which enhance their propensity to cause disease following infection compared to other clades. This hypothesis is supported by results from studies in BALB/c mice, co-infected with Beijing (HN878) and non-Beijing (CDC1551) strains, which have demonstrated differential growth rates with the Beijing strain growing more rapidly [34]. Further support comes from the recent observation that, while Beijing-type strains transmit equally as well as strains of other clades, they have a greater propensity to cause active disease after infection [35]. This may be explained by the Beijing strains' reported ability to modulate the host immune system in favour of a Th2 response while non-Beijing strains tend to induce a Th1 response [36]. Reed *et al.* determined that an important factor in this ability is the presence in Beijing strains of an intact *pks1-15* gene, responsible for the production of phenolic glycolipids, which inhibit the release of pro-inflammatory factors by monocyte-derived macrophages [37]. Subsequently, it has been shown that, while the *pks1-15* gene product plays a role in host immune modulation, it is neither required nor solely sufficient for hyper-virulence [38]. In addition, Theus *et al.* reported that virulent Beijing strains show increased growth in macrophages and inhibition of TNF-□ release [39].

It has also been suggested that the success of Beijing genotype strains is due to their ability to evade the protective effect of BCG vaccination [40]. It is therefore tempting to speculate that the same may apply to the relative success of the Beijing clade in high TB incidence communities where many, if not most, of the residents have been naturally vaccinated by infection. However, our attempt to substantiate this failed to demonstrate that Beijing genotype strains were more likely to cause re-infection after a previous episode of disease (data not shown). The possibility must be acknowledged, however, that recent, prior disease may (at least in the short-term) predispose the patient to susceptibility towards re-infection rather than providing protection against re-infection [41]. In addition, the extremely high rate of infection may itself militate against any attempt to demonstrate this phenomenon in this community.

We conclude that the Beijing clade is particularly well adapted to the conditions prevalent in this high incidence community. The emergence of Beijing genotype strains with increased pathogenicity may have important implications for the Tuberculosis Control Program. Early diagnosis and effective treatment is essential to curb the spread of these strains. In addition, contact tracing may have added value, although this poses great challenges in TB endemic regions. Furthermore, it is important to ensure that future vaccines provide adequate protection against these strains. The observation that drug-resistant Beijing strains retain a higher level of pathogenicity/fitness is a cause for concern and implies that novel intervention strategies are necessary to control the spread of drug-resistance. This can be achieved by the development and implementation of rapid diagnostics, provision of appropriate therapy, ensuring treatment adherence and intensified screening of contacts. However, in order for diagnosis and treatment to be effective it is essential that communities are educated to improve health seeking behaviour and thus minimise diagnostic delay and the opportunity for transmission.

Acknowledgements

The European Commission 6th framework programme on research technological development demonstration (project no. 037919) are thanked for financially supporting this project. The authors wish to thank Mrs E. Engelke, Mrs S. Carlini and Mrs M. De Kock, Ms K. Painczyk and Ms W. Brittle for their technical assistance and Dr Martin Kidd for his assistance with the statistical analysis of the data. We also wish to thank Department of Health, City of Cape Town for permission to do research in the clinics and aknowledge our indebtedness to the residents of the epidemiological field site.

Reference List

1. **Hermans PW, Messadi F, Guebrexabher H, van Soolingen D, de Haas PE, Heersma H, de Neeling H, Ayoub A, Portaels F, Frommel D.** Analysis of the population structure of Mycobacterium tuberculosis in Ethiopia, Tunisia, and The Netherlands: usefulness of DNA typing for global tuberculosis epidemiology.*J Infect Dis* 1995;**171**:1504-1513.

2. **Small PM, Hopewell PC, Singh SP, Paz A, Parsonnet J, Ruston DC, Schecter GF, Daley CL, Schoolnik GK.** The epidemiology of tuberculosis in San Francisco. A population-based study using conventional and molecular methods. *N Engl J Med* 1994;**330**:1703-1709.

3. **Alland D, Kalkut GE, Moss AR, McAdam RA, Hahn JA, Bosworth W, Drucker E, Bloom BR.** Transmission of tuberculosis in New York City. An analysis by DNA fingerprinting and conventional epidemiologic methods. *N Engl J Med* 1994;**330**:1710-1716.

4. **Gagneux S, DeRiemer K, Van T, Kato-Maeda M, de Jong BC, Narayanan S, Nicol M, Niemann S, Kremer K, Gutierrez MC, Hilty M, Hopewell PC, Small PM.** Variable host-pathogen compatibility in Mycobacterium tuberculosis. *Proc Natl Acad Sci U S A* 2006;**103**:2869-2873.

5. **Dwyer B, Jackson K, Raios K, Sievers A, Wilshire E, Ross B.** DNA restriction fragment analysis to define an extended cluster of tuberculosis in homeless men and their associates. *J Infect Dis* 1993;**167**:490-494.

6. **van Rie A, Warren R, Richardson M, Victor TC, Gie RP, Enarson DA, Beyers N, van Helden PD.** Exogenous reinfection as a cause of recurrent tuberculosis after curative treatment. *N Engl J Med* 1999;**341**:1174-1179.

7. **Hutton MD, Cauthen GM, Bloch AB.** Results of a 29-state survey of tuberculosis in nursing homes and correctional facilities. *Public Health Rep* 1993;**108**:305-314.

8. **Valway SE, Sanchez MP, Shinnick TF, Orme I, Agerton T, Hoy D, Jones JS, Westmoreland H, Onorato IM.** An outbreak involving extensive transmission of a virulent strain of Mycobacterium tuberculosis. *N Engl J Med* 1998;**338**:633-639.

9. **Small PM, McClenny NB, Singh SP, Schoolnik GK, Tompkins LS, Mickelsen PA.** Molecular strain typing of Mycobacterium tuberculosis to confirm cross-contamination in the mycobacteriology laboratory and modification of procedures to minimize occurrence of false-positive cultures.*J Clin Microbiol* 1993;**31**:1677-1682.

10. **Warren R, Richardson M, van der SG, Victor T, Sampson S, Beyers N, van Helden P.** DNA fingerprinting and molecular epidemiology of tuberculosis: use and interpretation in an epidemic setting. *Electrophoresis* 1999;**20**:1807-1812.

11. **Hanekom M, van der Spuy GD, Streicher E, Ndabambi SL, McEvoy CR, Kidd M, Beyers N, Victor TC, van Helden PD, Warren RM.** A recently evolved sublineage of the Mycobacterium tuberculosis Beijing strain family is associated with an increased ability to spread and cause disease.*J Clin Microbiol* 2007;**45**:1483-1490.

12. **Supply P, Mazars E, Lesjean S, Vincent V, Gicquel B, Locht C.** Variable human minisatellite-like regions in the Mycobacterium tuberculosis genome. *Mol Microbiol* 2000;**36**:762-771.

13. **Tsolaki AG, Hirsh AE, DeRiemer K, Enciso JA, Wong MZ, Hannan M, Goguet de la Salmoniere YO, Aman K, Kato-Maeda M, Small PM.** Functional and evolutionary genomics of Mycobacterium tuberculosis: insights from genomic deletions in 100 strains. *Proc Natl Acad Sci U S A* 2004;**101**:4865-4870.

14. **Lopez B, Aguilar D, Orozco H, Burger M, Espitia C, Ritacco V, Barrera L, Kremer K, Hernandez-Pando R, Huygen K, van Soolingen D.** A marked difference in pathogenesis and immune response induced by different Mycobacterium tuberculosis genotypes. *Clin Exp Immunol* 2003;**133**:30-37.

15. **Hanekom M, van der Spuy GD, Gey van Pittius NC, McEvoy CR, Ndabambi SL, Victor TC, Hoal EG, van Helden PD, Warren RM.** Evidence that the spread of Mycobacterium tuberculosis strains with the Beijing genotype is human population dependent.*J Clin Microbiol* 2007;**45**:2263-2266.

16. **Jasmer RM, Hahn JA, Small PM, Daley CL, Behr MA, Moss AR, Creasman JM, Schecter GF, Paz EA, Hopewell PC.** A molecular epidemiologic analysis of tuberculosis trends in San Francisco, 1991-1997. *Ann Intern Med* 1999;**130**:971-978.

17. **Geng E, Kreiswirth B, Driver C, Li J, Burzynski J, DellaLatta P, LaPaz A, Schluger NW.** Changes in the transmission of tuberculosis in New York City from 1990 to 1999. *N Engl J Med* 2002;**346**:1453-1458.

18. **Cruz-Ferro E, Fernandez-Nogueira E.** Epidemiology of tuberculosis in Galicia, Spain, 1996-2005. *Int J Tuberc Lung Dis* 2007;**11**:1073-1079.

19. **Lopez-Calleja AI, Lezcano MA, Vitoria MA, Iglesias MJ, Cebollada A, Lafoz C, Gavin P, Aristimuno L, Revillo MJ, Martin C, Samper S.** Genotyping of Mycobacterium tuberculosis

over two periods: a changing scenario for tuberculosis transmission. *Int J Tuberc Lung Dis* 2007;**11**:1080-1086.

20. **Borgdorff MW, Behr MA, Nagelkerke NJ, Hopewell PC, Small PM.** Transmission of tuberculosis in San Francisco and its association with immigration and ethnicity. *Int J Tuberc Lung Dis* 2000;**4**:287-294.

21. **Verver S, Warren RM, Munch Z, Vynnycky E, van Helden PD, Richardson M, van der Spuy GD, Enarson DA, Borgdorff MW, Behr MA, Beyers N.** Transmission of tuberculosis in a high incidence urban community in South Africa. *Int J Epidemiol* 2004;**33**:351-357.

22. **Beyers N, Gie RP, Zietsman HL, Kunneke M, Hauman J, Tatley M, Donald PR.** The use of a geographical information system (GIS) to evaluate the distribution of tuberculosis in a high-incidence community. *S Afr Med J* 1996;**86**:40-1, 44.

23. **Verver S, Warren RM, Munch Z, Richardson M, van der Spuy GD, Borgdorff MW, Behr MA, Beyers N, van Helden PD.** Proportion of tuberculosis transmission that takes place in households in a high-incidence area. *Lancet* 2004;**363**:212-214.

24. **Warren R, de KM, Engelke E, Myburgh R, Gey van PN, Victor T, van Helden P.** Safe Mycobacterium tuberculosis DNA extraction method that does not compromise integrity. *J Clin Microbiol* 2006;**44**:254-256.

25. **van Embden JD, Cave MD, Crawford JT, Dale JW, Eisenach KD, Gicquel B, Hermans P, Martin C, McAdam R, Shinnick TM.** Strain identification of Mycobacterium tuberculosis by DNA fingerprinting: recommendations for a standardized methodology. *J Clin Microbiol* 1993;**31**:406-409.

26. **Warren RM, Richardson M, Sampson SL, van der Spuy GD, Bourn W, Hauman JH, Heersma H, Hide W, Beyers N, van Helden PD.** Molecular evolution of Mycobacterium tuberculosis: phylogenetic reconstruction of clonal expansion. *Tuberculosis (Edinb)* 2001;**81**:291-302.

27. **Kamerbeek J, Schouls L, Kolk A, van Agterveld M, van Soolingen D, Kuijper S, Bunschoten A, Molhuizen H, Shaw R, Goyal M, van Embden J.** Simultaneous detection and strain differentiation of Mycobacterium tuberculosis for diagnosis and epidemiology. *J Clin Microbiol* 1997;**35**:907-914.

28. **van der Spuy GD, Warren RM, Richardson M, Beyers N, Behr MA, van Helden PD.** Use of genetic distance as a measure of ongoing transmission of Mycobacterium tuberculosis. *J Clin Microbiol* 2003;**41**:5640-5644.

29. **Streicher EM, Victor TC, van der SG, Sola C, Rastogi N, van Helden PD, Warren RM.** Spoligotype signatures in the Mycobacterium tuberculosis complex. *J Clin Microbiol* 2007;**45**:237-240.

30. **Warren RM, Victor TC, Streicher EM, Richardson M, van der Spuy GD, Johnson R, Chihota VN, Locht C, Supply P, van Helden PD.** Clonal expansion of a globally disseminated lineage of Mycobacterium tuberculosis with low IS6110 copy numbers. *J Clin Microbiol* 2004;**42**:5774-5782.

31. **Glynn JR, Kremer K, Borgdorff MW, Rodriguez MP, van Sooligen D.** Beijing/W genotype Mycobacterium tuberculosis and drug resistance. *Emerg Infect Dis* 2006;**12**:736-743.

32. **Billington OJ, McHugh TD, Gillespie SH.** Physiological cost of rifampin resistance induced in vitro in Mycobacterium tuberculosis. *Antimicrob Agents Chemother* 1999;**43**:1866-1869.

33. **van Soolingen D, Qian L, de Haas PE, Douglas JT, Traore H, Portaels F, Qing HZ, Enkhsaikan D, Nymadawa P, van Embden JD.** Predominance of a single genotype of Mycobacterium tuberculosis in countries of east Asia.*J Clin Microbiol* 1995;**33**:3234-3238.

34. **Barczak AK, Domenech P, Boshoff HI, Reed MB, Manca C, Kaplan G, Barry CE, III.** In Vivo Phenotypic Dominance in Mouse Mixed Infections with Mycobacterium tuberculosis Clinical Isolates.*J Infect Dis* 2005;**192**:600-606.

35. **de Jong BC, Hill PC, Aiken A, Awine T, Antonio M, Adetifa IM, Jackson-Sillah DJ, Fox A, Deriemer K, Gagneux S, Borgdorff MW, McAdam KP, Corrah T, Small PM, Adegbola RA.** Progression to Active Tuberculosis, but Not Transmission, Varies by Mycobacterium tuberculosis Lineage in The Gambia.*J Infect Dis* 2008;**198**:1037-1043.

36. **Manca C, Tsenova L, Bergtold A, Freeman S, Tovey M, Musser JM, Barry CE, III, Freedman VH, Kaplan G.** Virulence of a Mycobacterium tuberculosis clinical isolate in mice is determined by failure to induce Th1 type immunity and is associated with induction of IFN-alpha /beta. *Proc Natl Acad Sci U S A* 2001;**98**:5752-5757.

37. **Reed MB, Domenech P, Manca C, Su H, Barczak AK, Kreiswirth BN, Kaplan G, Barry CE, III.** A glycolipid of hypervirulent tuberculosis strains that inhibits the innate immune response. *Nature* 2004;**431**:84-87.

38. **Sinsimer D, Huet G, Manca C, Tsenova L, Koo MS, Kurepina N, Kana B, Mathema B, Marras SA, Kreiswirth BN, Guilhot C, Kaplan G.** The phenolic glycolipid of Mycobacterium tuberculosis differentially modulates the early host cytokine response but does not in itself confer hypervirulence. *Infect Immun* 2008.

39. **Theus S, Eisenach K, Fomukong N, Silver RF, Cave MD.** Beijing family Mycobacterium tuberculosis strains differ in their intracellular growth in THP-1 macrophages. *Int J Tuberc Lung Dis* 2007;**11**:1087-1093.

40. **van Soolingen D, Qian L, de Haas PE, Douglas JT, Traore H, Portaels F, Qing HZ, Enkhsaikan D, Nymadawa P, van Embden JD.** Predominance of a single genotype of Mycobacterium tuberculosis in countries of east Asia.*J Clin Microbiol* 1995;**33**:3234-3238.

41. **Verver S, Warren RM, Beyers N, Richardson M, van der Spuy GD, Borgdorff MW, Enarson DA, Behr MA, van Helden PD.** Rate of reinfection tuberculosis after successful treatment is higher than rate of new tuberculosis. *Am J Respir Crit Care Med* 2005;**171**:1430-1435.

Table S1. Clinical and demographic analysis by clade of drug-susceptible and drug-resistant cases.

Clade	Beijing n (%)		Haarlem n (%)		LAM n (%)		LC n (%)		Quebec n (%)		Other n (%)	
	S	R	S	R	S	R	S	R	S	R	S	R
Cases	381 (19.8)	71 (3.7)	130 (6.8)	4 (0.2)	513 (26.7)	31 (1.6)	299 (15.6)	44 (2.3)	131 (6.8)	10 (0.5)	283 (14.7)	24 (1.3)
Doubling Time (years) 95% CI	3.9 2.89 to 6.19	252.0 11.3 to +∞	-82.6 -∞ to -18.5	-1.4 -∞ to -0.6	25.5 7.9 to +∞	-8.9 -∞ to -3.4	13.0 6.3 to +∞	-4.5 -15.0 to -2.6	-12.7 -∞ to -4.9	-3.2 -∞ to -1.4	-21.7 -∞ to -6.8	-10.1 -∞ to -3.7
Reproductive rate {K + 1} (1/years)	1.178	1.003	0.992	0.507	1.027	0.924	1.053	0.847	0.947	0.782	0.97	0.933
Male	239 (62.7)	35 (49.3)	60 (46.2)	4 (100)	311 (60.6)	22 (71.0)	169 (56.5)	23 (52.3)	65 (49.6)	7 (70.0)	169 (59.7)	17 (70.8)
Smear positive	301 (79.0)	51 (71.8)	96 (73.9)	4 (100)	370 (72.1)	22 (71.0)	227 (75.9)	30 (68.2)	89 (67.9)	8 (80.0)	199 (70.3)	19 (79.2)
Median Age (years)	33.6	37.0	35.7	32.9	34.2	33.5	35.8	28.4	35.7	36.1	36.7	33.1
New Cases	225 (59.1)	23 (32.4)	79 (60.8)	0 (0.0)	331 (64.5)	6 (19.4)	165 (55.2)	17 (38.6)	80 (61.1)	4 (40.0)	155 (54.8)	3 (12.5)
Successfully treated	177 (78.7)	-	62 (78.5)	-	252 (76.1)	-	112 (67.9)	-	57 (71.3)	-	112 (72.3)	-
Treatment Interrupted	21 (9.3)	-	8 (10.1)	-	30 (9.1)	-	26 (15.8)	-	15 (18.8)	-	15 (9.7)	-
Treatment Failed	2 (0.9)	-	0 (0.0)	-	4 (1.2)	-	2 (1.2)	-	1 (1.3)	-	3 (1.9)	-
Died	3 (1.3)	2 (8.7)	2 (2.5)	0 (0.0)	7 (2.1)	0 (0.0)	9 (5.5)	0 (0.0)	2 (2.5)	1 (25.0)	5 (3.2)	1 (33.3)
Median Treatment Duration (days)	186		195		190		189		189		191	
Retreatment Cases	145	48	50	4	169	24	129	27	48	6	126	21
Successfully treated	76 (52.4)	-	32 (64.0)	-	116 (68.6)	-	81 (62.8)	-	29 (60.4)	-	72 (57.1)	-
Treatment Interrupted	34 (23.5)	-	5 (10.0)	-	31 (18.3)	-	29 (22.5)	-	11 (22.9)	-	21 (16.7)	-
Treatment Failed	0 (0.0)	-	1 (2.0)	-	1 (0.6)	-	1 (0.8)	-	0 (0.0)	-	5 (4.0)	-
Died	5 (3.5)	6 (12.5)	2 (4.0)	0 (0.0)	4 (2.4)	1 (4.2)	5 (2.3)	4 (14.8)	2 (4.2)	1 (16.8)	5 (4.0)	0 (0.0)
Median Treatment Duration (days)	222		227		229		231		214		219	

With the exception of 'death', standard outcome definitions were not available for drug-resistant cases and results could therefore not be reported. Results for other, non-listed outcomes are also not reported. Median treatment period could not be reported for drug-resistant cases as they are not necessarily exclusively treated at the community clinic. Cases that could not be classified as either 'New' or 'Retreatment' are not reported.

Table S2. Molecular epidemiological analysis by clade of drug-susceptible and drug-resistant cases

Clade	**Beijing**		**Haarlem**		**LAM**		**LC**		**Quebec**		**Other**	
	S	R	S	R	S	R	S	R	S	R	S	R
Strains	76	17	42	2	198	19	16	9	59	9	130	18
Transmission chains	95	17	54	2	256	20	23	10	74	80	159	18
Max transmission chain clustersize	128	37	41	1	42	4	-	-	8	2	13	4
Unique cases	64	12	42	2	185	16	11	5	55	6	117	15
Clustered transmission chains (clusters)	31	5	12	0	71	4	12	5	19	2	42	3
Clustered transmission chain cases	288	51	83	0	396	11			66	4	151	9
(%)	(81.8)	(81.0)	(66.4)	(0.0)	(61.5)	(40.7)	-	-	(54.5)	(40.0)	(56.3)	(37.5)
Recent Transmission (%)	73.0	73.0	56.8	0.0	46.8	25.9	-	-	38.8	20.0	40.7	25.0

Clustering data and analysis is derived from transmission chains defined as cases having identical strains with inter-strain intervals of <2 years.

Cases classified as retreatment after failure/interruption were excluded.

