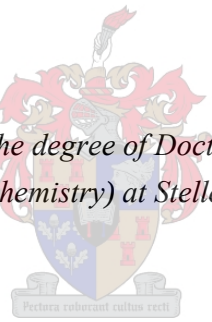
The Molecular Epidemiology of *Mycobacterium tuberculosis*: Role in understanding disease dynamics in high prevalence settings in Southern Africa Region

by
Violet Chihota



Dissertation presented for the degree of Doctor of Philosophy (Medical Biochemistry) at Stellenbosch University

Promoter: Prof Robin M. Warren
Co-promotor: Prof Paul D. van Helden
Faculty of Health Sciences
Department Biomedical Sciences

March 2011

Declaration

By submitting this thesis/dissertation electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the sole author thereof (save to the extent explicitly otherwise stated), that reproduction and publication thereof by Stellenbosch University will not infringe any third party rights and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

March 2011

Signature…… Date: ……

SUMMARY

The tuberculosis (TB) incidence has increased in Southern Africa and the situation isworsened by the emergence of drug-resistant *Mycobacterium tuberculosis* strains. Molecular biological techniques have been used to understand the disease dynamics of TB. In a series of studies we describe the use of these techniques to understand the disease dynamics of TB in Southern Africa.

Using spoligotyping and IS*6110*-restriction fragment length polymorphism (RFLP) to characterize *M. tuberculosis* strains from TB patients in Zimbabwe, we identified a genotype causing a disproportionate number of TB cases. The genotype belonged to the Latin American Mediterranean (LAM) lineage and we named it the Southern Africa1 (SAF1) family and later renamed it SAF1/RD^{Rio}, also reflecting its predominance in South America. To establish if this family of strains was predominant elsewhere in Southern Africa, genotypes were compared to those from Western Cape, South Africa and Zambia. The SAF1/RD^{Rio} strains were highly prevalent in Zambia but were only a minor fraction of the strains in South Africa. The geographical distribution of SAF1/RD^{Rio} strains was determined in Gweru, Zimbabwe, and was found to be spread in high incidence areas. From these two studies it was hypothesized that certain host and bacterial factors were associated with disease due to SAF1/RD^{Rio}.

Subsequently potential risk factors and clinical outcomes of disease due to SAF1/RD^{Rio} strains were explored. An association was found with smoking and cavitary pulmonary disease suggesting that SAF1/RD^{Rio} caused a more severe and highly transmissible formof TB

Using IS*6110*-RFLP, principal genetic grouping, spoligotyping, IS*6110* insertion-site mapping and variable-number tandem repeats (VNTR) typing, low IS*6110* copy clade (LCC) identified in Zimbabwe were characterized and compared to the strains from Cape Town, South Africa and other regions. The LCC strains from Cape Town, South Africa, were found to have close evolutionary relationship with strains from Zimbabwe and other regions and were widely distributed suggesting they play an important role in the global TB epidemic.

Observations from these studies and those from other studies led to the hypothesis that specific genotypes of *M. tuberculosis* predominate in regions of Southern Africa. To gain an insight on the population structure of *M. tuberculosis* strains in Southern Africa, spoligotyping and/or IS*6110*-RFLP data from eight countries were compared. This is the first study to describe the *M. tuberculosis* population structure in Southern Africa. Distinct genotypes were associated with specific geographic regions. These findings have important implications for TB diagnostics, anti-TB drug and vaccine development.

The population structure of multidrug-resistant (MDR), pre-extensively drug-resistant (pre-XDR) and extensively drug-resistant (XDR) *M. tuberculosis* isolates from provinces in South Africa was also determined. This is again the first study to describe the population structure of drug-resistant *M. tuberculosis* in South Africa. The results also showed geographic localization of genotypes and an association with resistance class. However, decreasing strain diversity was observed as the isolates evolved from MDR-TB to XDR-TB suggesting selection for the specific genotypes. These findings highlight the

importance of identifying genetic markers in drug-resistant strains, to enhance early detection of those at risk of developing XDR-TB.