Clustering analysis for the LC clade could not be done due to the poor resolution of IS*6110* with fewer than 6 copies

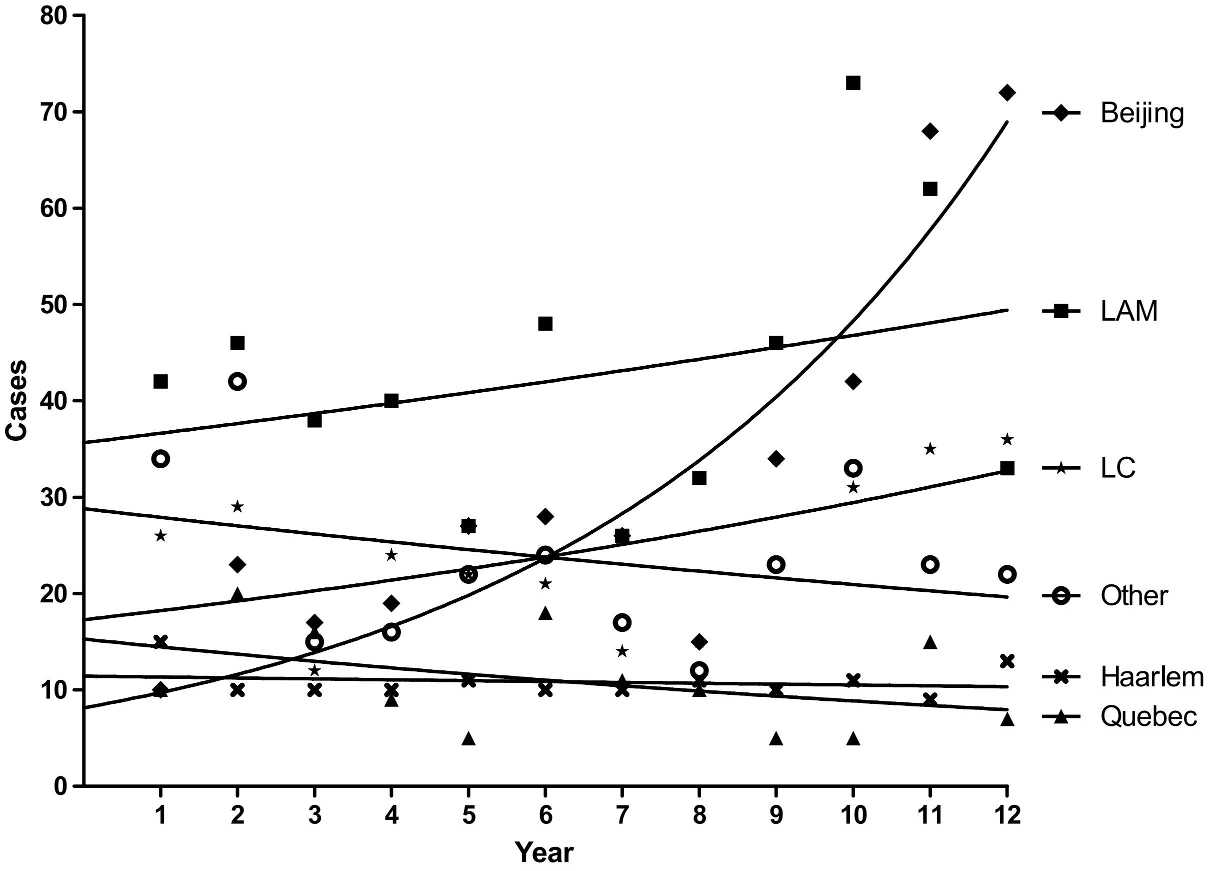

Figure S1. Annual incidence of drug-susceptible cases belonging to the 5 major *M. tuberculosis* clades with the remaining cases grouped as 'Other'. Calculated doubling times in years: Beijing (♦) = 3.89, LAM (■) = 25.49, LC (★) = 13.01, Haarlem (×) = -82.57, Quebec (▲) = -12.72, Other (○) = -21.72

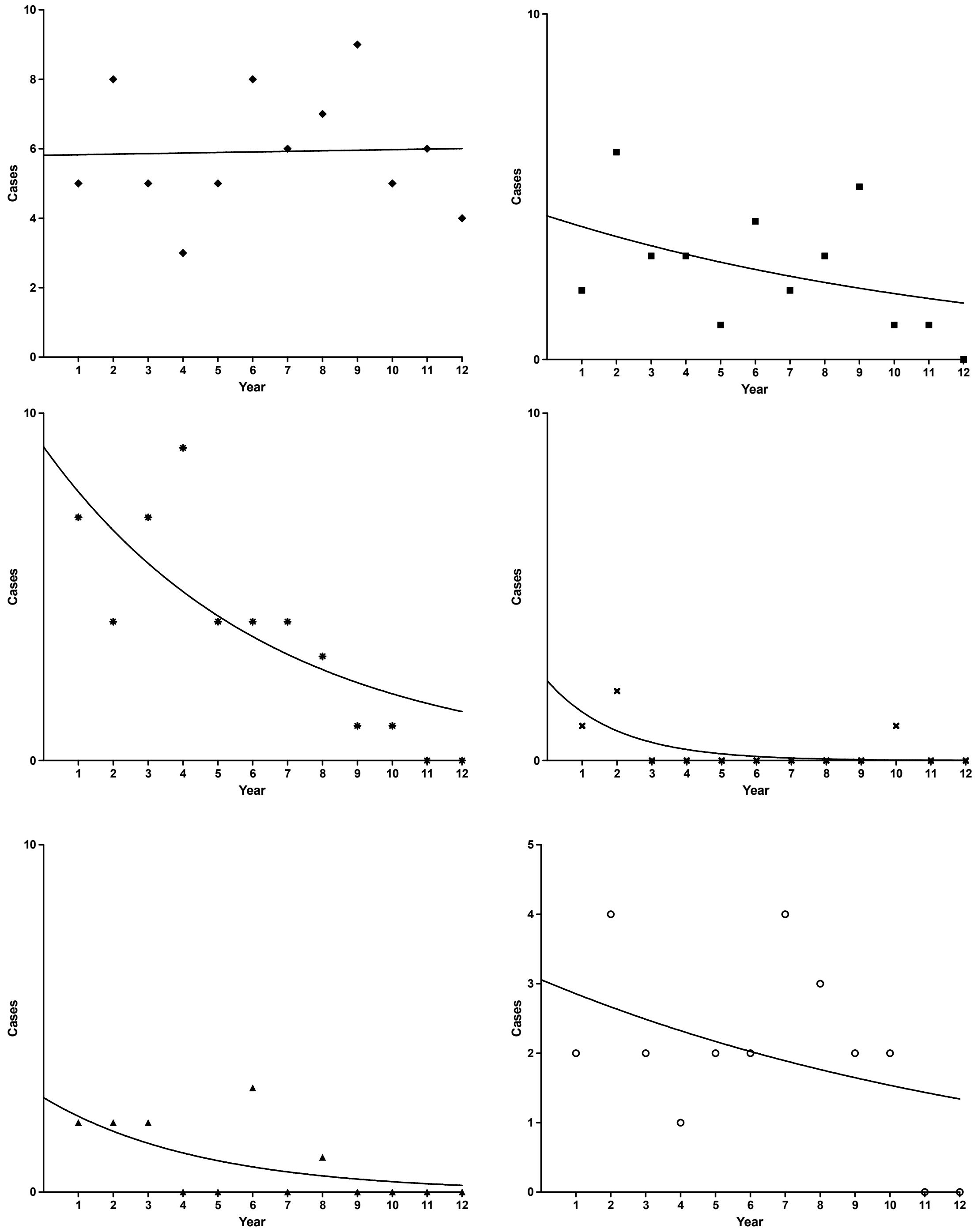

Figure S2. Annual incidence of drug-resistant cases belonging to the 5 major *M. tuberculosis* clades with the remaining cases grouped as 'Other'. Calculated doubling times in years: Beijing (♦) = 252, LAM (■) = -8.90, LC (★) = -4.48, Haarlem (✖) = -1.40, Quebec (▲) = -3.15

7

Evidence that the spread of *Mycobacterium tuberculosis* strains with the Beijing genotype is human population dependent.

Hanekom M., van der Spuy G.D., Gey van Pittius N.C., McEvoy C.R.E, Ndabambi S.L., Victor T.C., Hoal E. G., van Helden P.D., Warren R.M.

J. Clin. Microbiol. (2007) 45(7):2263-2266

Abstract

This study describes a comparative analysis of Beijing Mycobacterial Interspersed Repetitive Unit (MIRU) types of *Mycobacterium tuberculosis* isolates from Cape Town (South Africa) and East Asia. The results show a significant association between the frequency of occurrence of strains from defined Beijing sublineages and the human population from whom they were cultured [$P<0.0001$].

Main Text

Mycobacterium tuberculosis strains with the Beijing genotype have been shown to be globally widespread and are particularly prevalent in East Asia, where over 80% of strains from the Beijing region are of this genotype (5). It has been hypothesized that Beijing strains have evolved unique properties including the ability to evade the protective effect of BCG vaccination (19) and the ability to spread more efficiently than non-Beijing strains (2). However, clinical presentation of tuberculosis caused by a Beijing strain was found to vary between different geographical settings (3-5,16). Currently it is not known whether the observed variability in clinical presentation is a function of the Beijing strain population found in particular geographical settings, a function of the genetic composition of the human population, or a combination of these two variables.

This study aimed to test the hypothesis that host-pathogen compatibility determined the Beijing strain population structure in different host populations in different geographical settings. *M. tuberculosis* cultures from patients of mixed ancestry (14) who were resident in Cape Town, South Africa (20), were classified as Beijing genotype by spoligotyping (10). Beijing strains were assigned to phylogenetic sublineages as previously described (8) and were genotyped by MIRU typing (17).

During the study period January 1993 to December 2004, twenty-five MIRU types were identified among 321 tuberculosis cases with a Beijing strain (Table 1). A comparison between the MIRU type data from Beijing strains from Cape Town and previously published MIRU type data from Beijing strains from East Asia (1,9,12,13,15,18) showed that 9 of the Beijing MIRU types were shared between these geographical settings (MT01, MT08, MT11, MT18, MT19, MT21, MT28, MT33 and MT54) (Table 1). This suggests that the 9 shared Beijing MIRU types represent founder strains that were introduced into Cape Town from East Asia, as the latter is thought to be the evolutionary origin of strains with a Beijing genotype (5,7,12). The definition of founder MIRU types was supported by their disproportionately high number (n = 267) as compared to those with non-founder MIRU types (n = 54) in tuberculosis patients from Cape Town [z-test for the hypothesis that proportion of founder MIRU types = 0.5, P = 0.001].

Superposition of the Beijing MIRU type data onto the previously described phylogenetic tree of the Beijing strain family (8) provided a framework to predict the evolutionary order in which the 25 Beijing MIRU types had evolved (Figure 1). From this prediction the Beijing MIRU types could be partitioned into 7 Beijing sublineages. The number of founder Beijing MIRU types was variable among the different Beijing sublineages (Figure 1). Twenty-four of the Beijing MIRU types were unique to their respective sublineages, while the remaining Beijing MIRU type (MT11) was shared by three different sublineages 2, 3 and 6 (Figure 1 and Table 1). Suggesting that MT11 was an ancestral Beijing MIRU type (12)

To determine the propensity of Beijing strains from different sublineages to spread in the human population in Cape Town, the number of cases in circulation within each sublineage was compared to the number of founder strains for that sublineage (Table 1). The number of representatives of these founder strains was shown to be over-represented in sublineage 7 (n = 233 cases from 1 founder strain) vs. sublineage 1 to 6 (n = 88 cases from 8 founder strains) {z-test for the hypothesis that proportion of sublineage 7 cases = 0.11 (1/9), P =0.001}. In comparison, the founder Beijing MIRU types MT01, MT08, MT11, MT18, MT19 MT21, MT33 and MT54 were over-represented in the human population in East Asia as compared to South Africa {China (73/130), Hong Kong (108/211), Vietnam (25/37) and Singapore (45/56) vs. Cape Town (79/321); Fisher's exact test Odds Ratio (OR) = 4.20; CI95% 3.06 to 5.77, P =0.0001)}.

A significant association was observed between the frequency of occurrence of strains from defined Beijing sublineages and the human population from whom they were isolated {sublineage 1 to 6, Cape Town (n = 88) and East Asia (n = 253) vs. sublineage 7, Cape Town (n = 233) and East Asia (n = 43); Fisher's exact test OR = 15.58; CI95% 10.38 to 23.38, P =0.0001}.

It is unlikely that these findings can be explained by multiple importations of founder strains from sublineage 7 in preference to founder strains from sublineages 1 to 6 given that immigrants to South Africa were derived from many different geographical regions in East Asia and that sublineage 7 founder strains are less frequently observed in East Asia. Accordingly, we propose that the situation in Cape Town represents an approximation to a common starting point for all the introduced founder strains, with those best adapted to the local population spreading most efficiently. This could be due to the innate characteristics of the strains within defined Beijing sublineages or the local host population. Susceptibility to *M. tuberculosis per se* has frequently been associated with HLA genotype (11), and HLA allele frequencies are known to differ widely between human populations having different histories, with certain alleles totally absent in some populations. Our conclusion differs from that of Gagneux *et al.* (6) as we demonstrate that strains from a defined sublineage (subset of strains from an evolutionary lineage) may have been selected by a human population in a defined geographical setting.

Table 1. Geographical distribution of Beijing MIRU types from Asia and South Africa.

		Number of copies in polymorphic MIRU loci												Number of strains (%)						
MIRU type	Beijing sublineage [b]	2	4	10	16	20	23	24	26	27	31	39	40	RUS[c,d]	CHN[d]	HK[e]	VNM[d]	SGP[f]	BGD[g]	CT-SA
MT01[a]	4	2	2	3	3	2	5	1	7	3	4	3	3	1 (2.2)	5 (3.8)	7 (3.3)			2 (16.7)	1 (0.3)
MT02	NA	2	2	3	3	2	5	1	5	3	5	3	3	27 (60.0)	2 (1.5)	4 (1.9)		2 (3.6)	7 (58.3)	
MT04	NA	2	2	3	2	2	5	1	5	3	5	3	3	1 (2.2)			1 (2.7)			
MT05	NA	2	2	3	3	2	5	1	4	3	5	3	3	1 (2.2)						
MT07	NA	2	2	3	3	2	5	1	5	3	6	3	3	2 (4.4)						
MT08[a]	6	2	2	3	2	2	5	1	7	3	5	3	3	1 (2.2)	2 (1.5)	3 (1.4)				2 (0.6)
MT09	NA	2	2	3	3	2	5	1	7	3	5	3	1	1 (2.2)			1 (2.7)			
MT11[a]	2	2	2	3	3	2	5	1	7	3	5	3	3	10 (22.2)	42 (32.3)	77 (36.5)	10 (27.0)	39 (69.6)		3 (0.9)
	3																			2 (0.6)
	6																			8 (2.5)
MT12	NA	2	2	1	3	2	5	1	7	3	5	3	3	1 (2.2)	1 (0.8)					
MT13	NA	2	2	3	3	2	6	1	7	1	5	3	1			8 (3.8)	1 (2.7)			
MT14	NA	2	2	3	3	2	5	1	7	3	6	3	3		2 (1.5)	3 (1.4)	1 (2.7)	2 (3.6)		
MT16	NA	2	2	3	3	2	5	1	7	3	5	3	2		5 (3.8)	12 (5.7)		5 (8.9)		
MT17	NA	2	2	3	4	2	5	1	7	3	5	3	3		1 (0.8)			1 (1.8)		
MT18[a]	6	2	2	3	3	2	5	1	7	3	5	4	3		2 (1.5)	5 (2.4)		2 (3.6)		7 (2.2)
MT19[a]	5	2	0	3	3	2	5	1	7	3	5	3	3			2 (0.9)	1 (2.7)			39 (12.1)
MT20	7	2	2	2	3	2	5	1	7	4	5	4	3							19 (5.9)
MT21[a]	6	2	2	3	3	2	5	1	8	3	5	3	3		4 (3.1)	6 (2.8)	2 (5.7)	4 (7.1)		13 (4.0)
MT25	7	2	2	2	3	2	5	1	7	3	3	4	3							18 (5.6)
MT26	6	2	2	3	3	2	5	1	7	3	5	4	4							1 (0.3)
MT27	7	2	2	2	3	2	5	1	3	3	5	4	3							1 (0.3)
MT28[a]	7	2	2	2	3	2	5	1	7	3	5	4	3		16 (12.3)	23 (10.9)	4 (10.8)			189 (58.9)
MT29	6	2	2	3	3	2	5	1	7	4	5	3	3							1 (0.3)
MT33[a]	6	NA	2	3	3	NA	NA	NA	6	3	5	3	3		11 (8.5)	4 (1.9)	12 (32.4)			3 (0.9)
MT37	NA	NA	2	3	3	NA	NA	NA	2	3	5	3	3		2 (1.5)				1 (8.3)	
MT43	NA	NA	2	3	3	NA	NA	NA	6	3	4	3	3						2 (16.7)	
MT44	NA	NA	2	2	3	NA	NA	NA	7	3	5	3	3		6 (4.6)	9 (4.3)	1 (2.7)			
MT47	NA	NA	2	3	3	NA	NA	NA	7	1	5	3	1		3 (2.3)	6 (2.8)	3 (8.1)			
MT48	NA	NA	2	3	3	NA	NA	NA	4	3	4	3	3		2 (1.5)					

		Number of copies in polymorphic MIRU loci												Number of strains (%)						
MIRU type	**Beijing sublineage** [b]	**2**	**4**	**10**	**16**	**20**	**23**	**24**	**26**	**27**	**31**	**39**	**40**	**RUS**[c,d]	**CHN**[d]	**HK**[e]	**VNM**[d]	**SGP**[f]	**BGD**[g]	**CT-SA**
MT49	NA	NA	2	7	2	NA	NA	NA	7	3	4	3	3		3 (2.3)					
MT50	NA	NA	2	5	2	NA	NA	NA	7	3	4	3	3		2 (1.5)					
MT51	NA	NA	2	3	2	NA	NA	NA	5	1	5	3	3		2 (1.5)					
MT52	NA	NA	2	3	3	NA	NA	NA	6	3	5	3	2		2 (1.5)					
MT53	NA	NA	2	3	3	NA	NA	NA	7	3	5	1	3		2 (1.5)					
MT54[a]	6	NA	2	3	3	NA	NA	NA	7	3	5	2	3		7 (5.4)	4 (1.9)				1 (0.3)
MT55	NA	NA	2	2	3	NA	NA	NA	5	3	3	4	3		2 (1.5)					
MT56	NA	NA	2	2	3	NA	NA	NA	9	3	5	4	3		2 (1.5)					
MT57	NA	NA	2	2	3	NA	NA	NA	7	3	5	2	3		2 (1.5)					
MTSing76	NA	2	2	6	2	2	5	1	7	3	4	3	3					1 (1.8)		
MTZAF1	1	2	2	3	3	2	5	1	6	3	5	3	1							1 (0.3)
MTZAF2	5	2	3	3	3	2	5	1	7	3	5	3	3							1 (0.3)
MTZAF3	6	2	2	3	3	2	5	1	7	2	5	3	3							3 (0.9)
MTZAF4	6	2	2	3	3	2	5	1	5	3	5	3	3							1 (0.3)
MTZAF5	6	2	2	4	3	2	6	1	5	3	3	2	3							1 (0.3)
MTZAF6	7	2	2	2	3	2	5	1	7	4	4	4	3							1 (0.3)
MTZAF7	7	2	2	2	3	2	5	1	8	4	5	4	3							1 (0.3)
MTZAF8	7	2	2	2	3	2	5	1	9	3	5	3	3							1 (0.3)
MTZAF9	7	2	2	2	3	2	5	1	7	3	4	4	3							1 (0.3)
MTZAF10	7	2	2	2	3	2	5	1	5	3	5	4	3							1 (0.3)
MTZAF11	7	2	2	4	3	2	5	1	5	3	3	2	3							1 (0.3)
MTHK1	NA	2	2	3	3	2	5	1	7	3	5	3	4			6 (2.8)				
MTHK2	NA	2	2	3	3	2	5	1	7	3	5	5	3			5 (2.4)				
MTHK3	NA	2	2	3	4	2	5	1	7	3	5	3	2			5 (2.4)				
MTHK4	NA	2	1	3	3	2	5	1	7	3	5	3	3			4 (1.9)				
MTHK5	NA	2	2	3	3	2	5	1	6	3	5	3	4			3 (1.4)				
MTHK6	NA	2	0	3	3	2	5	1	7	3	5	3	4			3 (1.4)				
MTHK7	NA	2	2	3	3	2	5	1	7	3	5	4	2			2 (0.9)				
MTHK8	NA	2	2	3	3	2	5	1	8	3	5	5	3			2 (0.9)				
MTHK9	NA	2	2	2	3	2	5	1	7	3	5	3	2			2 (0.9)				
MTHK10	NA	2	2	3	3	2	5	1	6	3	7	3	3			2 (0.9)				
MTHK11	NA	2	2	3	3	2	2	1	7	4	4	3	3			2 (0.9)				
MTHK12	NA	2	2	2	3	2	5	1	7	3	5	4	4			2 (0.9)				

RUS – Russia; CHN – China; HK – Hong Kong; VNM – Vietnam; SGP – Singapore; BGD – Bangladesh; CT-SA – Cape Town South Africa

NA – not available, [a] founder MIRU types , [b] according to (8), [c] according to (13), [d] according to (12), [e] according to (9), [f] according to (15), [g] according to (1)

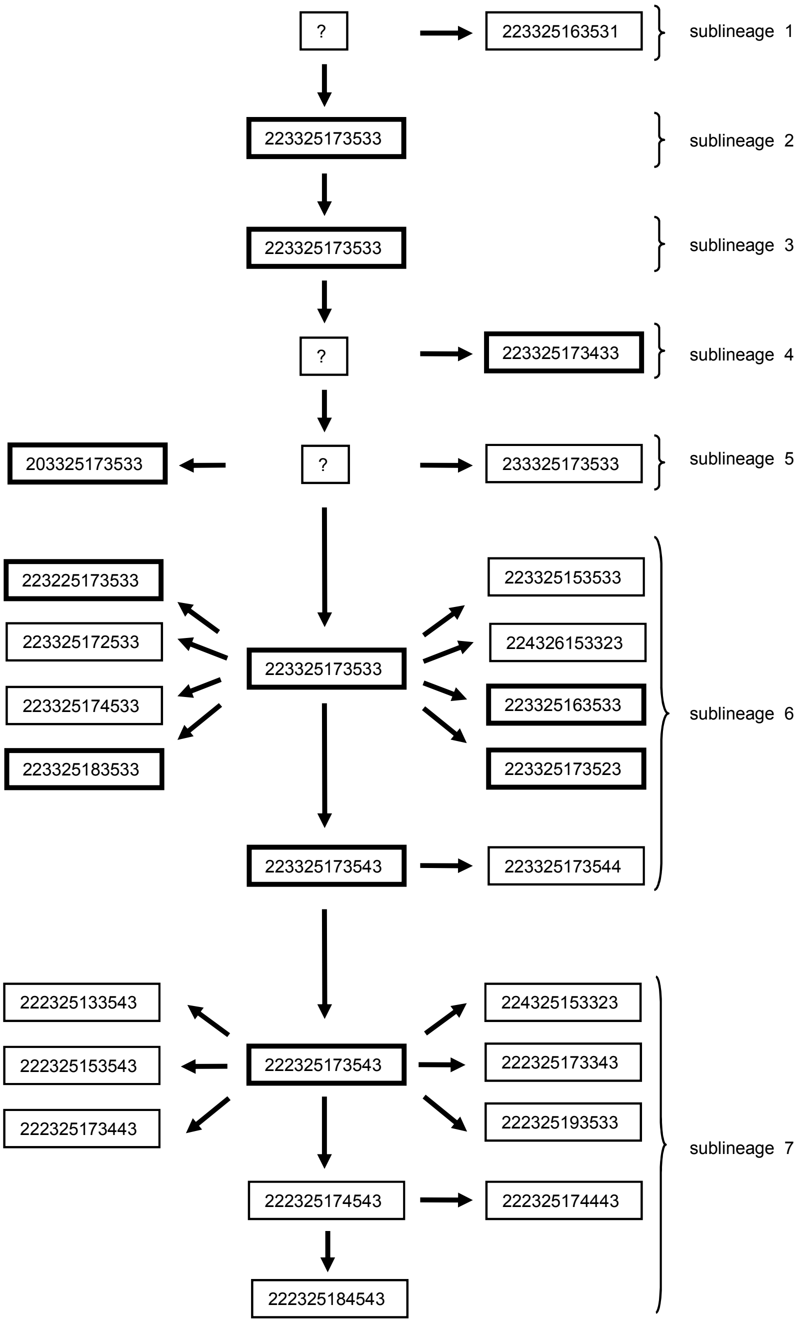

Figure 1. Evolutionary scenario of Beijing MIRU types according to Beijing sublineages. Beijing MIRU types were grouped according to their respective Beijing sublineages (8) and the most parsimonious evolutionary order was proposed. Beijing MIRU types are indicated within each box. Founder Beijing MIRU types are indicated by a bold box. Unknown Beijing MIRU types are indicated by "?".

In summary, the global success of the Beijing lineage may reflect either the selection of defined sublineages in different geographical settings by distinct human populations, or the adaptation of strains in a defined sublineage to spread more readily in a distinct human population. We acknowledge that these contrasting conclusions cannot be easily distinguished with the data available. However, the emergence of a sublineage of Beijing strains with increased pathogenicity may have important implications for the Tuberculosis Control Program. Early diagnosis and contact tracing will be essential to curb the spread of these strains. Furthermore, it will be important to ensure that future vaccines protect against these strains.

Acknowledgements

TB in the 21st Century Consortium, the Harry Crossley Foundation and the European Commission 6th framework programme on research technological development demonstration (project no. 037919) are thanked for financially supporting this project. Dr M. Kidd is thanked for statistical advice.

Reference List

1. **Banu, S., S. V. Gordon, S. Palmer, M. R. Islam, S. Ahmed, K. M. Alam, S. T. Cole, and R. Brosch**. 2004. Genotypic analysis of *Mycobacterium tuberculosis* in Bangladesh and prevalence of the Beijing strain. J.Clin.Microbiol. 42:674-682.

2. **Bifani, P. J., B. Mathema, N. E. Kurepina, and B. N. Kreiswirth**. 2002. Global dissemination of the *Mycobacterium tuberculosis* W-Beijing family strains. Trends Microbiol. 10:45-52.

3. **Borgdorff, M. W., H. van Deutekom, P. E. de Haas, K. Kremer, and D. van Soolingen**. 2004. *Mycobacterium tuberculosis*, Beijing genotype strains not associated with radiological presentation of pulmonary tuberculosis. Tuberculosis. (Edinb.) 84:337-340.

4. **Drobniewski, F., Y. Balabanova, V. Nikolayevsky, M. Ruddy, S. Kuznetzov, S. Zakharova, A. Melentyev, and I. Fedorin**. 2005. Drug-resistant tuberculosis, clinical virulence, and the dominance of the Beijing strain family in Russia. JAMA 293:2726-2731.

5. **European Concerted Action on New Generation Genetic Markers and Techniques for the Epidemiology and Control of Tuberculosis**. 2006. Beijing/W genotype *Mycobacterium tuberculosis* and drug resistance. Emerg.Infect.Dis. 12:736-743.

6. **Gagneux, S., K. DeRiemer, T. Van, M. Kato-Maeda, B. C. de Jong, S. Narayanan, M. Nicol, S. Niemann, K. Kremer, M. C. Gutierrez, M. Hilty, P. C. Hopewell, and P. M. Small**. 2006. Variable host-pathogen compatibility in *Mycobacterium tuberculosis*. Proc.Natl.Acad.Sci.U.S.A .

7. **Glynn, J. R., J. Whiteley, P. J. Bifani, K. Kremer, and D. van Soolingen**. 2002. Worldwide occurrence of Beijing/W strains of *Mycobacterium tuberculosis*: a systematic review. Emerg.Infect.Dis. 8:843-849.

8. **Hanekom, M., G. D. van der Spuy, E. Streicher, S. L. Ndabambi, C. R. McEvoy, M. Kidd, N. Beyers, T. C. Victor, P. D. van Helden, and R. M. Warren**. 2007. A recently evolved sublineage of the *Mycobacterium tuberculosis* Beijing strain family was associated with an increased ability to spread and cause disease. J.Clin.Microbiol. (Epub)

9. **Kam, K. M., C. W. Yip, L. W. Tse, K. L. Wong, T. K. Lam, K. Kremer, B. K. Au, and D. van Soolingen**. 2005. Utility of mycobacterial interspersed repetitive unit typing for differentiating multidrug-resistant *Mycobacterium tuberculosis* isolates of the Beijing family. J.Clin.Microbiol. 43:306-313.

10. **Kamerbeek, J., L. Schouls, A. Kolk, M. van Agterveld, D. van Soolingen, S. Kuijper, A. Bunschoten, H. Molhuizen, R. Shaw, M. Goyal, and J. Van Embden**. 1997. Simultaneous detection and strain differentiation of *Mycobacterium tuberculosis* for diagnosis and epidemiology. J.Clin.Microbiol. 35:907-914.

11. **Lombard, Z., A. E. Brune, E. G. Hoal, C. Babb, P. D. van Helden, J. T. Epplen, and L. Bornman**. 2006. HLA class II disease associations in southern Africa. Tissue Antigens 67:97-110.

12. **Mokrousov, I., H. M. Ly, T. Otten, N. N. Lan, B. Vyshnevskyi, S. Hoffner, and O. Narvskaya**. 2005. Origin and primary dispersal of the *Mycobacterium tuberculosis* Beijing genotype: clues from human phylogeography. Genome Res. 15:1357-1364.

13. **Mokrousov, I., O. Narvskaya, E. Limeschenko, A. Vyazovaya, T. Otten, and B. Vyshnevskiy**. 2004. Analysis of the allelic diversity of the mycobacterial interspersed repetitive units in *Mycobacterium tuberculosis* strains of the Beijing family: practical implications and evolutionary considerations. J.Clin.Microbiol. 42:2438-2444.

14. **Nurse, G. T., J. S. Weiner, and T. Jenkins**. 1985. The Peoples of Southern Africa and their Affinities, p. 410. *In* Clarendon Press, Oxford (United Kingdom).

15. **Sun, Y. J., R. Bellamy, A. S. Lee, S. T. Ng, S. Ravindran, S. Y. Wong, C. Locht, P. Supply, and N. I. Paton**. 2004. Use of mycobacterial interspersed repetitive unit-variable-number tandem repeat typing to examine genetic diversity of *Mycobacterium tuberculosis* in Singapore. J.Clin.Microbiol. 42:1986-1993.

16. **Sun, Y. J., T. K. Lim, A. K. Ong, B. C. Ho, G. T. Seah, and N. I. Paton**. 2006. Tuberculosis associated with *Mycobacterium tuberculosis* Beijing and non-Beijing genotypes: A clinical and immunological comparison. BMC.Infect.Dis. 6:105.

17. **Supply, P., S. Lesjean, E. Savine, K. Kremer, D. van Soolingen, and C. Locht**. 2001. Automated high-throughput genotyping for study of global epidemiology of *Mycobacterium tuberculosis* based on mycobacterial interspersed repetitive units. J.Clin.Microbiol. 39:3563-3571.

18. **Supply, P., R. M. Warren, A. L. Banuls, S. Lesjean, G. D. van der Spuy, L. A. Lewis, M. Tibayrenc, P. D. van Helden, and C. Locht**. 2003. Linkage disequilibrium between minisatellite loci supports clonal evolution of *Mycobacterium tuberculosis* in a high tuberculosis incidence area. Mol.Microbiol. 47:529-538.

19. **van Soolingen, D., L. Qian, P. E. de Haas, J. T. Douglas, H. Traore, F. Portaels, H. Z. Qing, D. Enkhsaikan, P. Nymadawa, and J. D. van Embden**. 1995. Predominance of a

single genotype of *Mycobacterium tuberculosis* in countries of east Asia. J.Clin.Microbiol. 33:3234-3238.

20. **Verver, S., R. M. Warren, Z. Munch, E. Vynnycky, P. D. van Helden, M. Richardson, G. D. van der Spuy, D. A. Enarson, M. W. Borgdorff, M. A. Behr, and N. Beyers**. 2004. Transmission of tuberculosis in a high incidence urban community in South Africa. Int.J.Epidemiol. 33:351-357.

8

Discordance between MIRU-VNTR and IS*6110* RFLP genotyping when analyzing *Mycobacterium tuberculosis* Beijing strains in a high incidence setting

Hanekom M.[§], van der Spuy G.D.[§], Gey van Pittius N.C., McEvoy C.R.E., Hoek K.G.P., Ndabambi S.L., Jordaan A.M., Victor T.C., van Helden P.D., Warren R.M.

[§] Joint First Authorship

J. Clin. Microbiol. (Accepted, August 2008)

ABSTRACT

IS*6110*-Restriction Fragment Length Polymorphism (RFLP) genotyping is the most widely used genotyping method to study the epidemiology of *Mycobacterium tuberculosis*. However, due to the complexity of the IS*6110*-RFLP genotyping technique and the interpretation of RFLP data, Mycobacterial Interspersed Repetitive-Unit–Variable-Number Tandem-Repeat (MIRU-VNTR) genotyping has been proposed as the new genotyping standard. This study aimed to determine the discriminatory power of different MIRU-VNTR locus combinations relative to IS*6110*-RFLP genotyping, using a collection of Beijing genotype *M. tuberculosis* strains with a well established phylogenetic history. Clustering, diversity index, clustering concordance, concordance among unique genotypes, divergent and convergent evolution were calculated for 7 combinations of 27 different MIRU-VNTR loci and compared to IS*6110*-RFLP. Our results confirmed previous findings that MIRU-VNTR genotyping could be used to estimate the extent of recent or ongoing transmission. However, molecular epidemiological linking of cases varied significantly depending on the genotyping method used. We conclude that IS*6110*-RFLP and MIRU-VNTR loci evolve independently and at different rates which leads to discordance between transmission chains predicted by the respective genotyping methods. Concordance between the two genotyping methods could be improved by the inclusion of genetic distance into the clustering formula for some of the MIRU-VNTR loci combinations. In summary, our findings differ from previous reports, which may be explained by the fact that in low incidence settings, the genetic distance between epidemiologically unrelated isolates was sufficient to define a strain using either marker, whereas in high incidence settings, continuous evolution and persistence of strains revealed the weaknesses inherent to these markers.

INTRODUCTION

Over the past two decades, molecular genotyping methods have enhanced our understanding of the epidemiology of tuberculosis (TB) in numerous geographical settings. These methods have enabled geo-temporal tracking of *Mycobacterium tuberculosis* strains with the view to identifying source cases responsible for TB outbreaks (2), tracking of recent and ongoing disease transmission (31), distinguishing between re-infection and relapse (28), evaluation of the effectiveness of Direct Observed Therapy Short-course (DOTS) based TB control programs (4,16), and identification of global genetic lineages (6). Ideally, molecular genotyping tools should be inexpensive, highly discriminative, deliver rapid results, be straightforward to perform and produce easily interpretable results that allow for accurate inter-laboratory comparisons (universally comparable databases).

Three genotyping methods are currently widely used in molecular epidemiological studies of TB: IS*6110*-Restriction Fragment Length Polymorphism (RFLP) genotyping (27), spoligotyping (13) and Mycobacterial Interspersed Repetitive-Unit–Variable-Number Tandem-Repeat (MIRU-VNTR) genotyping (21,22). Currently, IS*6110*-RFLP genotyping is the most widely used genotyping method (27). However, this method is time-consuming, laborious and complex. Furthermore, differences in application can make inter-laboratory comparisons difficult and the data generated may have limitations (i.e. comparison of strains with high *vs.* low IS*6110* copy numbers). More recently, the validity of the calculation of IS*6110*-RFLP clustering, as a surrogate for transmission, has been questioned as the IS*6110* banding pattern may change during transmission (33,35). A nearest genetic distance (NGD) model has been evaluated to incorporate IS*6110* banding changes into the calculation of ongoing transmission (24). The term cluster has also been questioned in studies which have compared contact tracing data with IS*6110*-RFLP data (3,26). In response, numerous studies have been conducted to try to identify alternative methods that have the ability to accurately describe epidemiological events in different settings at a similar discriminatory level to that of IS*6110*-RFLP genotyping. One of the most promising methods is MIRU-VNTR genotyping, a PCR-based method for detecting the number of tandem repeats at a given genetic locus. Supply *et al.* (21) defined a set of 15 MIRU-VNTR loci for molecular epidemiological investigations and a set of 24 MIRU-VNTR loci for phylogenetic analysis of *M. tuberculosis* strains worldwide. In support of this, another study concluded that this “real-time” MIRU-VNTR genotyping approach was highly applicable for population-based studies (18). This view was reinforced by a study conducted in the Brussels region, where the authors concluded that a standardized MIRU-VNTR genotyping method could be a new reference for epidemiological and phylogenetic screening of *M. tuberculosis* strains (15).

A study from Japan (9) investigated the differentiation power of the proposed 15- and 24-loci MIRU-VNTR genotyping methods for strains with the Beijing genotype and concluded that the analysis of these loci were of limited use for discriminating strains of this genotype. In their study they showed that VNTR loci 3820, 3232 and 4120 were highly polymorphic in Beijing genotype strains and thus proposed the use of these loci to enhance the discriminatory power of the proposed 15-MIRU-VNTR genotyping method. However, other studies have excluded these loci due to difficulties associated with the reproducibility of PCR amplification (14,21,36).

Subsequently, a study in Hong Kong, which also examined strains of the Beijing genotypes, showed that a different combination of 12 VNTR and QUB (Queen's University of Belfast) loci gave a Hunter-Gaston discriminatory index value which was almost equal to that obtained in IS*6110*-RFLP genotyping (11,12). However, this was refuted by a more recent study from China which suggests that MIRU-VNTR genotyping may over-estimate transmission in isolates with the Beijing genotype (10). Collectively, these findings suggest that the selection of MIRU-VNTR loci for optimal differentiation of *M. tuberculosis* requires further validation in different geographical settings. To date, the performance of the MIRU-VNTR genotyping method has not been evaluated in an epidemic setting, nor has it been tested within the context of a robust *M. tuberculosis* phylogeny.

In this study the discriminatory power of different MIRU-VNTR locus combinations was determined as previously described (7,9,21,22) and compared to the IS*6110*-RFLP genotyping method using a collection of Beijing genotype *M. tuberculosis* strains with a well established phylogenetic history (8). The results are discussed in the context of concordance between the different genotyping methods in their ability to define a strain and to accurately describe the epidemiology of TB in a high incidence setting.

METHODS

Study Population

Sputum samples were collected during the period from January 1993 to December 2004 from new and retreatment TB patients who were resident and attending healthcare clinics in an epidemiological field site in Cape Town, South Africa (31). This study forms part of a larger, long-term molecular epidemiological project which has been approved by the ethics committee of Stellenbosch University.

IS*6110* RFLP fingerprinting

M. tuberculosis isolates were cultured on MGIT (Becton Dickinson, USA) or Löwenstein-Jensen media and DNA was extracted as previously described (32). Each isolate was classified by IS*6110*-RFLP genotyping (27) and spoligotyping (13) using internationally standardized protocols. IS*6110*-RFLP patterns were analyzed using Gelcompar II (Applied-Maths, Sint_Martens_latem, Belgium) with tolerance settings allowing a 5% shift in lane position and a 0.6% variation in individual band position to compensate for minor technical error. Isolates were assigned as members of the Beijing genotype if they had the characteristic Beijing spoligotype (30). Only the first *M. tuberculosis* isolate from each case was included for subsequent analysis. Each Beijing isolate was grouped into one of seven phylogenetic sublineages according to 40 different genetic markers, as previously described (8).

DNA Sequencing

The DNA sequence of the *katG, rpoB*, *embB* and *rrs* genes of isolates classified as members of the Beijing sublineage 5 were determined as previously described (19,25).

MIRU-VNTR typing

Twenty seven MIRU-VNTR loci were amplified by PCR as described previously (7,9,21,22). The number of repeats at each genomic locus was calculated according to the electrophoretic mobility of the corresponding PCR product (23). Alleles were assigned numerical values according to the number of repeats present in that genomic locus. Isolates were genotypically classified according to seven different MIRU-VNTR locus combinations (Table 1).

Analytical Calculations

Estimation of clustering: A cluster (representing either recent or ongoing transmission: < 2 year and unrestricted interval, respectively) was defined as a series of isolates having the same genotype (IS*6110*-RFLP or MIRU-VNTR), while isolates with unique IS*6110*-RFLP or MIRU-VNTR genotypes were considered as representing reactivation or influx of disease into the study community (20). Secondary analyses which incorporated the concept of evolution during transmission were done using datasets (genotypes according to IS*6110*-RFLP or a particular MIRU-VNTR locus combination) in which isolates separated by a single evolutionary event were combined into transmission chains with a genetic distance of one (24).

Estimation of genetic diversity: The genetic diversity for each individual MIRU-VNTR loci, each of the seven MIRU-VNTR locus combinations (Table 1) and the IS*6110*-RFLP fingerprints was calculated as $h = 1 - \sum x_i^2 (n/(n-1))$, where x_i is the frequency of the *i*th allele at the locus, *n* is the number of isolates in the sample, and the term *n/(n - 1)* is a correction for bias in small samples (17).

Estimation of matching and mismatching concordance: Concordance between the IS*6110*-RFLP genotypes and the respective MIRU-VNTR genotypes was calculated as follows: Each isolate was paired with every other isolate in the dataset and their genotypes (IS*6110*-RFLP and MIRU-VNTR) were scored as either a match (identical) or mismatch (non-identical). Matching concordance between the respective genotyping methods was calculated according to the number of paired isolates having a match for both of the methods as a proportion of the total number of pairs having matching IS*6110*-RFLP genotypes. This is a measure of agreement between two methods as to whether any two isolates form part of the same transmission chain. Mismatching concordance was calculated as the number of paired isolates having non-matching genotypes for both of the methods as a proportion of the total number of pairs having non-matching IS*6110*-RFLP genotypes. This is a measure of agreement between two methods that any two isolates do not form part of the same transmission chain.