OPSOMMING

Die voorkoms van tuberkulose (TB) in Suider Afrika word vererger deur stamme van *Mycobacterium tuberculosis* wat weerstandig is teen die beskikbare anti-tuberkulose middels. Molekulêre tegnieke word gebruik om in hierdie reeks studies die dinamika van TB in Suider Afrika te ondersoek

Deur spoligotipering en IS*6110* restriksie fragment lengte polimorfisme (RFLP) tegnieke te gebruik om *M. tuberculosis* stamme van pasiente in Zimbabwe te beskryf, het ons ‘n genotipe gevind wat ‘n buitengewone aantal TB gevalle veroorsaak het. Hierdie genotipe is deel van die internasionaal beskryfde Latyns Amerikaase en Meditereense (LAM) stam familie. Ons het dit die Suider Afrikaanse Familie1 (SAF1) genoem, maar later hernoem na SAF1/RD^{Rio}, omdat dieselfde genotipe in ook volop is in Suid Amerika. Om vas te stel of hierdie familie ook oorheesend is in die res van Suider Afrika, is dit vergelyk met beskikbare databasisse van die Wes-Kaap, Suid-Afrika en Zambië. Alhoewel SAF1/RD^{Rio} in die Wes-Kaap gevind is, dra dit slegs tot ‘n mindere mate by tot die plaaslike TB epidemie. Aan die anderkant kom SAF1/RD^{Rio} baie algemeen in Zambië voor. ‘n Verdere studie wys ook dat die SAF1/RD^{Rio} familie eweredig en wyd verspreid voorkom in hoë insidensie gebiede in Gweru, Zimbabwe. Vanuit die bevindings van hierdie 2 studies, kan ons aflei dat sekere gasheer- en bakteriële eienskappe geassosieer is met SAF1/RD^{Rio}-TB-infeksie.

Hierna is potensiële risiko faktore en kliniese uitkomste van siekte as gevolg van infeksie met SAF1/RD^{Rio} ondersoek. ‘n Assosiasie met rook en kaviterende pulmonale infeksie is

gevind,wat daarop dui dat SAF1/RD^{Rio} erger vorm van TB veroorsaak en hoogs oordraagbaar is.

Deur gebruik te maak van IS*6110*- (RFLP), hoof groep groepering, spoligotipering, IS*6110* invoegings kaartering en veranderlike getal tandem herhaling (VNTR) tipering kon lae IS*6110* invoeginsgetal (LCC) stamme van Kaapstad, Zimbabwe en ander gebiede vergelyk word. Al die LCC stamme in die studie is evolusionêr naby verwant aan mekaar en is wyd verspreid, wat dui op hulle belangrike rol in die wêreldwye TB epidemie.

Waarnemings in hierdie asook ander studies het tot die hipotese gely dat spesifieke genotipes van *M. tuberculosis* dominant is in verskillende gebiede van Suider Afrika. Om meer insig tot die populasie samestelling van *M. tuberculosis* stamme in Suider Afrika in te win is spoligotipes en RFLP-data van 8 lande vergelyk. Hierdie is die eerste studie om die populasie samestelling van *M. tuberculosis* in Suider Afrika te beskryf en is belangrike fir toekomstige ontwikkeling van nuwe TB diagnose tegnieke, anti-TB middels en TB entstowwe.

Die populasie samestelling van multiweerstandige (MDR), pre-ekstreme weerstandige (pre-XDR) en ekstreme weerstandige (XDR) *M. tuberculosis* van verskillende provinsies in Suid-Afrika is ook bepaal. Hierdie studie is ook die eerste wat die populasie samestelling van weerstandige *M. tuberculosis* in Suid-Afrika beskryf. Die resultate wys geografiese lokalisering van genotipes en ‘n assosiasie met weerstandigheidsklas. ‘n Afname in stam diversiteit soos die isolate van MDR-TB tot XDR-TB ontwikkel, dui op

seleksie van spesifieke genotipes. Hierdie bevinding lê die klem op die belangrikheid van die identifisering van genetiese merkers in weerstandige stamme om die risiko vir die ontwikkeling van XDR-TB te verminder deur vroë deteksie.

Acknowledgements

I would like to express my sincere thanks and gratitude to the following people with whom the studies described in this thesis would not have been possible.

I would like to thank my promoter Professor Robin Mark Warren without whom these achievements would not have been accomplished. His expertise, mentorship, understanding, and patience over the past 6 years and dedication to work has added considerably to my experience and has inspired me. I would also like to thank my co-promotor Professor Paul van Helden for giving me the opportunity to be part of the Division of Molecular Biology and Human Genetics. His eagerness to develop scientists from all over the world, in particular the African region is acknowledged.

Thanks to all my fellow scientists in the laboratory who have contributed to my learning of the laboratory techniques, Professor Tommie Victor and Dr Gian van der Spuy for advice and Dr. Elizabeth Streicher for the Afrikaans translations in this thesis.