Table 1. MIRU-VNTR locus combinations.

	MIRU Locus Combinations						
	12-MIRU[a]	12-MIRU + ETR A, B, C[b]	12-MIRU + hyper-variable loci	15-MIRU-VNTR[c]	15-MIRU-VNTR + hyper-variable loci[d]	24-MIRU-VNTR[c]	24-MIRU-VNTR + hyper-variable loci
MIRU02	●	●	●			●	●
MIRU04	●	●	●	●	●	●	●
MIRU10	●	●	●	●	●	●	●
MIRU16	●	●	●	●	●	●	●
MIRU20	●	●	●			●	●
MIRU23	●	●	●			●	●
MIRU24	●	●	●			●	●
MIRU26	●	●	●	●	●	●	●
MIRU27	●	●	●			●	●
MIRU31	●	●	●	●	●	●	●
MIRU39	●	●	●			●	●
MIRU40	●	●	●	●	●	●	●
VNTR1955				●	●	●	●
VNTR2165/ETR-A		●		●	●	●	●
QUB11b				●	●	●	●
QUB26b				●	●	●	●
VNTR0424				●	●	●	●
VNTR2401				●	●	●	●
VNTR4156				●	●	●	●
VNTR 3690				●	●	●	●
ETR-C		●		●	●	●	●
VNTR 2347						●	●
ETR-B		●				●	●
Mtub 34						●	●
QUB3232			●		●		●
VNTR3820			●		●		●
VNTR 4120			●		●		●

●locus included

[a] according to (22), [b] according to (7), [c] according to (21), [d] according to (9), [e] this study

Estimation of concordance among unique genotypes: Concordance between uniquely occurring IS*6110* genotypes and the MIRU-VNTR genotypes was calculated as the proportion of isolates having unique IS*6110* genotypes that also had unique MIRU-VNTR genotypes.

Estimation of the number of convergent events: Convergent evolution was identified by drawing connecting lines between each IS*6110*-RFLP genotype and each MIRU-VNTR genotype for which isolates had been found having that genotype combination (Figure 1). Convergent evolution was defined, conservatively, as the existence of isolates representing each of the four possible combinations of two IS*6110*-RFLP genotypes (*e.g.* IS_1 and IS_2) and two MIRU-VNTR genotypes (*e.g.* M_1 and M_2) (Figure 1).

This scenario would only be possible if one of the MIRU-VNTR genotypes had evolved more than once, assuming that the chance of IS*6110*-RFLP genotype convergence was significantly lower than that of MIRU-VNTR genotype convergence. The validity of this method was confirmed by plotting the IS*6110*-RFLP genotypes onto a phylogenetic tree constructed using the MIRU-VNTR data in combination with the neighbor-joining algorithm (data not shown) (34).

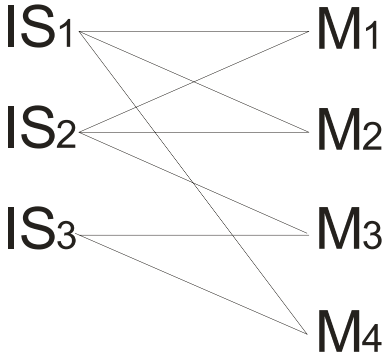

Figure 1. An example of MIRU-VNTR (Mx) and IS*6110* (ISx) genotypes. The connecting lines represent the MIRU-VNTR and IS*6110* genotype combinations observed in *M. tuberculosis* isolates in the study setting. M_1 and M_2 are both linked to IS_1 and IS_2 and therefore represent a convergent event. Neither M_3 nor M_4 share common connections to more than one IS_x with any other M_x. Their connecting lines therefore indicate simple, linear evolution.

Estimation of the number of divergent events: A divergent evolutionary event was scored for each MIRU-VNTR genotype which existed in combination with only one IS*6110*-RFLP genotype and where this IS*6110*-RFLP genotype was found in combination with more than one MIRU-VNTR genotype (Figure 1). This implies that the MIRU-VNTR genotype arose subsequent to the IS*6110*-RFLP genotype. A divergence event was also added for each convergent event, since a convergent event implies a prior divergent event.

Sensitivity and specificity calculations: The sensitivity and specificity (and positive and negative predictive values) of the IS*6110*-RFLP and respective MIRU-VNTR genotyping methods were calculated using Graphpad Prism 5 software (La Jolla, CA USA) for their ability to correctly identify an independently genotyped drug-resistant cluster.

RESULTS

IS*6110*-RFLP genotyping identified 74 different strains among the 321 isolates with the Beijing spoligotype collected over a 12 year period (Table 2). Of these strains, 272 were grouped into 25 clusters (containing between 2 and 100 isolates) and 49 were unique strains. The overall percent clustering was calculated to be 84.7% using the n/T formula (1). Each isolate was subsequently genotyped with 27 MIRU-VNTR loci and analyzed according to 7 different MIRU-VNTR locus combinations (Table 1 and

Supplemental data). The performance of these locus combinations, in relation to the IS*6110*-RFLP genotyping method, was determined either over a 12-year period (Table 2) or over six consecutive 2-year periods (Table 3). In both analyses the traditional 12-MIRU loci genotyping method under-estimated the number of genotypes (strains) identified and thereby over-estimated the percentage of clustering (Table 2 and 3). The inclusion of Exact Tandem Repeat (ETR) alleles A, B and C to the 12-MIRU loci set did not significantly improve the number of strains detected or the estimate of clustering (Table 2 and 3). Analysis of the isolates using the newly proposed 15- and 24-MIRU-VNTR locus combinations increased the number of strains identified, however, the discriminatory power of these locus combinations remained lower than that observed using IS*6110*-RFLP genotyping (Table 2 and 3). Consequently, these locus combinations over-estimated clustering. The addition of the VNTR loci 3232, VNTR 4120 and VNTR 3820 to the 12-, 15- and 24-MIRU-VNTR locus combinations increased the number of strains detected and thereby produced clustering estimates similar to or slightly lower than that of IS*6110*-RFLP genotyping (Table 2 and 3). This implies that some MIRU-VNTR loci combinations could be selected as epidemiological markers to estimate the extent of both recent (<2 year interval) and ongoing transmission (unrestricted interval) in settings with a high incidence of strains with the Beijing genotype.

To determine whether a correlation existed between the definitions of a strain according to IS*6110*-RFLP or MIRU-VNTR genotyping methods, the respective genotypes were compared. From the results shown in Table 2 it is evident that a strain classified as a cluster according to IS*6110*-RFLP genotyping, may in some instances be classified as unique according to the different MIRU-VNTR locus combinations or vice versa. Using a pair-wise analysis, we estimated the degree of matching concordance between the IS*6110*-RFLP and MIRU-VNTR genotyping methods to range between 39% and 68% depending on the locus combinations used (Table 2 and 3). The inclusion of additional MIRU-VNTR loci decreased the degree of matching concordance, as a result of an increased rate of divergence caused by more rapid evolution, with the hyper-variable loci having the greatest effect. Conversely, the inclusion of additional loci increased the degree of mismatching concordance, as well as concordance between strains identified as having unique genotypes according to both genotyping methods (IS*6110*-RFLP and MIRU-VNTR). A consequence of more rapid evolution was the increased risk of convergent evolutionary events (Table 2).

To determine whether concordance between the respective genotyping methods could be improved, the analysis was repeated to allow for genetic distance (GD) = 1, *i.e.* evolution of single MIRU-VNTR loci or single band changes in the IS*6110* pattern within the definition of a cluster. The results showed that the inclusion of genetic distance had a significant influence on the MIRU-VNTR definition of a cluster, collapsing many of the genotypes (Table 2 and Supplemental data 1). This was less pronounced for IS*6110*-RFLP (Table 2 and Supplemental data 1). Matching concordance was improved by allowing for evolution of the MIRU-VNTR genotypes; however, mismatching concordance was concomitantly reduced for genotypes based on the 12-MIRU loci combinations. This may be explained by the loss of discriminatory power as a result of the collapsing of genotypes, associated with a low rate of evolution. In contrast, mismatching concordance was improved for 15- and 24-MIRU-VNTR combinations due to the higher evolutionary rate of these markers. However, the concordance among unique genotypes remained low (Table 2).

To establish which of the genotyping methods provided the most accurate description of ongoing transmission in the study setting, the largest group of drug-resistant isolates (found within sublineage 5) was selected, based on identical mutations conferring resistance to isoniazid, rifampicin, ethambutol and streptomycin (see Supplemental data 2). These isolates represent the continuing spread of a previously described MDR-TB outbreak (29). A total of 35 isolates were identified with the *katG*315 AGC to ACC, *rpoB*531 TCG to TTG, *embB*306 ATG to ATA and *rrs*513 CAG to CCG mutations, forming a single drug-resistance-based cluster (Figure 2). The sensitivity, specificity and positive and negative predictive values related to the ability of the different markers to identify the drug-resistant cluster are given in Table 4. While the sensitivities of all the markers were high, with some of those based on MIRU-VNTR loci outperforming IS*6110*-RFLP, the specificity of all MIRU-VNTR markers was substantially lower than that of IS*6110*-RFLP. The inclusion of genetic distance (single events) within the definition of a cluster appeared to improve the sensitivity of most of the markers, but concomitantly decreased the specificity of the MIRU-VNTR markers. The specificity of IS*6110* was not affected by the inclusion of genetic distance. Positive predictive values were not significantly affected by allowing for evolution of the markers, however, with the exceptions of IS*6110*, which increased, and the 24 MIRU + 3 hyper-variable loci, which remained unchanged, the negative predictive values for all markers were reduced to zero.

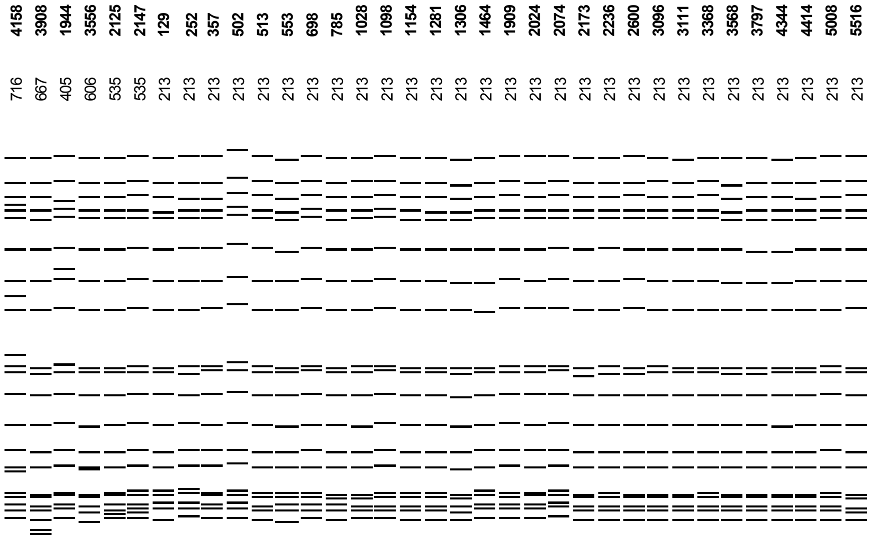

Figure 2. IS*6110*-RFLP banding patterns of Beijing sublineage 5 isolates sharing identical *katG, rpoB*, *embB* and *rrs* gene mutations. The isolate numbers are indicated in bold, while the IS*6110*-RFLP cluster number are indicated in standard text (these numbers correspond to the numbers given in the Supplemental data 2).

To determine whether MIRU-VNTR genotyping could be used as a method to phylogenetically group strains with the Beijing genotype, the correlation between MIRU-VNTR genotype and Beijing sublineage was quantified. As sublineages 3 and 4 and sublineages 5 and 6 were distinguished solely on the basis of IS*6110* in our dataset, these two pairs of sublineages were combined for the purposes of this analysis. Table 2 shows that the respective MIRU-VNTR locus combinations correctly grouped >96% of the isolates according to their sublineage designation in comparison to the 100% of IS*6110*-RFLP genotyping. The incorporation of genetic distance reduced the ability of genotyping methods based on the 12-MIRU locus combinations to correctly group isolates (Table 2).

Table 2. Comparison between molecular epidemiological data generated over a 12 year interval by IS*6110*-RFLP and MIRU-VNTR genotyping methods

			MIRU-VNTR Locus Combinations						
		IS*6110*-RFLP	**12-MIRU[a]**	**12-MIRU + ETR A, B, C[b]**	**12-MIRU + hyper-variable loci[e]**	**15-MIRU-VNTR[c]**	**15-MIRU-VNTR + hyper-variable loci[d]**	**24-MIRU-VNTR[c]**	**24-MIRU-VNTR + hyper-variable loci[e]**
GD = 0	**Genotypes (n)**	74	27	39	67	47	83	57	91
	Clustering (%)	84.7	95.0	91.2	84.1	89.4	78.8	87.2	77.3
	Unique IS*6110* genotypes (n = 49) with unique MIRU-VNTR genotypes (n)	N/A	6	8	14	12	19	12	19
	Unique IS*6110* genotypes (n = 49) with clustered MIRU-VNTR genotypes (n)	N/A	43	41	35	37	30	37	30
	Clustered IS*6110* genotype (n = 272) with unique MIRU-VNTR genotypes (n)	N/A	10	18	38	22	49	28	54
	Clustered IS*6110* genotype (n = 272) with clustered MIRU-VNTR genotypes (n)	N/A	262	254	234	250	223	244	218
	Pair-wise matching concordance (%)	100	68	60	53	67	51	52	39
	Pair-wise mismatching concordance (%)	N/A	69	72	71	68	71	73	81
	Concordance between unique strains (%)	100	20.4	22.4	28.6	34.7	40.8	40.8	42.9
	Converged genotypes (n)	N/A	3	5	5	3	5	7	9
	Diverged genotypes (n)	N/A	14	26	47	27	58	38	68
	Diversity index[f]	0.85	0.63	0.67	0.7	0.63	0.7	0.72	0.78
GD = 1	**Genotypes (n)**	40	3	4	11	11	20	15	27
	Pair-wise matching concordance (%)	100	99	99	96	99	96	99	96
	Pair-wise mismatching concordance (%)	N/A	2	3	8	80	82	83	88
	Concordance between unique strains (%)	100	0	4	4	15	30	22	41
	Beijing sublineage discrimination GD=0 (%)	100	96	96	99	99	99	99	99
	Beijing sublineage discrimination GD=1 (%)	99	1	1	3	99	99	99	99

[a] according to (22), [b] according to (7), [c] according to (21), [d] according to (9), [e] this study , [f] according to (15), N/A not applicable

Table 3. Comparison between molecular epidemiological data generated over six consecutive 2 year intervals by IS*6110*-RFLP and MIRU-VNTR genotyping methods

		MIRU-VNTR Locus Combinations						
	IS*6110*-RFLP	**12-MIRU**[a]	**12-MIRU + ETR A, B, C**[b]	**12-MIRU + hyper-variable loci**[e]	**15-MIRU-VNTR**[c]	**15-MIRU-VNTR + hyper-variable loci**[d]	**24-MIRU-VNTR**[c]	**24-MIRU-VNTR + hyper-variable loci**[e]
Average No strains (range)	17.9 (7-20)	9.9 (5-14)	10.9 (7-14)	15.4 (8-22)	12.0 (5-17)	16.6 (7-23)	13.6 (7-19)	18.1 (9-25)
Average No clusters (range)	6.1 (4-10)	4.6 (2-7)	4.9 (3-7)	4.0 (2-6)	3.7 (2-5)	3.6 (2-6)	4.6 (3-6)	4.3 (3-7)
Average % clustering (range)	74.5 (55.4-86.7)	85.1 (60.0-93.1)	83.1 (60.0-91.7)	73.1 (60.0-77.8)	80.3 (66.7-93.1)	70.7 (57.6-76.4)	77.8 (60.0-89.7)	67.7 (54.5-75.0)
Average % pair-wise matching concordance (range)	100	69.7 (60-80)	63.7 (44-72)	52.7 (40-58)	68.2 (45-78)	51.0 (22-63)	55.0 (44-69)	40.2 (22-51)
Average % pair-wise mis-matching concordance (range)	N/A	92.8 (72-99)	95.0 (84-99)	94.0 (77-99)	92.3 (70-99)	94.0 (76-100)	95.2 (84-99)	95.0 (84-99)

[a] according to (22); [b] according to (7); [c] according to (21); [d] according to (9); [e] this study; N/A not applicable

Table 4. Sensitivity and specificity (and positive and negative predictive values) of the IS*6110*-RFLP and respective MIRU-VNTR genotyping methods for their ability to correctly identify an independently genotyped drug-resistant cluster characterized by unique mutations in the *katG*, *rpoB*, *embB* and *rrs* genes

			MIRU-VNTR Locus Combinations						
		IS*6110*-RFLP	**12-MIRU**[a]	**12-MIRU + ETR A, B, C**[b]	**12-MIRU + hyper-variable loci**[e]	**15-MIRU-VNTR**[c]	**15-MIRU-VNTR + hyper-variable loci**[d]	**24-MIRU-VNTR**[c]	**24-MIRU-VNTR + hyper-variable loci**[e]
GD=0	**Sensitivity**	0.83 0.66 - 0.93	1.00 0.90 - 1.00	0.97 0.85 - 1.00	0.80 0.63 - 0.92	0.91 0.77 - 0.98	0.74 0.57 - 0.88	0.91 0.77 - 0.98	0.74 0.57 - 0.88
	Specificity	1.00 0.48 - 1.00	0.20 0.01 - 0.72	0.20 0.01 - 0.72	0.20 0.01 - 0.72	0.40 0.05 - 0.85	0.40 0.05 - 0.85	0.40 0.05 - 0.85	0.40 0.05 - 0.85
	Positive Predictive Value	1.00 0.88 - 1.00	0.90 0.76 - 0.97	0.89 0.75 - 0.97	0.88 0.71 - 0.96	0.91 0.77 - 0.98	0.90 0.73 - 0.98	0.91 0.77 - 0.98	0.90 0.73 - 0.98
	Negative Predictive Value	0.45 0.17 - 0.77	1.00 0.03 - 1.00	0.50 0.01 - 0.99	0.13 0.00 - 0.53	0.40 0.05 - 0.85	0.18 0.02 - 0.52	0.400 0.0534 - 0.85	0.18 0.02 - 0.52
GD=1	**Sensitivity**	0.94 0.81 - 0.99	1.00 0.72 – 1.00	1.00 0.72 – 1.00	0.97 0.85 - 1.00	1.00 0.72 – 1.00	0.94 0.81 - 0.99	1.00 0.72 – 1.00	0.74 0.57 - 0.88
	Specificity	1.00 0.48 - 1.00	0.00 0.00 - 0.54	0.00 0.00 - 0.54	0.00 0.00 - 0.52	0.00 0.00 - 0.537	0.00 0.00 - 0.52	0.00 0.00 - 0.54	0.40 0.05 - 0.85
	Positive Predictive Value	1.00 0.89 - 1.00	0.88 0.72 – 0.95	0.88 0.72 – 0.95	0.87 0.73 - 0.96	0.88 0.72 – 0.95	0.87 0.72 - 0.96	0.88 0.72 – 0.95	0.90 0.73 - 0.98
	Negative Predictive Value	0.71 0.29 - 0.96	ND	ND	0.00 0.00 - 0.98	ND	0.00 0.00 - 0.84	ND	0.18 0.02 - 0.52

Sensitivity, specificity and predictive values are given with their 95% confidence intervals

[a] according to (22); [b] according to (7); [c] according to (21); [d] according to (9); [e] this study; ND: not determinable

DISCUSSION

IS*6110*-RFLP genotyping is the most widely used genotyping method used to investigate and understand the epidemiology of *M. tuberculosis* (27). However, studies comparing IS*6110*-RFLP molecular epidemiological and contact tracing data, have questioned the validity of the definition of transmission (3,26). In order to address these concerns MIRU-VNTR genotyping using either 15- or 24-MIRU-VNTR loci combinations have been extensively evaluated as the new genotyping standard for molecular epidemiological studies of *M. tuberculosis* (21). Concordance between MIRU-VNTR genotyping and contact tracing data was found to be superior to that of IS*6110*-RFLP in low incidence settings (15,18). However, these MIRU-VNTR locus combinations have not been fully tested in geographical regions where TB is endemic or within a robust *M. tuberculosis* phylogeny. Our results confirm previous findings (9,15,18,21) which have suggested that MIRU-VNTR genotyping, using carefully-selected locus combinations, could be used to estimate the extent of recent or ongoing transmission. The inclusion of the 3 hyper-variable loci improved the discriminatory power of the MIRU-VNTR genotyping method in this Beijing lineage, thereby supporting a previous suggestion for their inclusion (9). However, the use of these loci needs further evaluation in other evolutionary lineages as difficulties associated with amplification reproducibility have been reported (14,21,36).

We conclude that the PCR-based MIRU-VNTR genotyping method could be applied as an epidemiological tool to measure the performance of a TB control program over time in a defined geographical setting. However, the observed concordance in the estimate of recent and ongoing transmission when using the IS*6110*-RFLP or MIRU-VNTR genotyping methods was only coincidental. A subsequent analysis of the MIRU-VNTR data, in comparison to the IS*6110*-RFLP genotyping data, revealed that the classification of a strain according to its genotype differed significantly depending on the genotyping method used. Accordingly our study showed that the degree of matching and mismatching concordance as well as concordance among unique strains was low. This led to discordance between the transmission chains predicted by the respective genotyping methods. Matching concordance increased when genetic distance was incorporated into the clustering calculation for all of the MIRU-VNTR combinations. However, this effect was offset in the case of 12-MIRU-based markers by the concomitant reduction in mismatching concordance which was not the case for the 15- and 24-MIRU-VNTR combinations. From this, it is apparent that the additional loci included in the 15- and 24-MIRU-VNTR combinations (with or without the addition of the hyper-variable loci) improved the overall concordance of MIRU-VNTR with respect to IS*6110*-RFLP. This may be due to these loci being inherently less stable and therefore more informative. However, a caveat to the inclusion of genetic distance in the clustering formula is that epidemiologically unrelated cases may be incorrectly linked within a transmission chain.

Our analysis of the drug-resistant cluster to elucidate which of the genotyping methods provided the most accurate reflection of the epidemiology, highlighted shortcomings of both the IS*6110*-RFLP and MIRU-VNTR genotyping methods. This analysis supported a previous study which demonstrated that ongoing transmission was characterized by the evolution of variant IS*6110*-RFLP genotypes while simultaneously preserving existing genotypes (33). A similar observation was found when using the different MIRU-VNTR locus combinations. This could be explained by the fact that the evolution of different loci could take place both convergently and divergently. Together, these results substantiate previous findings which have suggested that the definition of ongoing transmission according to IS*6110*-RFLP or MIRU-VNTR genotyping should include closely related genotypes (18,24,35). However, when allowing for single MIRU-VNTR changes within the definition of a cluster, the MIRU-VNTR genotyping method collapsed many of the sublineage 5 isolates into a limited number of clusters. As a result, most of the isolates were grouped as resistant, giving the method a high sensitivity, but in doing so, compromising specificity. In contrast, the identification of isolates within the drug-resistant cluster was largely retained by IS*6110*-RFLP despite the inclusion of genetic distance. This suggests that IS*6110*-RFLP in combination with genetic distance provides a more accurate reflection of ongoing transmission of this MDR-TB outbreak in this setting. This finding is important for the interpretation of molecular epidemiological data in settings where contact tracing is extremely difficult. However we acknowledge that the concordance between IS*6110*-RFLP and transmission needs further investigation in different settings and in *M. tuberculosis* strains with different genetic backgrounds.

Our results differ from previous studies (15,18) which have demonstrated a close correlation between IS*6110*-RFLP and MIRU-VNTR genotyping. These studies were conducted in low incidence Western European settings where the TB epidemic is primarily driven by reactivation and immigration (5). In these settings, efficient TB control programs would largely prevent recent and ongoing transmission and the subsequent generation of closely related clonal variants. Thus genetic diversity is predicted to be preserved. In most instances, this would imply that the strains cultured from TB cases would be genetically distantly related and thus would not share either IS*6110*-RFLP banding patterns or MIRU-VNTR genotypes. Accordingly, MIRU-VNTR genotyping would discriminate strains at a level similar to that of IS*6110*-RFLP genotyping. In contrast, our high incidence setting has promoted the evolution of a large number of genetically closely related strains which are maintained within the host population. The genetic distance between these strains is often of such a nature that strains either have identical IS*6110*-RFLP genotypes and variant MIRU-VNTR genotypes or *vice versa*. Accordingly, we hypothesize that the degree of discordance between IS*6110*-RFLP and MIRU-VNTR genotyping is dependent on the genetic distance between isolates. This is supported by the observation that distantly related isolates from the different Beijing sublineages have evolved distinct IS*6110*-RFLP and MIRU-VNTR genotypes.

In summary, we conclude that both IS*6110*-RFLP and MIRU-VNTR genotyping methods have limitations in defining chains of transmission of Beijing genotype *M. tuberculosis* strains in this high incidence setting.

ACKNOWLEDGEMENTS

The Harry Crossley Foundation and the European Commission 6th framework programme on research technological development demonstration (project no. 037919) are thanked for financially supporting this project.

Reference List

1. **Alland, D., G. E. Kalkut, A. R. Moss, R. A. McAdam, J. A. Hahn, W. Bosworth, E. Drucker, and B. R. Bloom**. 1994. Transmission of tuberculosis in New York City. An analysis by DNA fingerprinting and conventional epidemiologic methods. N.Engl.J.Med. **330**:1710-1716.

2. **Bifani, P. J., B. B. Plikaytis, V. Kapur, K. Stockbauer, X. Pan, M. L. Lutfey, S. L. Moghazeh, W. Eisner, T. M. Daniel, M. H. Kaplan, J. T. Crawford, J. M. Musser, and B. N. Kreiswirth.** 1996. Origin and interstate spread of a New York City multidrug-resistant *Mycobacterium tuberculosis* clone famil. JAMA 275:452-457.

3. **Cowan, L. S., L. Diem, T. Monson, P. Wand, D. Temporado, T. V. Oemig, and J. T. Crawford**. 2005. Evaluation of a two-step approach for large-scale, prospective genotyping of *Mycobacterium tuberculosis* isolates in the United States. J.Clin.Microbiol. **43**:688-695.

4. **Cruz-Ferro, E. and E. Fernandez-Nogueira**. 2007. Epidemiology of tuberculosis in Galicia, Spain, 1996-2005. Int.J.Tuberc.Lung Dis. **11**:1073-1079.

5. **Dahle, U. R., V. Eldholm, B. A. Winje, T. Mannsaker, and E. Heldal**. 2007. Impact of immigration on the molecular epidemiology of *Mycobacterium tuberculosis* in a low-incidence country. Am.J.Respir.Crit Care Med. **176**:930-935.

6. **Gagneux, S., K. DeRiemer, T. Van, M. Kato-Maeda, B. C. de Jong, S. Narayanan, M. Nicol, S. Niemann, K. Kremer, M. C. Gutierrez, M. Hilty, P. C. Hopewell, and P. M. Small.** 2006. Variable host-pathogen compatibility in *Mycobacterium tuberculosis*. Proc.Natl.Acad.Sci.U.S.A 103:2869-2873.

7. **Gibson, A., T. Brown, L. Baker, and F. Drobniewski**. 2005. Can 15-locus mycobacterial interspersed repetitive unit-variable-number tandem repeat analysis provide insight into the evolution of *Mycobacterium tuberculosis*? Appl.Environ.Microbiol. **71**:8207-8213.

8. **Hanekom, M., G. D. van der Spuy, E. Streicher, S. L. Ndabambi, C. R. McEvoy, M. Kidd, N. Beyers, T. C. Victor, P. D. van Helden, and R. M. Warren**. 2007. A recently evolved sublineage of the *Mycobacterium tuberculosis* Beijing strain family is associated with an increased ability to spread and cause disease. J.Clin.Microbiol. **45**:1483-1490.

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

10. **Jiao, W. W., I. Mokrousov, G. Z. Sun, Y. J. Guo, A. Vyazovaya, O. Narvskaya, and A. D. Shen**. 2008. Evaluation of new variable-number tandem-repeat systems for typing *Mycobacterium tuberculosis* with Beijing genotype isolates from Beijing, China. J.Clin.Microbiol. 46:1045-1049.

11. **Kam, K. M., C. W. Yip, L. W. Tse, K. L. Leung, K. L. Wong, W. M. Ko, and W. S. Wong**. 2006. Optimization of variable number tandem repeat typing set for differentiating *Mycobacterium tuberculosis* strains in the Beijing family. FEMS Microbiol.Lett. 256:258-265.

12. **Kam, K. M., C. W. Yip, L. W. Tse, K. L. Wong, T. K. Lam, K. Kremer, B. K. Au, and D. van Soolingen**. 2005. Utility of mycobacterial interspersed repetitive unit typing for differentiating multidrug-resistant *Mycobacterium tuberculosis* isolates of the Beijing family. J.Clin.Microbiol. 43:306-313.

13. **Kamerbeek, J., L. Schouls, A. Kolk, M. van Agterveld, D. van Soolingen, S. Kuijper, A. Bunschoten, H. Molhuizen, R. Shaw, M. Goyal, and J. Van Embden.** 1997. Simultaneous detection and strain differentiation of *Mycobacterium tuberculosis* for diagnosis and epidemiology. J.Clin.Microbiol. 35:907-914.

14. **Kremer, K., C. Arnold, A. Cataldi, M. C. Gutierrez, W. H. Haas, S. Panaiotov, R. A. Skuce, P. Supply, A. G. van der Zanden, and D. van Soolingen.** 2005. Discriminatory power and reproducibility of novel DNA typing methods for *Mycobacterium tuberculosis* complex strains. J.Clin.Microbiol. 43:5628-5638.

15. **Allix-Beguec, C., M. Fauville-Dufaux, and P. Supply**. 2008. Three-year population-based evaluation of standardized mycobacterial interspersed repetitive-unit-variable-number tandem-repeat typing of *Mycobacterium tuberculosis*. J.Clin.Microbiol. 46:1398-1406.

16. **Lopez-Calleja, A. I., M. A. Lezcano, M. A. Vitoria, M. J. Iglesias, A. Cebollada, C. Lafoz, P. Gavin, L. Aristimuno, M. J. Revillo, C. Martin, and S. Samper.** 2007. Genotyping of *Mycobacterium tuberculosis* over two periods: a changing scenario for tuberculosis transmission. Int.J.Tuberc.Lung Dis. 11:1080-1086.

17. **Nei, M.** 1978. Estimation of Average Heterozygosity and Genetic Distance from a Small Number of Individuals. Genetics 89:583-590.

18. **Oelemann, M. C., R. Diel, V. Vatin, W. Haas, S. Rusch-Gerdes, C. Locht, S. Niemann, and P. Supply**. 2007. 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. 45:691-697.

19. **Pretorius, G. S., P. D. van Helden, F. Sirgel, K. D. Eisenach, and T. C. Victor**. 1995. Mutations in katG gene sequences in isoniazid-resistant clinical isolates of *Mycobacterium tuberculosis* are rare. Antimicrob.Agents Chemother. 39:2276-2281.

20. **Small, P. M., P. C. Hopewell, S. P. Singh, A. Paz, J. Parsonnet, D. C. Ruston, G. F. Schecter, C. L. Daley, and G. K. Schoolnik**. 1994. The epidemiology of tuberculosis in San Francisco. A population-based study using conventional and molecular methods. N.Engl.J.Med. 330:1703-1709.

21. **Supply, P., C. Allix, S. Lesjean, M. Cardoso-Oelemann, S. Rusch-Gerdes, E. Willery, E. Savine, H. P. de, D. H. van, S. Roring, P. Bifani, N. Kurepina, B. Kreiswirth, C. Sola, N. Rastogi, V. Vatin, M. C. Gutierrez, M. Fauville, S. Niemann, R. Skuce, K. Kremer, C. Locht, and D. van Soolingen**. 2006. Proposal for standardization of optimized mycobacterial interspersed repetitive unit-variable-number tandem repeat typing of *Mycobacterium tuberculosis*. J.Clin.Microbiol. 44:4498-4510.

22. **Supply, P., S. Lesjean, E. Savine, K. Kremer, D. van Soolingen, and C. Locht**. 2001. Automated high-throughput genotyping for study of global epidemiology of *Mycobacterium tuberculosis* based on mycobacterial interspersed repetitive units. J.Clin.Microbiol. 39:3563-3571.

23. **Supply, P., E. Mazars, S. Lesjean, V. Vincent, B. Gicquel, and C. Locht**. 2000. Variable human minisatellite-like regions in the *Mycobacterium tuberculosis* genome. Mol.Microbiol. 36:762-771.

24. **van der Spuy, G. D., R. M. Warren, M. Richardson, N. Beyers, M. A. Behr, and P. D. van Helden**. 2003. Use of genetic distance as a measure of ongoing transmission of *Mycobacterium tuberculosis.* J.Clin.Microbiol. 41:5640-5644.

25. **van der Zanden, A. G., E. M. Te Koppele-Vije, B. N. Vijaya, D. van Soolingen, and L. M. Schouls**. 2003. Use of DNA extracts from Ziehl-Neelsen-stained slides for molecular detection of rifampin resistance and spoligotyping of *Mycobacterium tuberculosis*. J.Clin.Microbiol. 41:1101-1108.

26. **van Deutekom H., P. Supply, P. E. de Haas, E. Willery, S. P. Hoijng, C. Locht, R. A. Coutinho, and S. D. van**. 2005. Molecular typing of *Mycobacterium tuberculosis* by mycobacterial interspersed repetitive unit-variable-number tandem repeat analysis, a more accurate method for identifying epidemiological links between patients with tuberculosis. J.Clin.Microbiol. 43:4473-4479.

27. **van Embden, J. D., M. D. Cave, J. T. Crawford, J. W. Dale, K. D. Eisenach, B. Gicquel, P. Hermans, C. Martin, R. McAdam, and T. M. Shinnick.** 1993. Strain identification of

Mycobacterium tuberculosis by DNA fingerprinting: recommendations for a standardized methodology [see comments]. J.Clin.Microbiol. 31:406-409.

28. **van Rie, A., R. Warren, M. Richardson, T. C. Victor, R. P. Gie, D. A. Enarson, N. Beyers, and P. D. van Helden**. 1999. Exogenous reinfection as a cause of recurrent tuberculosis after curative treatment [see comments]. N.Engl.J.Med. 341:1174-1179.

29. **van Rie, A., R. M. Warren, N. Beyers, R. P. Gie, C. N. Classen, M. Richardson, S. L. Sampson, T. C. Victor, and P. D. van Helden.** 1999. Transmission of a multidrug-resistant *Mycobacterium tuberculosis* strain resembling "strain W" among noninstitutionalized, human immunodeficiency virus-seronegative patients. J.Infect.Dis. 180:1608-1615.

30. **van Soolingen, D., L. Qian, P. E. de Haas, J. T. Douglas, H. Traore, F. Portaels, H. Z. Qing, D. Enkhsaikan, P. Nymadawa, and J. D. van Embden.** 1995. Predominance of a single genotype of *Mycobacterium tuberculosis* in countries of east Asia. J.Clin.Microbiol. 33:3234-3238.

31. **Verver, S., R. M. Warren, Z. Munch, E. Vynnycky, P. D. van Helden, M. Richardson, G. D. van der Spuy, D. A. Enarson, M. W. Borgdorff, M. A. Behr, and N. Beyers.** 2004. Transmission of tuberculosis in a high incidence urban community in South Africa. Int.J.Epidemiol. 33:351-357.

32. **Warren, R., M. de Kock, E. Engelke, R. Myburgh, v. P. Gey, T. Victor, and P. van Helden**. 2006. Safe *Mycobacterium tuberculosis* DNA extraction method that does not compromise integrity. J.Clin.Microbiol. 44:254-256.

33. **Warren, R. M., G. D. van der Spuy, M. Richardson, N. Beyers, C. Booysen, M. A. Behr, and P. D. van Helden**. 2002. Evolution of the IS*6110*-based restriction fragment length polymorphism pattern during the transmission of *Mycobacterium tuberculosis*. Journal of Clinical Microbiology 40:1277-1282.

34. **Warren, R. M., T. C. Victor, E. M. Streicher, M. Richardson, G. D. van der Spuy, R. Johnson, V. N. Chihota, C. Locht, P. Supply, and P. D. van Helden.** 2004. Clonal expansion of a globally disseminated lineage of *Mycobacterium tuberculosis* with low IS*6110* copy numbers. J.Clin.Microbiol. 42:5774-5782.

35. **Yeh, R. W., d. L. Ponce, C. B. Agasino, J. A. Hahn, C. L. Daley, P. C. Hopewell, and P. M. Small.** 1998. Stability of *Mycobacterium tuberculosis* DNA genotypes. J.Infect.Dis. 177:1107-1111.

36. **Yokoyama, E., K. Kishida, M. Uchimura, and S. Ichinohe**. 2007. 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. 7:499-508.

Table S1. Cross-tabulation of marker genotypes

The tables below show cross-tabulations of the numbers of cases classified into transmission chains according to IS*6110* and each of the MIRU-VNTR combinations, allowing for evolution by single events in either marker.

IS*6110*-RFLP genotypes GD=1	12-MIRU genotypes (GD=1)			
	1	2	3	Total
1	219	1		220
2	1			1
3	1			1
4	1			1
5	4			4
6	33			33
7	1			1
8	9			9
9	8			8
10	1			1
11	1			1
12	1			1
13	1			1
14	1			1
15	2			2
16	1			1
17	2			2
18	1			1
19	3			3
20	1			1
21	4		1	5
22	1			1
23	1			1
24	1			1
25	1			1
26	1			1
27	2			2
28	2			2
29	1			1
30	1			1
31	2			2
32	1			1
33	2			2
34	1			1
35	1			1
36	1			1
37	1			1
38	1			1
39	1			1
40	1			1
Total	319	1	1	321

IS*6110*-RFLP genotypes GD=1	12-MIRU + ETR A, B, C (GD=1)				
	1	2	3	4	Total
1	219		1		220
2	1				1
3	1				1
4	1				1
5	4				4
6	33				33
7	1				1
8	9				9
9	8				8
10	1				1
11	1				1
12	1				1
13	1				1
14	1				1
15	2				2
16	1				1
17	2				2
18	1				1
19	3				3
20	1				1
21	4			1	5
22	1				1
23	1				1
24	1				1
25	1				1
26	1				1
27	2				2
28	2				2
29	1				1
30		1			1
31	2				2
32	1				1
33	2				2
34	1				1
35	1				1
36	1				1
37	1				1
38	1				1
39	1				1
40	1				1
Total	318	1	1	1	321

IS*6110*-RFLP genotypes GD=1	12-MIRU + hyper-variable loci (GD=1)											
	1	2	3	4	5	6	7	8	9	10	11	Total
1	216		1	1	1					1		220
2	1											1
3	1											1
4	1											1
5	4											4
6	32	1										33
7	1											1
8	9											9
9	8											8
10	1											1
11	1											1
12	1											1
13	1											1
14	1											1
15	2											2
16	1											1
17	2											2
18	1											1
19	2								1			3
20	1											1
21	4										1	5
22	1											1
23	1											1
24	1											1
25	1											1
26	1											1
27	2											2
28	2											2
29	1											1
30								1				1
31						1	1					2
32	1											1
33	2											2
34	1											1
35	1											1
36	1											1
37	1											1
38	1											1
39	1											1
40	1											1
Total	311	1	1	1	1	1	1	1	1	1	1	321

IS*6110*-RFLP genotypes GD=1	15-MIRU-VNTR (GD-1)											
	1	2	3	4	5	6	7	8	9	10	11	Total
1		219								1		220
2		1										1
3	1											1
4	1											1
5		4										4
6	33											33
7	1											1
8	9											9
9	8											8
10	1											1
11	1											1
12	1											1
13	1											1
14	1											1
15	2											2
16	1											1
17	2											2
18	1											1
19	3											3
20	1											1
21	4										1	5
22									1			1
23		1										1
24	1											1
25					1							1
26						1						1
27							2					2
28								2				2
29		1										1
30				1								1
31			2									2
32		1										1
33		2										2
34	1											1
35	1											1
36		1										1
37	1											1
38		1										1
39		1										1
40	1											1
Total	77	232	2	1	1	1	2	2	1	1	1	321

IS*6110*-RFLP genotypes GD=1	15-MIRU-VNTR + hyper-variable loci (GD=1)																				
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	Total
1				216	1	1	1												1		220
2				1																	1
3																1					1
4															1						1
5				4																	4
6	32	1																			33
7	1																				1
8	9																				9
9	8																				8
10	1																				1
11	1																				1
12	1																				1
13	1																				1
14	1																				1
15	2																				2
16	1																				1
17	2																				2
18	1																				1
19	2													1							3
20	1																				1
21	4																			1	5
22																		1			1
23				1																	1
24	1																				1
25										1											1
26											1										1
27												2									2
28													2								2
29				1																	1
30									1												1
31								2													2
32				1																	1
33				2																	2
34	1																				1
35			1																		1
36				1																	1
37																	1				1
38				1																	1
39				1																	1
40	1																				1
Total	71	1	1	229	1	1	1	2	1	1	1	2	2	1	1	1	1	1	1	1	321

IS*6110*-RFLP genotypes GD=1	24-MIRU-VNTR (GD=1)															
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Total
1		219												1		220
2		1														1
3	1															1
4	1															1
5		4														4
6	33															33
7	1															1
8	9															9
9	8															8
10								1								1
11	1															1
12	1															1
13	1															1
14							1									1
15								2								2
16	1															1
17	2															2
18	1															1
19	3															3
20								1								1
21									4						1	5
22													1			1
23		1														1
24	1															1
25						1										1
26										1						1
27											2					2
28												2				2
29		1														1
30				1												1
31			2													2
32		1														1
33		2														2
34	1															1
35	1															1
36		1														1
37	1															1
38		1														1
39		1														1
40					1											1
Total	67	232	2	1	1	1	1	4	4	1	2	2	1	1	1	321

IS*6110*-RFLP genotypes GD=1	24-MIRU-VNTR + hyper-variable loci (GD=1)																											Total
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	
1				216	1	1	1																			1		220
2				1																								1
3																						1						1
4																					1							1
5				4																								4
6	32	1																										33
7	1																											1
8									9																			9
9										8																		8
10																1												1
11	1																											1
12	1																											1
13	1																											1
14															1													1
15																2												2
16	1																											1
17	2																											2
18	1																											1
19	2												1															3
20																1												1
21																	4										1	5
22																								1				1
23				1																								1
24																									1			1
25														1														1
26																		1										1
27																			2									2
28																				2								2
29				1																								1
30											1																	1
31								2																				2
32				1																								1
33				2																								2
34	1																											1
35			1																									1
36				1																								1
37																							1					1
38				1																								1
39				1																								1
40												1																1
Total	43	1	1	229	1	1	1	2	9	8	1	1	1	1	1	4	4	1	2	2	1	1	1	1	1	1	1	321

Table S2. Sublineage 5 drug-resistance mutations and MIRU-VNTR repeat numbers

								Number of Repeats (hexadecimal Notation)																											
Isolate No	IS6110 RFLP	Inh	Rif	*katG* mutation	*rpoB* mutation	*embB*	*rrs*	SubLineage	MIRU02	MIRU04	MIRU10	MIRU16	MIRU20	MIRU23	MIRU24	MIRU26	MIRU27	MIRU31	MIRU39	MIRU40	VNTR1955	VNTR2165	QUB11b	QUB26b	VNTR0424	VNTR2401	VNTR4156	VNTR3690	ETR-C	VNTR2347	ETR-B	Mtub34	QUB3232	VNTR3820	VNTR4120
1945	551	S	R					1	2	2	3	3	2	5	1	6	3	5	3	1	4	5	5	9	4	4	4	3	4	3	2	3	D	A	9
2701	515	NT	NT					2	2	2	3	3	2	5	1	7	3	5	3	3	4	5	6	9	4	4	6	3	3	3	2	3	D	C	9
2704	515	NT	NT					2	2	2	3	3	2	5	1	7	3	5	3	3	4	5	6	9	4	4	6	3	3	3	2	3	D	C	9
4690	660	S	S					2	2	2	3	3	2	5	1	7	3	5	3	3	4	5	6	9	4	4	6	3	3	3	2	3	D	C	A
1453	907	NT	NT					3	2	2	3	3	2	5	1	7	3	5	3	3	4	5	6	9	4	4	6	3	3	3	2	3	D	C	9
4510	907	NT	NT					3	2	2	3	3	2	5	1	7	3	5	3	3	4	5	6	9	4	4	5	3	3	3	2	3	D	C	9
1176	484	NT	NT					4	2	2	3	3	2	5	1	7	3	4	3	3	6	5	5	7	4	4	4	3	3	3	2	3	F	D	B
129	213	R	R	315 (AGC to ACC)	531 (TCG to TTG)	306 (ATG to ATA)	513 (CAG to CCG)	5	2	0	3	3	2	5	1	7	3	5	3	3	5	5	4	8	4	4	3	3	3	3	2	3	E	D	A
252	213	R	R	315 (AGC to ACC)	531 (TCG to TTG)	306 (ATG to ATA)	513 (CAG to CCG)	5	2	0	3	3	2	5	1	7	3	5	3	3	5	5	4	8	4	4	3	3	3	3	2	3	E	D	A
357	213	R	R	315 (AGC to ACC)	531 (TCG to TTG)	306 (ATG to ATA)	513 (CAG to CCG)	5	2	0	3	3	2	5	1	7	3	5	3	3	5	5	4	8	4	4	3	3	3	3	2	3	E	D	A
502	213	R	R	315 (AGC to ACC)	531 (TCG to TTG)	306 (ATG to ATA)	513 (CAG to CCG)	5	2	0	3	3	2	5	1	7	3	5	3	3	5	5	4	8	4	4	3	3	3	3	2	3	8	D	2
513	213	R	R	315 (AGC to ACC)	531 (TCG to TTG)	306 (ATG to ATA)	513 (CAG to CCG)	5	2	0	3	3	2	5	1	7	3	5	3	3	5	5	4	8	4	4	3	3	3	3	2	3	E	D	A
553	213	R	R	315 (AGC to ACC)	531 (TCG to TTG)	306 (ATG to ATA)	513 (CAG to CCG)	5	2	0	3	3	2	5	1	7	3	5	3	3	5	5	4	8	4	4	3	3	3	3	2	3	E	D	A
698	213	R	R	315 (AGC to ACC)	531 (TCG to TTG)	306 (ATG to ATA)	513 (CAG to CCG)	5	2	0	3	3	2	5	1	7	3	5	3	3	5	5	4	8	4	4	3	3	3	3	2	3	E	D	A
785	213	R	R	315 (AGC to ACC)	531 (TCG to TTG)	306 (ATG to ATA)	513 (CAG to CCG)	5	2	0	3	3	2	5	1	7	3	5	3	3	5	5	4	8	4	4	3	3	3	3	2	3	E	D	A
1028	213	R	R	315 (AGC to ACC)	531 (TCG to TTG)	306 (ATG to ATA)	513 (CAG to CCG)	5	2	0	3	3	2	5	1	7	3	5	3	3	5	5	4	8	4	4	3	3	3	3	2	3	E	D	A
1098	213	R	R	315 (AGC to ACC)	531 (TCG to TTG)	306 (ATG to ATA)	513 (CAG to CCG)	5	2	0	3	3	2	5	1	7	3	5	3	3	5	5	4	8	4	4	3	3	3	3	2	3	E	D	A
1154	213	R	R	315 (AGC to ACC)	531 (TCG to TTG)	306 (ATG to ATA)	513 (CAG to CCG)	5	2	0	3	3	2	5	1	7	3	5	3	3	5	5	4	8	4	4	3	3	3	3	2	3	E	D	A
1281	213	R	R	315 (AGC to ACC)	531 (TCG to TTG)	306 (ATG to ATA)	513 (CAG to CCG)	5	2	0	3	3	2	5	1	7	3	5	3	3	5	5	4	8	4	4	3	3	3	3	2	3	E	D	A
1306	213	R	R	315 (AGC to ACC)	531 (TCG to TTG)	306 (ATG to ATA)	513 (CAG to CCG)	5	2	0	3	3	2	5	1	7	3	5	3	3	5	5	4	8	4	4	3	3	3	3	2	3	3	D	A
1464	213	R	R	315 (AGC to ACC)	531 (TCG to TTG)	306 (ATG to ATA)	513 (CAG to CCG)	5	2	0	3	3	2	5	1	7	3	5	3	3	5	5	4	8	4	4	3	3	3	3	2	3	E	D	A
1909	213	R	R	315 (AGC to ACC)	531 (TCG to TTG)	306 (ATG to ATA)	513 (CAG to CCG)	5	2	0	3	3	2	5	1	7	3	5	3	3	5	5	4	8	4	4	3	3	3	3	2	3	E	D	A
2024	213	R	R	315 (AGC to ACC)	531 (TCG to TTG)	306 (ATG to ATA)	513 (CAG to CCG)	5	2	0	3	3	2	5	1	7	3	5	3	3	5	5	4	8	4	4	3	3	3	3	2	3	E	D	A
2074	213	R	R	315 (AGC to ACC)	531 (TCG to TTG)	306 (ATG to ATA)	513 (CAG to CCG)	5	2	0	3	3	2	5	1	7	3	5	3	3	5	5	4	8	4	4	3	3	1	3	2	3	E	D	A
2173	213	R	R	315 (AGC to ACC)	531 (TCG to TTG)	306 (ATG to ATA)	513 (CAG to CCG)	5	2	0	3	3	2	5	1	7	3	5	3	3	5	5	4	8	4	4	3	3	3	3	2	3	3	D	A
2236	213	R	R	315 (AGC to ACC)	531 (TCG to TTG)	306 (ATG to ATA)	513 (CAG to CCG)	5	2	0	3	3	2	5	1	7	3	5	3	3	5	5	4	8	4	4	3	3	3	3	2	3	E	D	9
2600	213	R	R	315 (AGC to ACC)	531 (TCG to TTG)	306 (ATG to ATA)	513 (CAG to CCG)	5	2	0	3	3	2	5	1	7	3	5	3	3	5	5	4	8	4	4	3	3	3	3	2	3	3	D	A
3096	213	R	R	315 (AGC to ACC)	531 (TCG to TTG)	306 (ATG to ATA)	513 (CAG to CCG)	5	2	0	3	3	2	5	1	7	3	5	3	3	5	5	4	8	4	4	3	3	3	3	2	3	E	D	A

								Number of Repeats (hexadecimal Notation)																											
Isolate No	IS6110 RFLP	Inh	Rif	*katG* mutation	*rpoB* mutation	*embB*	*rrs*	SubLineage	MIRU02	MIRU04	MIRU10	MIRU16	MIRU20	MIRU23	MIRU24	MIRU26	MIRU27	MIRU31	MIRU39	MIRU40	VNTR1955	VNTR2165	QUB11b	QUB26b	VNTR0424	VNTR2401	VNTR4156	VNTR3690	ETR-C	VNTR2347	ETR-B	Mtub34	QUB3232	VNTR3820	VNTR4120
3111	213	R	R	315 (AGC to ACC)	531 (TCG to TTG)	306 (ATG to ATA)	513 (CAG to CCG)	5	2	0	3	3	2	5	1	7	3	5	3	3	5	5	4	8	4	4	3	3	3	3	2	3	E	D	A
3368	213	R	R	315 (AGC to ACC)	531 (TCG to TTG)	306 (ATG to ATA)	513 (CAG to CCG)	5	2	0	3	3	2	5	1	7	3	5	3	3	5	5	4	8	4	4	3	3	3	3	2	3	E	D	A
3568	213	R	R	315 (AGC to ACC)	531 (TCG to TTG)	306 (ATG to ATA)	513 (CAG to CCG)	5	2	0	3	3	2	5	1	7	3	5	3	3	5	5	4	8	4	4	3	3	3	3	2	3	E	D	A
3797	213	R	R	315 (AGC to ACC)	531 (TCG to TTG)	306 (ATG to ATA)	513 (CAG to CCG)	5	2	0	3	3	2	5	1	7	3	5	3	3	5	5	4	8	4	4	3	3	3	3	2	3	E	D	A
4344	213	R	R	315 (AGC to ACC)	531 (TCG to TTG)	306 (ATG to ATA)	513 (CAG to CCG)	5	2	0	3	3	2	5	1	7	3	5	3	3	5	5	4	8	4	4	3	3	3	3	2	3	E	D	A
4414	213	R	R	315 (AGC to ACC)	531 (TCG to TTG)	306 (ATG to ATA)	513 (CAG to CCG)	5	2	0	3	3	2	5	1	7	3	5	3	3	5	5	4	8	4	2	3	3	3	3	2	3	E	D	A
5008	213	R	R	315 (AGC to ACC)	531 (TCG to TTG)	306 (ATG to ATA)	513 (CAG to CCG)	5	2	0	3	3	2	5	1	7	3	5	3	3	5	5	4	8	4	4	3	3	3	3	2	3	E	D	A
5516	213	R	R	315 (AGC to ACC)	531 (TCG to TTG)	306 (ATG to ATA)	513 (CAG to CCG)	5	2	0	3	3	2	5	1	7	3	5	3	3	5	5	4	8	4	4	3	3	3	3	2	3	E	D	A
953	218	S	S	WT	WT	not tested	not tested	5	2	3	3	3	2	5	1	7	3	5	3	3	5	5	3	8	4	4	3	3	3	3	2	3	C	D	A
1646	240	S	S	WT	WT	not tested	not tested	5	2	0	3	3	2	5	1	7	3	5	3	3	5	5	4	8	4	4	3	3	3	3	2	3	E	D	A
819	294	S	S	WT	WT	not tested	not tested	5	2	0	3	3	2	5	1	7	3	5	3	3	5	5	4	8	4	4	3	3	3	3	2	3	E	D	A
1149	316	S	R	WT	526 (CAC to TAC)	not tested	not tested	5	2	0	3	3	2	5	1	7	3	5	3	3	5	5	3	8	4	4	3	3	3	3	2	3	E	D	A
1944	405	R	R	315 (AGC to ACC)	531 (TCG to TTG)	306 (ATG to ATA)	513 (CAG to CCG)	5	2	0	3	3	2	5	1	7	3	5	3	3	5	5	4	8	4	2	3	3	3	3	2	3	C	3	A
2125	535	R	R	315 (AGC to ACC)	531 (TCG to TTG)	306 (ATG to ATA)	513 (CAG to CCG)	5	2	0	3	3	2	5	1	7	3	5	3	3	5	5	4	8	4	4	3	3	3	3	2	3	E	D	A
2147	535	R	R	315 (AGC to ACC)	531 (TCG to TTG)	306 (ATG to ATA)	513 (CAG to CCG)	5	2	0	3	3	2	5	1	7	3	5	3	3	5	5	4	8	4	4	3	3	3	3	2	3	E	D	A
3556	606	R	R	315 (AGC to ACC)	531 (TCG to TTG)	306 (ATG to ATA)	513 (CAG to CCG)	5	2	0	3	3	2	5	1	7	3	5	3	3	5	5	4	8	4	4	3	3	3	3	2	3	E	D	A
3908	667	R	R	315 (AGC to ACC)	531 (TCG to TTG)	306 (ATG to ATA)	513 (CAG to CCG)	5	2	0	3	3	2	5	1	7	3	5	3	3	5	5	4	8	4	4	3	3	3	3	2	3	E	D	A
4158	716	R	R	315 (AGC to ACC)	531 (TCG to TTG)	306 (ATG to ATA)	513 (CAG to CCG)	5	2	0	3	3	2	5	1	7	3	5	3	3	5	5	4	8	4	4	3	3	3	3	2	3	B	3	7
4570	810	NT	NT	WT	WT	not tested	not tested	5	2	0	3	3	2	5	1	7	3	5	3	3	5	5	4	8	4	4	3	3	3	3	2	3	E	D	A
1125	205	S	S					6	2	2	3	3	2	5	1	7	3	5	3	3	5	5	5	6	4	4	3	3	3	3	2	3	D	D	3
407	206	S	S					6	2	2	3	3	2	5	1	7	3	5	3	3	5	5	5	6	4	4	3	3	3	3	2	3	D	D	3
261	219	NT	NT					6	2	2	3	3	2	5	1	6	3	5	3	3	5	4	5	8	4	4	3	3	3	3	2	3	D	D	A
354	219	S	S					6	2	2	3	3	2	5	1	8	3	5	3	3	5	5	5	8	4	4	3	3	3	3	2	3	D	D	A
1307	219	NT	NT					6	2	2	3	3	2	5	1	6	3	5	3	3	5	5	5	8	4	4	3	3	3	3	2	3	D	D	A
1641	219	NT	NT					6	2	2	3	3	2	5	1	6	3	5	3	3	5	5	5	8	4	4	3	3	3	3	2	3	D	D	A
5573	219	S	S					6	2	2	3	3	2	5	1	5	3	5	3	3	5	5	5	8	4	4	3	3	3	3	2	3	C	D	A
954	220	S	S					6	2	2	3	3	2	5	1	8	3	5	3	3	5	4	5	8	4	4	3	3	3	3	2	3	D	D	A
1716	220	S	R					6	2	2	3	3	2	5	1	8	3	5	3	3	5	5	4	8	4	4	3	3	3	3	2	3	D	D	A
1973	220	NT	NT					6	2	2	3	3	2	5	1	8	3	5	3	3	5	5	5	8	4	4	3	3	3	3	2	3	D	D	A
2332	220	NT	NT					6	2	2	3	3	2	5	1	8	3	5	3	3	5	5	5	8	4	2	3	3	3	3	2	3	D	C	A

								Number of Repeats (hexadecimal Notation)																											
Isolate No	IS6110 RFLP	Inh	Rif	*katG* mutation	*rpoB* mutation	*embB*	*rrs*	SubLineage	MIRU02	MIRU04	MIRU10	MIRU16	MIRU20	MIRU23	MIRU24	MIRU26	MIRU27	MIRU31	MIRU39	MIRU40	VNTR1955	VNTR2165	QUB11b	QUB26b	VNTR0424	VNTR2401	VNTR4156	VNTR3690	ETR-C	VNTR2347	ETR-B	Mtub34	QUB3232	VNTR3820	VNTR4120
3166	220	R	S					6	2	2	3	3	2	5	1	8	3	5	3	3	5	5	5	8	4	4	3	3	3	3	2	3	D	D	A
4036	220	R	S					6	2	2	3	3	2	5	1	8	3	5	3	3	5	4	5	8	4	4	3	3	3	3	2	3	C	D	A
4104	220	NT	NT					6	2	2	3	3	2	5	1	8	3	5	3	3	5	5	5	8	4	4	3	3	3	3	2	3	C	D	A
4473	220	R	NT					6	2	2	3	3	2	5	1	8	3	5	3	3	5	5	5	8	4	4	3	3	3	3	2	3	D	D	A
4493	220	NT	NT					6	2	2	3	3	2	5	1	8	3	5	3	3	5	5	5	8	4	4	3	3	3	3	2	3	C	D	A
4576	220	R	S					6	2	2	3	3	2	5	1	8	3	5	3	3	5	5	5	8	4	4	3	3	3	3	2	3	C	D	2
4880	220	NT	NT					6	2	2	3	3	2	5	1	8	3	5	3	3	5	5	5	8	4	4	3	3	3	3	2	3	C	D	A
1189	223	R	S					6	2	2	3	3	2	5	1	7	3	5	4	4	5	5	1	8	4	4	3	3	3	1	2	3	D	D	A
318	307	S	S					6	2	2	3	3	2	5	1	7	3	5	4	3	5	5	1	8	4	4	3	3	3	1	2	3	D	D	A
1507	307	NT	NT					6	2	2	3	3	2	5	1	7	3	5	4	3	5	5	1	8	4	4	3	3	3	1	2	3	D	D	A
204	319	S	S					6	2	2	3	3	2	5	1	7	2	5	3	3	5	5	4	8	4	4	3	3	3	3	2	3	E	3	9
919	319	S	S					6	2	2	3	3	2	5	1	7	3	5	3	3	5	5	4	8	4	4	3	3	3	3	2	3	3	D	A
2387	319	S	S					6	2	2	3	3	2	5	1	7	3	5	3	3	5	5	4	8	4	2	3	3	3	3	2	3	D	D	9
3514	319	NT	NT					6	2	2	3	3	2	5	1	7	3	5	3	3	5	5	4	8	4	4	3	3	3	3	2	3	D	3	9
700	322	NT	NT					6	2	2	3	3	2	5	1	8	3	5	3	3	5	5	5	8	4	4	3	3	3	3	2	3	D	D	B
1844	323	R	NT					6	2	2	3	3	2	5	1	7	3	5	4	3	5	5	1	8	4	4	3	3	3	1	2	3	D	D	A
3027	323	NT	NT					6	2	2	3	3	2	5	1	7	3	5	4	3	5	4	1	8	4	4	3	4	3	1	2	3	B	D	A
4456	323	S	S					6	2	2	3	3	2	5	1	7	3	5	4	3	5	5	1	8	4	4	3	3	3	1	2	3	C	D	A
1696	343	NT	NT					6	2	2	3	3	2	5	1	7	3	5	4	3	5	5	1	8	4	4	3	3	3	1	2	3	D	D	A
1823	345	NT	NT					6	2	2	3	3	2	5	1	7	3	5	3	3	8	5	5	8	4	4	3	3	3	3	2	3	D	D	9
1892	402	NT	NT					6	2	2	3	3	2	5	1	7	4	5	3	3	5	5	4	8	4	4	3	3	3	3	2	3	B	D	A
3255	618	NT	NT					6	2	2	4	3	2	6	1	5	3	3	2	3	4	5	3	7	4	4	3	1	3	3	2	1	3	8	2
3361	618	S	S					6	2	2	3	3	2	5	1	7	3	5	2	3	5	5	1	8	4	4	3	3	3	1	2	3	D	D	A
3654	618	NT	NT					6	2	2	3	3	2	5	1	7	3	5	4	3	5	4	1	8	4	4	3	3	3	1	2	3	D	D	A
3257	650	NT	NT					6	2	2	3	2	2	5	1	7	3	5	3	3	5	3	3	8	4	4	3	3	3	3	2	3	C	D	8
4126	650	NT	NT					6	2	2	3	2	2	5	1	7	3	5	3	3	5	5	3	8	4	4	3	3	3	3	2	3	D	D	8
3518	669	R	S					6	2	2	3	3	2	5	1	7	2	5	3	3	5	4	4	8	4	4	3	3	3	3	2	3	3	8	9
3954	670	NT	NT					6	2	2	3	3	2	5	1	7	3	5	3	3	5	5	4	8	4	4	3	3	3	3	2	3	D	D	9
3047	823	NT	NT					6	2	2	3	3	2	5	1	7	3	5	3	3	5	5	4	6	4	4	3	3	3	3	2	3	A	D	3
5610	985	NT	NT					6	2	2	3	3	2	5	1	7	2	5	2	3	5	5	4	7	4	4	3	3	3	3	2	3	D	C	A

Isolate No	IS6110 RFLP	Inh	Rif	*katG* mutation	*rpoB* mutation	*embB*	*rrs*	Number of Repeats (hexadecimal Notation)																											
								SubLineage	MIRU02	MIRU04	MIRU10	MIRU16	MIRU20	MIRU23	MIRU24	MIRU26	MIRU27	MIRU31	MIRU39	MIRU40	VNTR1955	VNTR2165	QUB11b	QUB26b	VNTR0424	VNTR2401	VNTR4156	VNTR3690	ETR-C	VNTR2347	ETR-B	Mtub34	QUB3232	VNTR3820	VNTR4120
649	2	S	S					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
600	187	S	S					7	2	2	2	3	2	5	1	7	3	3	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	9
1639	207	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
2189	207	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	3	5	6	4	4	3	4	5	3	2	3	B	D	A
4048	207	S	S					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
137	208	NT	NT					7	2	2	2	3	2	5	1	7	4	5	4	3	5	5	5	6	4	4	3	4	5	3	3	3	B	D	A
151	208	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
313	208	S	S					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
439	208	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
638	208	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	3	3	B	D	A
650	208	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	3	3	B	D	A
714	208	S	S					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	6
717	208	S	S					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
764	208	S	S					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
767	208	S	S					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
768	208	S	S					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
815	208	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	C	A
823	208	S	S					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
993	208	S	S					7	2	2	2	3	2	5	1	7	4	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
1042	208	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
1058	208	S	S					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
1130	208	S	S					7	2	2	2	3	2	5	1	7	4	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
1232	208	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
1293	208	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
1376	208	NT	NT					7	2	2	2	3	2	5	1	7	4	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
1382	208	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	N/A	N/A
1411	208	NT	NT					7	2	2	2	3	2	5	1	7	4	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
1432	208	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
1524	208	NT	NT					7	2	2	2	3	2	5	1	7	4	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
1537	208	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	5	4	3	4	5	3	2	3	B	D	A

								Number of Repeats (hexadecimal Notation)																											
Isolate No	IS6110 RFLP	Inh	Rif	*katG* mutation	*rpoB* mutation	*embB*	*rrs*	SubLineage	MIRU02	MIRU04	MIRU10	MIRU16	MIRU20	MIRU23	MIRU24	MIRU26	MIRU27	MIRU31	MIRU39	MIRU40	VNTR1955	VNTR2165	QUB11b	QUB26b	VNTR0424	VNTR2401	VNTR4156	VNTR3690	ETR-C	VNTR2347	ETR-B	Mtub34	QUB3232	VNTR3820	VNTR4120
1609	208	NT	NT					7	2	2	2	3	2	5	1	3	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
1620	208	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	8
1763	208	S	S					7	2	2	2	3	2	5	1	7	4	4	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
1779	208	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	2
1818	208	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
2011	208	S	S					7	2	2	2	3	2	5	1	7	4	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
2112	208	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
2229	208	NT	NT					7	2	2	2	3	2	5	1	7	3	3	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
2256	208	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	2	D	A
2377	208	S	S					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	C	A
2890	208	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
3222	208	S	S					7	2	2	2	3	2	5	1	7	3	5	4	3	4	5	5	6	4	4	3	4	5	3	2	3	B	D	A
3259	208	S	S					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
3305	208	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
3327	208	NT	NT					7	2	2	2	3	2	5	1	7	4	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
3339	208	R	S					7	2	2	2	3	2	5	1	7	4	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	B
3496	208	NT	NT					7	2	2	2	3	2	5	1	7	4	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	1	D	A
3641	208	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
3646	208	NT	NT					7	2	2	2	3	2	5	1	7	4	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
3688	208	S	S					7	2	2	2	3	2	5	1	8	4	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
3835	208	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
4146	208	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
4153	208	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
4174	208	NT	NT					7	2	2	2	3	2	5	1	7	4	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
4225	208	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	9
4290	208	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
4329	208	NT	NT					7	2	2	2	3	2	5	1	7	4	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
4424	208	NT	NT					7	2	2	2	3	2	5	1	7	3	3	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
4449	208	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
4551	208	S	S					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A

								Number of Repeats (hexadecimal Notation)																											
Isolate No	IS6110 RFLP	Inh	Rif	*katG* mutation	*rpoB* mutation	*embB*	*rrs*	SubLineage	MIRU02	MIRU04	MIRU10	MIRU16	MIRU20	MIRU23	MIRU24	MIRU26	MIRU27	MIRU31	MIRU39	MIRU40	VNTR1955	VNTR2165	QUB11b	QUB26b	VNTR0424	VNTR2401	VNTR4156	VNTR3690	ETR-C	VNTR2347	ETR-B	Mtub34	QUB3232	VNTR3820	VNTR4120
5016	208	S	S					7	2	2	2	3	2	5	1	7	3	5	3	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
5187	208	S	S					7	2	2	2	3	2	5	1	9	3	5	3	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
5357	208	S	S					7	2	2	2	3	2	5	1	7	4	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
212	209	S	S					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	3	3	B	D	A
215	209	S	S					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	3	3	B	D	A
284	209	NT	NT					7	2	2	2	3	2	5	1	7	3	3	4	3	5	5	5	6	4	4	3	4	5	3	3	3	B	D	A
469	209	S	S					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
495	209	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	3	3	B	D	A
599	209	R	R					7	2	2	2	3	2	5	1	7	3	3	4	3	5	5	5	6	4	4	3	4	5	3	2	3	4	D	A
605	209	S	S					7	2	2	2	3	2	5	1	7	3	3	4	3	5	5	5	6	4	4	3	4	5	3	3	3	B	D	A
618	209	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
634	209	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	3	3	B	D	A
687	209	R	R					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
715	209	S	S					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
735	209	S	S					7	2	2	2	3	2	5	1	7	3	3	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
759	209	S	S					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
795	209	R	R					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
972	209	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	2	4	5	3	2	3	B	D	A
1151	209	S	S					7	2	2	2	3	2	5	1	7	3	3	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
1158	209	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
1170	209	S	S					7	2	2	2	3	2	5	1	7	3	3	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	C	9
1182	209	R	S					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
1259	209	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
1267	209	S	S					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
1279	209	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
1326	209	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
1360	209	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
1383	209	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
1421	209	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
1427	209	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A

Isolate No	IS6110 RFLP	Inh	Rif	*katG* mutation	*rpoB* mutation	*embB*	*rrs*	Number of Repeats (hexadecimal Notation)																											
								SubLineage	MIRU02	MIRU04	MIRU10	MIRU16	MIRU20	MIRU23	MIRU24	MIRU26	MIRU27	MIRU31	MIRU39	MIRU40	VNTR1955	VNTR2165	QUB11b	QUB26b	VNTR0424	VNTR2401	VNTR4156	VNTR3690	ETR-C	VNTR2347	ETR-B	Mtub34	QUB3232	VNTR3820	VNTR4120
1547	209	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	9
1555	209	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
1575	209	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
1588	209	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
1736	209	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
1777	209	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
1780	209	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	3	D	2
1852	209	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
1871	209	S	S					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
1924	209	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
1948	209	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
1949	209	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	8
2076	209	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
2087	209	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	1	3	2	3	B	D	A
2206	209	NT	NT					7	2	2	2	3	2	5	1	7	3	3	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
2221	209	S	S					7	2	2	2	3	2	5	1	7	3	5	4	3	4	5	5	6	4	4	3	4	5	3	2	3	B	D	A
2307	209	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	A	D	A
2330	209	NT	NT					7	2	2	2	3	2	5	1	7	3	3	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
2391	209	S	S					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
2412	209	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	4	5	5	6	4	4	3	4	5	3	2	3	B	D	A
2413	209	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
2463	209	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
2598	209	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
2891	209	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	4	5	5	6	4	4	3	4	5	3	2	3	B	D	A
2903	209	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	4	5	5	6	4	4	3	4	5	3	2	3	B	D	A
3005	209	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
3033	209	S	S					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
3039	209	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
3052	209	S	S					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
3188	209	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A

								Number of Repeats (hexadecimal Notation)																											
Isolate No	IS6110 RFLP	Inh	Rif	*katG* mutation	*rpoB* mutation	*embB*	*rrs*	SubLineage	MIRU02	MIRU04	MIRU10	MIRU16	MIRU20	MIRU23	MIRU24	MIRU26	MIRU27	MIRU31	MIRU39	MIRU40	VNTR1955	VNTR2165	QUB11b	QUB26b	VNTR0424	VNTR2401	VNTR4156	VNTR3690	ETR-C	VNTR2347	ETR-B	Mtub34	QUB3232	VNTR3820	VNTR4120
3207	209	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
3314	209	S	S					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	2	2	3	B	D	A
3324	209	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
3369	209	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
3414	209	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
3416	209	S	S					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
3506	209	NT	NT					7	2	2	2	3	2	5	1	7	3	3	4	3	5	5	5	6	4	4	3	4	5	3	2	3	3	D	7
3582	209	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	3	D	8
3589	209	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
3594	209	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	4	D	A
3687	209	S	S					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
3752	209	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
3847	209	NT	NT					7	2	2	2	3	2	5	1	7	3	3	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
3885	209	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	3	4	3	4	5	3	2	3	B	D	A
3973	209	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
4002	209	S	S					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
4004	209	S	S					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
4122	209	NT	NT					7	2	2	2	3	2	5	1	7	3	3	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
4151	209	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	F	A
4186	209	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
4187	209	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
4195	209	NT	NT					7	2	2	2	3	2	5	1	7	3	3	4	3	5	5	5	6	4	4	3	4	5	3	2	3	3	7	2
4220	209	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
4221	209	S	S					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
4235	209	S	S					7	2	2	2	3	2	5	1	7	3	4	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
4271	209	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
4276	209	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
4292	209	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
4376	209	S	S					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
4395	209	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A

								Number of Repeats (hexadecimal Notation)																											
Isolate No	IS6110 RFLP	Inh	Rif	*katG* mutation	*rpoB* mutation	*embB*	*rrs*	SubLineage	MIRU02	MIRU04	MIRU10	MIRU16	MIRU20	MIRU23	MIRU24	MIRU26	MIRU27	MIRU31	MIRU39	MIRU40	VNTR1955	VNTR2165	QUB11b	QUB26b	VNTR0424	VNTR2401	VNTR4156	VNTR3690	ETR-C	VNTR2347	ETR-B	Mtub34	QUB3232	VNTR3820	VNTR4120
4430	209	S	S					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
4436	209	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
4447	209	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
4459	209	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
4463	209	NT	NT					7	2	2	2	3	2	5	1	7	3	3	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
4464	209	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
4477	209	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
4574	209	S	S					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
4942	209	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
5218	209	NT	NT					7	2	2	2	3	2	5	1	7	3	5	3	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
5240	209	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
5460	209	S	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
5527	209	NT	NT					7	2	2	2	3	2	5	1	5	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
5569	209	S	S					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
3189	211	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
4059	211	S	S					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
4896	211	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
3941	212	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
191	215	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
230	215	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	3	3	B	D	A
590	215	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
791	215	R	R					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	2	5	3
1081	215	S	S					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
1390	215	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
1396	215	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
1652	215	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
1708	215	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
3030	215	S	S					7	2	2	2	3	2	5	1	7	4	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
3114	215	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
3555	215	NT	NT					7	2	2	2	3	2	5	1	7	4	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A

								Number of Repeats (hexadecimal Notation)																											
Isolate No	IS6110 RFLP	Inh	Rif	*katG* mutation	*rpoB* mutation	*embB*	*rrs*	SubLineage	MIRU02	MIRU04	MIRU10	MIRU16	MIRU20	MIRU23	MIRU24	MIRU26	MIRU27	MIRU31	MIRU39	MIRU40	VNTR1955	VNTR2165	QUB11b	QUB26b	VNTR0424	VNTR2401	VNTR4156	VNTR3690	ETR-C	VNTR2347	ETR-B	Mtub34	QUB3232	VNTR3820	VNTR4120
4023	215	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
4166	215	S	S					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
4200	215	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
4256	215	NT	NT					7	2	2	2	3	2	5	1	7	4	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
2115	216	NT	NT					7	2	2	4	3	2	5	1	5	3	3	2	3	4	5	3	7	4	2	3	1	5	3	2	3	3	6	2
2274	216	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
2542	216	S	S					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
3915	216	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
4097	216	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
1389	217	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
1492	217	R	R					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
1821	217	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
3289	217	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
1428	255	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
1504	255	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	E	D	A
1522	255	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
1994	255	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
2958	255	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
4012	308	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
1515	310	S	S					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
2315	396	S	S					7	2	2	2	3	2	5	1	7	3	5	4	3	4	5	5	6	4	4	3	4	5	3	2	3	B	D	A
3050	411	S	S					7	2	2	2	3	2	5	1	7	4	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
2888	412	S	S					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
2895	426	S	S					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
3066	502	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
3181	502	R	R					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
2675	516	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	3	3	2	3	B	D	A
2057	537	S	S					7	2	2	2	3	2	5	1	7	3	3	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
2338	537	NT	NT					7	2	2	2	3	2	5	1	7	3	3	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
1015	548	S	S					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A

								Number of Repeats (hexadecimal Notation)																											
Isolate No	IS6110 RFLP	Inh	Rif	*katG* mutation	*rpoB* mutation	*embB*	*rrs*	SubLineage	MIRU02	MIRU04	MIRU10	MIRU16	MIRU20	MIRU23	MIRU24	MIRU26	MIRU27	MIRU31	MIRU39	MIRU40	VNTR1955	VNTR2165	QUB11b	QUB26b	VNTR0424	VNTR2401	VNTR4156	VNTR3690	ETR-C	VNTR2347	ETR-B	Mtub34	QUB3232	VNTR3820	VNTR4120
1463	548	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
793	595	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
1813	596	S	S					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
1518	598	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
4204	656	S	S					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
3815	657	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
3457	661	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	4	5	5	6	4	4	3	4	5	3	2	3	B	D	A
4352	661	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	4	5	5	6	4	4	3	4	5	3	2	3	B	D	A
3584	662	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
4006	664	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	4	5	5	6	4	4	3	4	5	3	2	3	B	D	A
3393	665	S	S					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
3486	665	S	S					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
5189	698	S	S					7	2	2	2	3	2	5	1	7	4	5	4	3	5	5	4	6	4	4	3	4	5	3	2	3	B	D	A
4013	743	S	S					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
4047	750	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
2383	773	S	S					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
3809	785	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
571	851	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
4437	871	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
5144	871	NT	NT					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
5080	878	S	S					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	2	4	3	4	5	3	2	3	B	D	A
4457	879	S	S					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A
5629	938	S	S					7	2	2	2	3	2	5	1	7	3	5	4	3	5	5	5	6	4	4	3	4	5	3	2	3	B	D	A

R = Resistant

S = Susceptible

NT = Not Tested

Programming

concordance.pl

```
# This script accepts input from 'input.txt' in the form of one isolate per row with genotypes
# using two markers. The first row contains the column headers which are the marker names.
# e.g. IS    MIRU_12
#      208   9
#      213   1
#      215   7
#      208   7
#      319   17
#      209   7
#      209   7
#      209   4
#      213   1
#
# The output is a number of pairwise concordance measures between the markers.
# Matching concordance is the proportion of isolate pairs having identical marker1 genotypes
which also have identical marker2 genotypes
# Mismatching concordance is the proportion of isolate pairs having non-identical marker marker1
# genotypes which also have non-identical marker2 genotypes
# Total concordance is the combination of of both matching and mismatching concordances.
#
open INPUT,"<input.txt" or die "Can't open file: input.txt";

$count = 0;
$labels = <INPUT>;
chomp $labels;
($marker1,$marker2) = split /\t/, $labels;  # Column labels in 1st line
while (<INPUT>) {
    chomp;
    ($is,$miru)=split /\t/;
    $rec={};
    $rec->{IS} = $is;
    $rec->{MIRU} = $miru;
    push @dataset, $rec;
}

# Set up comparison matrix
for $isolate1 ( 0 .. $#dataset){ # Iterate through list
    for $isolate2 ( $isolate1 .. $#dataset){ # Iterate through rest of list
        if ($isolate1 == $isolate2) { # Don't match with self
            $matrix[$isolate1][$isolate2] = 0;
            next; # Skip self-match
        }
        # Match IS strains -> upper-right matrix sector
        if ($dataset[$isolate1]{IS} == $dataset[$isolate2]{IS}){
            $matrix[$isolate1][$isolate2] = 1;
        } else {
            $matrix[$isolate1][$isolate2] = 0;
        }
        # Match MIRU strains -> lower-left matrix sector
        if ($dataset[$isolate1]{MIRU} == $dataset[$isolate2]{MIRU}){
            $matrix[$isolate2][$isolate1] = 1;
        } else {
            $matrix[$isolate2][$isolate1] = 0;
        }
    }
}

$isolates = $matches = 0;

for $isolate1 ( 0 .. $#matrix){ # Iterate through list
    $isolates++;
    for $isolate2 ( $isolate1 .. $#matrix){ # Iterate through rest of list
        if ($isolate1 == $isolate2) { # Don't match with self
            next;
        }
        if($matrix[$isolate1][$isolate2] == $matrix[$isolate2][$isolate1]){
            $matches++;
        }
        if ($matrix[$isolate1][$isolate2] == 1){ #IS matching pairs with matching MIRU
            $CountIScluster++;
            if ($matrix[$isolate2][$isolate1] == 1){
                    $IS_MIRUclusterMatch++;
                }
            } else { # Non matching IS with non-matching MIRU
                $CountISnoncluster++;
                if ($matrix[$isolate2][$isolate1] == 0){
                    $CountNonClusterMatch++;
                }
            }
            if ($matrix[$isolate2][$isolate1] == 1){ # MIRU matching pairs with matching IS
                $CountMIRUcluster++;
                if ($matrix[$isolate1][$isolate2] == 1){
                    $MIRU_ISclusterMatch++;
                }
            }
        }
}
open OUTFILE,">>result.txt";
print OUTFILE "$marker1 vs $marker2\n";
print OUTFILE "Isolates: $isolates\n";
$pairs = ($isolates**2-$isolates)/2;
print OUTFILE "Pairs: $pairs\n";
print OUTFILE "$marker1 matching pairs with matching $marker2 pairs: $IS_MIRUclusterMatch\n";
print OUTFILE "$marker2 matching pairs: $CountMIRUcluster\n";
print OUTFILE "$marker1 matching pairs: $CountIScluster\n";
printf OUTFILE "%s%.2f%s","$marker2 on $marker1 matching concordance: ",
$IS_MIRUclusterMatch/$CountIScluster, "\n";
printf OUTFILE "%s%.2f%s","$marker1 on $marker2 matching concordance: ",
$MIRU_ISclusterMatch/$CountMIRUcluster, "\n";
printf OUTFILE "%s%.2f%s","Mismatching concordance: ", $CountNonClusterMatch/$CountISnoncluster,
"\n";
print OUTFILE "Total Matches: $matches\n";
printf OUTFILE "%s%.2f%s", "Total Concordance: ", $matches/$pairs, "\n";
print OUTFILE "\n";
close INPUT;
close OUTFILE;
```

9

Conclusion

Molecular epidemiology has established itself as a powerful weapon in the arsenal arrayed against infectious diseases in general and tuberculosis (TB) in particular. It has provided novel insights into the structure and dynamics of the TB epidemic at both local and global levels and continues to do so as the organism evolves under the evolutionary pressures to which it is subjected such as drug intervention, novel host populations and changing host behavioural patterns.

Although numerous genetic markers have been developed for use in this field (as discussed in Chapter 1), the most widely used is still the RFLP pattern provided by the insertion element, IS*6110*. We have shown, using serial isolates from chronically secreting patients, that this marker evolves over time with a half-life in the order of 8.74 years, which is in the range which makes it useable for distinguishing between epidemiologically unrelated cases of disease while linking together those which are linked by chains of ongoing transmission (Chapter 2). This rate was split into two components: a fast rate of change (t½ = 0.57 years), observed to be occurring in the early phase of disease, soon after diagnosis, and a later, slow rate (t½ = 10.69). We interpreted these vastly different evolutionary rates to reflect two independent evolutionary phenomena. We speculated that the fast rate is likely to have been due to the presence of both pre- and post-evolved variants where the evolutionary event had occurred at some time prior to diagnosis. Such evolutionary change may be influenced by active growth or adaptation to a new host environment. The slow rate could either reflect evolution as a consequence of exposure to anti-tuberculosis drugs or liquefaction of a lesion in which a subset of strains had undergone an evolutionary event. However, in either case, a direct link between IS*6110* transposition and strain fitness remains to be demonstrated. It also became clear from our observations that both the original and evolved strains of *M. tuberculosis* may be present simultaneously within the host and that either or both may appear in sputum culture at any time, thus complicating the interpretation of molecular epidemiological data.

In further investigating the nature of IS*6110* evolution in the context of the transmission of *M. tuberculosis* within households, we found that variant strains emerged at a rate of 0.14 per source case per year (Chapter 3). The manifestation of evolution was usually closely associated with transmission of disease. This supports the notion that the environmental pressures of a new host environment may select for an evolved strain, in which IS*6110* transposition was one of a number of events, of which only it was documented. Alternatively, such altered strains may be observed simply because they evolved early in the disease phase when their relative numbers are sufficient for them to form a significant proportion of the population. It is quite possible that many more evolutionary events occur within the bacterial population which are never observed as a result of their low representation. The rate of variant strain production implies that the overall strain population will change at a rate of 2.9% per annum. Conventional understanding of epidemiological linkage as being defined by clusters of identical strains will, therefore, underestimate the level of transmission as variant strains appear in the population without the loss of their progenitor genotypes.

The implications of these findings led us to consider the possibility of incorporating evolution of the molecular marker into the concept of a chain of transmission. We found that, by allowing for up to two changes in the IS*6110* RFLP fingerprint, it was possible to link strains together into 'superclusters' on the basis of Nearest Genetic Neighbour. As we anticipated, this significantly increased the estimate of disease attributable to ongoing transmission in the study community to between 73 and 88% (Chapter 4). The implications of this finding are twofold. Firstly. this demonstrates that the current TB control programme is unable to diagnose and treat cases before they have infected both their close and community contacts. Secondly, an intervention strategy which targets transmission has the potential to dramatically reduce the level of tuberculosis in the community. However, such a strategy will need to be vigorously maintained for decades, given that the high level of transmission has created a substantial pool of latent infection which may sporadically reactivate.

A secondary observation of note that emerged from the previous study was the substantial impact of the degree to which the estimation of clustering is affected by the temporal boundaries of the data under investigation. This suggested that the length of a study may have a substantial impact on such calculations, a phenomenon which has previously been largely ignored. The reason for this is that transmission chains which overlap the edges of the study window appear to be smaller than they in fact are. Our analysis of sub-windows of longer datasets (both real and simulated) confirmed this hypothesis, revealing that studies conducted over shorter time periods will significantly under-estimate the extent of clustering (Chapter 5). We also noted that this effect is highly correlated with the size of a cluster so that epidemics consisting predominantly of smaller clusters will be more severely affected. Accordingly, we proposed that molecular epidemiological studies be conducted over a minimum of four years. Where this is not possible, calculations of ongoing transmission should be viewed under the caveat that they will under-estimate the true value, particularly in epidemics dominated by small clusters.

While numerous studies have described the characteristics of epidemics in various settings as a static overview, not many that have addressed the dynamics of an epidemic in terms of a changing strain population structure, particularly in a high-incidence context. Having access to twelve years of molecular, demographic and clinical data from a study site in Cape Town, South Africa, enabled us to conduct such an analysis (Chapter 6). We found that, while the incidence of tuberculosis cases infected with four of the predominant strain clades remained relatively stable over this period, cases with Beijing clade strains increased exponentially. It was also apparent that, while this phenomenal growth was due to drug-sensitive strains, both drug-sensitive and –resistant strains enjoyed a selective advantage over their non-Beijing counterparts. Comparison of the clinical and demographic data of cases with Beijing and non-Beijing clade strains failed to identify major factors which could explain our observations with the exception that the Beijing clade strains had a greater proportion of smear positive sputa and were less likely to be successfully treated than other strains. Thus we suggested that the genetic make-up of strains within the Beijing clade may encode a higher level of fitness than those of other strains. This would also allow Beijing clade strains to more easily overcome the fitness cost imposed by the acquisition of mutations conferring resistance, thereby resulting in their more frequent transmission. However, we cannot exclude the possibility that host-pathogen compatibility allowed the Beijing clade to expand exponentially. It was in such a context that we used another genotyping marker (spoligotyping) to investigate the relationship between the success of different Beijing sublineages within the Beijing clade and their host population (Chapter 7). We found a significant correlation between prevalence of certain sublineages and the regions in which they are found (East Asia *vs.* South Africa) which could not be explained by founder effects. Our theory, therefore, is either, that different host populations have selected for different Beijing sublineages, or that specific sublineages have adapted to be more successful in particular human populations.

While IS*6110* remains a useful molecular epidemiological tool, it has certain limitations in that it requires time-consuming culturing of the sputum sample and is only applicable to strains of *M. tuberculosis* which possess more than five copies of the insertion element. For these reasons, alternative molecular markers have been sought and one which has gained favourable consideration is Mycobacterial Interspersed Repetitive-Unit–Variable-Number Tandem-Repeat (MIRU-VNTR) which is based on a PCR amplification of a combination of minisatellite-like variable-number repeat regions scattered around the bacterial genome. Any molecular epidemiological marker which is to replace IS*6110* as the standard must be able to perform at least as well as it in a variety of contexts. Variations of MIRU-VNTR genotyping, using different combinations of loci, have compared favourably with IS*6110* in studies done in low-incidence settings, however, it was not clear that this would be the case in a high-incidence community with high levels of transmission. We therefore set out to compare a number of MIRU-VNTR loci combinations with IS*6110* (Chapter 8). We used a variety of methods to evaluate the performance of the MIRU-VNTR markers with respect to that of IS*6110* within the context of a subset of strains found in the study population which belonged to the Beijing clade and which had a robust phylogeny. We also compared the ability of both marker types to correctly describe *M. tuberculosis* isolates from a Beijing sub-lineage which was well-characterised by four drug-resistance markers. We found that while MIRU-VNTR yielded clustering estimates that were similar to that of IS*6110*, this was merely coincidental. The independent evolution of the two types of marker resulted in discordance between both the predicted transmission chains and strains with unique genotypes. The analysis of the Beijing sublineage suggested that IS*6110* was more accurate marker in terms of discriminating between epidemiologically unrelated cases although it tended to split related cases into a number of subgroups. Allowing for limited evolution overcame this problem to a large degree in the case of IS*6110*, but was generally not useful when applied to MIRU-VNTR. Our findings differ from previous studies which have demonstrated a close correlation between IS*6110*-RFLP and MIRU-VNTR genotyping. However, these studies were conducted in low incidence, Western European settings where the TB epidemic is primarily driven by reactivation and immigration. In this context, efficient TB control programs would largely prevent recent and ongoing transmission with the subsequent generation of closely related clonal variants. In most instances, this would imply that strains cultured from TB cases would be genetically distantly related and would thus not share either IS*6110*-RFLP or MIRU-VNTR genotypes. In this situation, MIRU-VNTR and IS*6110*-RFLP genotyping would discriminate strains equally well. In contrast, our high incidence setting has promoted the evolution of a large number of genetically, closely related strains which are maintained within the host population. The evolutionary distance between them of such a nature that strains often differ in terms of only one of the two markers. Accordingly, we hypothesised that the degree of discordance between IS*6110*-RFLP and MIRU-VNTR genotyping is dependent on the genetic distance between isolates. This is supported by the observation that distantly related isolates from different Beijing sublineages have evolved distinct IS*6110*-RFLP and MIRU-VNTR genotypes. Therefore, in spite of its limitations, particularly when using fewer loci, MIRU-VNTR remains a valuable technique and is just as robust as IS*6110* in the elucidation of higher level phylogenetic relationships.

We have, in the course of these studies and those listed in Appendix B, accumulated a substantial, longitudinal dataset comprising both clinical and molecular strain data. This continuously growing database constitutes a valuable resource for both epidemiological and other TB-related studies. Much still remains to be done in the field of molecular epidemiology of tuberculosis. Although the value of such investigations have been well established, the tools and our understanding of what they tell us are still maturing. The need for a better marker for tracking transmission chains is accepted, but there is as yet, no obvious candidate which meets all the criteria. It may well be that there exists no single solution and that different markers will continue to be used for specific applications. Whatever marker is used, however, the data it provides require careful interpretation. Many factors besides those presented here, influence estimates of clustering. These would include: the proportion of TB cases that are sampled and are able to be genotyped, the level of drug-resistance, the types of cases included in the sample and whether or not they were infectious and the mobility of the community which leads to both influx and efflux of bacterial strains. The relatively recent phenomenon of HIV and its interaction with tuberculosis adds a new dimension to the epidemic, further complicating interpretation. In addition, *M. tuberculosis* continues to evolve.

Phthisis, Consumption, Scrofula, King's evil, Pott's disease – tuberculosis has been with us in its various guises for millennia. This ancient and resourceful enemy will not be easily defeated.

Appendix A

Candidate’s Contributions

Chapter 1 Molecular biology of tuberculosis

- Primary author

Chapter 2 Calculation of the stability of the IS*6110* banding pattern in patients with persistent *M. tuberculosis* disease

- Equal contribution with first author
- Conceptualisation and Planning of study
- Construction of database
- Analysis and interpretation of data
- Writing of manuscript

Chapter 3 Evolution of the IS*6110* based RFLP pattern during the transmission of *Mycobacterium tuberculosis*

- Equal contribution with first author
- Planning of study
- Construction of database
- Analysis and interpretation of data
- Writing of manuscript

Chapter 4 Genetic Distance: A measure of ongoing transmission of *Mycobacterium tuberculosis*

- Conceptualisation and planning of study
- Construction of database
- Development of analytical tools, analysis and interpretation of data
- Writing of manuscript

Chapter 5 The implications of study-duration in the molecular epidemiology of tuberculosis

- Conceptualisation and planning of study
- Construction of database
- Development of analytical tools, analysis and interpretation of data
- Writing of manuscript

Chapter 6 Population dynamics of *Mycobacterium tuberculosis* genotype families: A 12-year perspective of an epidemic

- Conceptualisation and planning of study
- Construction of database
- Analysis and interpretation of data
- Writing of manuscript

Chapter 7 Evidence that the spread of *Mycobacterium tuberculosis* strains with the Beijing genotype is human population dependent

- Analysis and interpretation of data
- Writing of manuscript

Chapter 8 Discordance between MIRU-VNTR genotyping and IS*6110* RFLP fingerprinting of Beijing genotype *Mycobacterium tuberculosis* strains

- Joint first author
- Planning of study
- Construction of database
- Development of analytical tools, analysis and interpretation of data
- Writing of manuscript

Appendix B

Publication List

In addition to the papers which form part of this thesis, I have also contributed to the publications listed below in various ways. Besides playing a part in the planning of these studies and the writing and editing of manuscripts, a major contribution I have made has been in the generation, management and classification of molecular data such as IS*6110* RFLP, spoligotype, PGRS and MIRU-VNTR. A substantial part of this has been the establishment of interconnected databases for the storage and retrieval of this data. A second important aspect has been in the analysis of molecular and clinical/demographic data. This has often involved the development of analytical techniques and software.

Hoek K G P, Gey van Pittius N C, Moolman-Smook H, Carelse-Tofa K, Jordaan A, van der Spuy G D, Streicher E, Victor T C, van Helden P D and Warren R M. *(2008)* Fluorometric assay for testing rifampin susceptibility of *Mycobacterium tuberculosis* complex. *J Clin Microbiol 46:1369-73.*

Hanekom M, van der Spuy G D, Streicher E, Ndabambi S L, McEvoy C R E, Kidd M, Beyers N, Victor T C, van Helden P D and Warren R M. *(2007)* A recently evolved sublineage of the *Mycobacterium tuberculosis* Beijing strain family is *associated* with an increased ability to spread and cause disease. *J Clin Microbiol 45:1483-90.*

Chihota V, Apers L, Mungofa S, Kasongo W, Nyoni I M, Tembwe R, Mbulo G, Tembo M, Streicher E M, van der Spuy G D, Victor T C, van Helden P and Warren R M. *(2007)* Predominance of a single genotype of *Mycobacterium tuberculosis* in *regions* of Southern Africa. *Int J Tuberc Lung Dis 11:311-8.*

Victor T C, Streicher E M, Kewley C, Jordaan A M, van der Spuy G D, Bosman M, Louw H, Murray M, Young D, van Helden P D and Warren R M. *(2007)* Spread of an emerging *Mycobacterium tuberculosis* drug-resistant strain in the western Cape of *South* Africa. *Int J Tuberc Lung Dis 11:195-201.*

Streicher E M, Victor T C, van der Spuy G, Sola C, Rastogi N, van Helden P D and Warren R M. *(2007)* Spoligotype signatures in the *Mycobacterium tuberculosis* complex. *J Clin Microbiol 45:237-40.*

Johnson R, Jordaan A M, Pretorius L, Engelke E, van der Spuy G, Kewley C, Bosman M, van Helden P D, Warren R and Victor T C. *(2006)* Ethambutol resistance testing by mutation detection. *Int J Tuberc Lung Dis 10:68-73.*

van Rie Annelies, Victor Thomas C, Richardson Madalene, Johnson Rabia, van der Spuy Gian D, Murray Emma J, Beyers Nulda, Gey van Pittius Nico C, van Helden Paul D and Warren Robin M. *(2005)* Reinfection and mixed infection cause changing *Mycobacterium tuberculosis* drug-resistance patterns. *Am J Respir Crit Care Med 172:636-42.*

Verver Suzanne, Warren Robin M, Beyers Nulda, Richardson Madalene, van der Spuy Gian D, Borgdorff Martien W, Enarson Donald A, Behr Marcel A and van Helden Paul D. *(2005)* Rate of reinfection tuberculosis after successful treatment is higher than rate of new tuberculosis. *Am J Respir Crit Care Med 171:1430-5.*

Verver S, Warren RM, Munch Z, Richardson M, van der Spuy GD, Borgdorff MW, Behr MA, Beyers N and van Helden PD. *(2004)* Proportion of tuberculosis transmission that takes place in households in a high-incidence area. *The Lancet 363:212-214.*

Verver S, Warren RM, Munch Z, Vynnycky E, van Helden PD, Richardson M, van der Spuy GD, Enarson DA, Borgdorff MW, Behr MA and Beyers N. *(2004)* Transmission of tuberculosis in a high incidence urban community in South Africa. *In J Epidemiol 33(2):351-357.*

Warren RM, Victor TC, Streicher EM, Richardson M, van der Spuy GD, Johnson R, Chihota VN, Locht C, Supply P and van Helden PD. *(2004)* Clonal Expansion of a Globally Disseminated Lineage of *Mycobacterium tuberculosis* with Low *IS6110* Copy Numbers. *J Clin Microbiol 42 (12):5774-5782.*

Richardson M, van der Spuy G, Sampson SL, Beyers N, van Helden PD and Warren RM. *(2004)* Stability of polymorphic GC-rich repeat *sequence*-containing regions of *Mycobacterium tuberculosis* . *Journal of Clinical Microbiology 42(3):1302-1304.*

Victor TC, de Haas PEW, Jordaan AM, van der Spuy G, Richardson M, van Soolingen D, van Helden PD and Warren R. *(2004)* Molecular characteristics and global spread of *Mycobacterium tuberculosis* with a Western Cape F11 genotype.*Journal of Clinical Microbiology 42(2):769-772.*

Streicher EM, Warren RM, Kewley C, Simpson J, Rastogi N, Sola C, Filliol I, van der Spuy G, van Helden PD and Victor TC. *(2004) Genotypic* and phenotypic characterization of drug-resistant *Mycobacterium tuberculosis* isolates from rural districts of the Western Cape Province of South Africa.*Journal of Clininal Microbiology 12(2):891-894.*

Van Helden PD, Warren RM, Uys P, van der Spuy GD and Victor TC. *(2004)* The molecular epidemiology of MDR-TB. *in* Management of Multiple Drug-Resistant Infections - Part III:225-242 *eds Gillespie SH*, Humana Press Inc.

Sampson SL, Warren RM, Richardson M, Victor TC, Jordaan AM, van der Spuy GD and van Helden PD. *(2003)* IS*6110*-mediated deletion polymorphism in the DR region of clinical isolates of *Mycobacterium tuberculosis* . *J.Bact 185(9):2856-2866.*

Supply P, Warren RM, Bañuls A-L, Lesjean S, van der Spuy GD, Lewis L-A, Tibayrenc M, van Helden PD and Locht C. *(2003)* Linkage disequilibrium between minisatellite loci supports

clonal evolution of *Mycobacterium tuberculosis* in a high tuberculosis incidence area. *Mol.Microbiol 47(2):529-538.*

Savine E, Warren RM, van der Spuy GD, Beyers N, van Helden PD, Locht C and Supply P. *(2002)* Stability of variably-number tandem repeats of mycobacterial interspersed repetitive units from 12 loci in serial isolates of *Mycobacterium tuberculosis .J.Clin.Micro 40(12):4561-4566.*

Warren RM, Streicher EM, Sampson SL, van der Spuy GD, Richardson M, Nguyen D, Behr MA, Victor TC and van Helden PD. *(2002)* Micro-evolution of the direct repeat region of *Mycobacterium tuberculosis* : Implications for interpretation of spoligotyping data. *J.Clin.Micro 40(12):4457-4465.*

Warren RM, Streicher EM, Charalambous S, Churchyard G, van der Spuy GD, Grant AD, van Helden PD and Victor TC. *(2002)* Use of spoligotyping for accurate classification of recurrent tuberculosis.*J.Clin.Micro 40(10):3851-3853.*

Victor TC, Lee H, Cho S-N, Jordaan AM, van der Spuy G, van Helden PD and Warren R. *(2002)* Molecular detection of early appearance of drug resistance during *Mycobacterium tuberculosis* infection. *Clin.Chem.Lab.Med 40(9):876-881.*

Richardson M, van Lill SWP, van der Spuy GD, Munch Z, Booysen C, Beyers N, van Helden PD and Warren RM. *(2002)* Historic and recent events contribute to the disease dynamics of Beijing-like *M. tuberculosis* isolates in a high incidence region. *Int.J.Tuberculosis and Lung Disease 6:1001-1011.*

Richardson M, Carroll NM, Engelke E, van der Spuy GD, Salker F, Munch Z, Gie RP, Warren RM, Beyers N and van Helden PD. *(2002)* Multiple *Mycobacterium tuberculosis* strains in early cultures from patients in a high incidence community setting.*J.Clin.Micro 40:2750-2754.*

Sharaf-Eldin GS, Saeed NS, Hamid ME, Jordaan AM, van der Spuy GD, Warren RM, van Helden PD and Victor TC. *(2002)* Molecular analysis of clinical isolates of *Mycobacterium tuberculosis* collected from patients with persistent disease in the Khartoum region of Sudan. *Journal of Infection 44:244-251.*

Van Helden P, Warren R, Victor T, van der Spuy G, Richardson M and Hoal van Helden E. *(2002)* Strain families of *Mycobacterium tuberculosis*. *Trends in Microbiology 10(4):167-168.*

Victor TC, van Rie A, Jordaan AM, Richardson M, van der Spuy GD, Beyers N, van Helden PD and Warren R. *(2001)* Sequence polymorphism in the rrs gene of *M. tuberculosis* is deeply rooted within an evolutionary clade and is not associated with streptomycin resistance. *J.Clin.Micro 39(11):4184-4186.*

Warren RM, Richardson M, Sampson SL, van der Spuy GD, Bourn W, Hauman JH, Heersma H, Hide W, Beyers N and van Helden PD. *(2001)* Molecular evolution of *Mycobacterium tuberculosis*: phylogenetic reconstruction of clonal expansion. *Tuberculosis 81(4):291-302.*

Sampson S, Warren R, Richardson M, van der Spuy G and van Helden P. *(2001)* IS*6110* insertions in *M. tuberculosis*: predominantly into coding regions.*J.Clin.Micro 39:3423-3424.*

van Rie A, Warren R, Mshanga I, Jordaan AM, van der Spuy GD, Richardson M, Simpson J, Gie RP, Enarson DA, Beyers N, van Helden PD and Victor TC. *(2001)* Analysis for a limited number of gene codons can predict drug resistance of *Mycobacterium tuberculosis* in a high incidence community.*J.Clin.Micro 39:636-641.*

Warren RM, Sampson SL, Richardson M, van der Spuy GD, Lombard CJ, Victor TC and van Helden PD. *(2000)* Mapping of IS*6110* flanking regions in clinical isolates of *M. tuberculosis* demonstrates genome plasticity. *Mol.Microbiol 37:(6-1405.*

Victor TC, Jordaan AM, van Rie A, van der Spuy GD, Richardson M, van Helden PD and Warren R. *(1999)* Detection of mutations in drug resistance genes of *Mycobacterium tuberculosis* by a dot-blot hybridization strategy. *Tubercle & Lung Disease 79:343-348.*

Sampson SL, Warren RM, Richardson M, van der Spuy GD and van Helden PD. *(1999)* Disruption of coding regions of IS*6110* insertion in *Mycobacterium tuberculosis*. *Tubercle & Lung Disease 79:349-359.*

Warren R, Richardson M, van der Spuy G, Victor T, Sampson S, Beyers N and van Helden P. *(1999)* DNA fingerprinting and molecular epidemiology of tuberculosis use and interpretation in an epidemic setting. *Electrophoresis 20:1807-1812.*

Biography

Gian van der Spuy was born and raised in Cape Town, South Africa. He graduated from the University of Cape Town with a BSc (hons) in Biochemistry and Microbiology in 1987. He then switched fields to obtain an MSc in Neurochemistry *(cum laude)* at Stellenbosch University in 1991. He subsequently altered course once again to work on the subject that comprises this PhD thesis. He is married to Dorothy, a paediatric dietician, who does her best to keep him healthy. Both are committed Christians. As well as enjoying the outdoors, he also cooks, reads widely, does a little woodwork and is an amateur jazz saxophonist.