I am greatly indebted to Dr Elizabeth Corbett and the team at Biomedical Research Institute TB laboratory in Harare, Zimbabwe for the isolates analysed in the studies. My sincere thanks go to Dr. Stanley Mungofa for all the help with setting up studies in the City of Harare and Dr, Ludwig Apers in Gweru, Zimbabwe. To our collaborators in Mozambique, Tanzania, Zambia, Gauteng South Africa, KwaZulu Natal South Africa; respectively for isolates analysed in this study.

To the Wellcome Trust for supporting my work at Stellenbosch and NIH Fogarty ICORTA TB/AIDS grant for training in epidemiology and biostatistics.

Thanks to Drs Katherine Fielding and James Lewis for all the statistical advice.

Finally I would like to thank my family for the support during this period and in particular I would like to acknowledge my husband Joseph, my two sons Benjamin and Tanaka and my parents Edward and Violet Nyatoti.

Table of contents

Chapter 1

Molecular Epidemiology of Tuberculosis in South Africa: Lessons learnt[1]

Violet N Chihota, Paul D van Helden, Thomas C Victor, Robin M Warren

1. **Review article published as a book chapter: VN Chihota**, PD van Helden, T. Victor and RM Warren. Molecular epidemiology of tuberculosis in South Africa: Lessons learnt: In Molecular Epidemiology and Population Genetics of Tuberculosis. Editors YF Ngeow and SF Yap: *Academy of Sciences Malaysia* 2006.

Table of Contents

Abstract

Over the past decade the incidence of tuberculosis in South Africa has increased. The situation has been worsened by human immunodeficiency virus (HIV) co-infection and the emergence of drug resistant strains. This places severe pressure on tuberculosis (TB) control efforts and has necessitated deviation from conventional epidemiological methods. Approaches to improve control efforts have included the use of molecular epidemiology techniques together with the conventional methods to better understand disease dynamics. The ability to distinguish between the different IS*6110*-restriction fragment length polymorphisms (-RFLP) of *Mycobacterium tuberculosis* isolates has greatly improved our understanding of TB epidemiology. In South Africa the use of molecular epidemiological techniques has enabled the definition of the population structure of both drug sensitive and resistant *M. tuberculosis* strains. Using both conventional and molecular epidemiological methods it has been possible to show that the TB epidemic in these settings has been attributed predominantly to ongoing transmission. IS*6110*-RFLP data from communities particularly in Cape Town shows that there are several distinct *M. tuberculosis* strain family groupings with characteristic IS*6110*-RFLP patterns and that a few of these families cause a disproportionately large number of both drug sensitive and drug resistant TB cases. Molecular techniques using IS*6110*-RFLP have also allowed documentation of exogenous reinfection and distinction between cases of TB drug resistance due to true acquisition and those due to reinfection with a drug resistant strain. Growing evidence from the molecular epidemiology studies also suggests that individuals may be infected with multiple strains of *M. tuberculosis*.

Not only has the data been used to answer immediate epidemiological questions but is beginning to be used to understand the evolution of this species.

1.0 Introduction

1.1 Global tuberculosis problem

Tuberculosis (TB) has plagued mankind for thousands of years. Today, TB is the world's second most common cause of death from infectious disease after HIV/AIDS. Despite significant scientific progress in understanding this disease, the global incidence of TB continues to increase. The estimated number of new cases of TB in the world has increased from 8.4 million in 1999 to 9.2 million in 2006 and 9.4 million in 2009(1-4). Of the new cases in 2006, 3-4 million cases were sputum smear positive at diagnosis (the most infectious form of tuberculosis (4).

Over the past decade South Africa has reported a rapid increase in TB incidence with new cases rising from less than 301 per 100 000 population in 1990 to 576 per 100 000 population in 2000 and continues to increase, with more than 900 per 100 000 from 2006 to 2009 (4). In terms of the estimated number of persons with active tuberculosis, South Africa is one of the five countries with the highest number of incident cases of TB (4) which is exacerbated by an increase in number of TB cases resistant to anti-TB drugs (5). Tuberculosis and human immunodeficiency virus (TB/HIV) co-infection is thought to be one of the main constraints to achieving TB control in the world. In South Africa it is estimated that more than 5.3 million people are living with HIV/AIDS, the highest in the world (6). Of the total 1.84 million deaths from tuberculosis worldwide in 2000, 246 000 were attributable to HIV infection with 203 000 occurring in Africa (1).

Infection due to *Mycobacterium tuberculosis* is commonly acquired via airborne transmission of infectious droplets (aerosols). Persons with pulmonary TB who are sputum smear positive, are thought to be the main sources of infection, but smear negative patients have also been shown to transmit *M. tuberculosis* (7). When susceptible hosts inhale the aerosols, they may become infected and an estimated 10-20% develop TB at some time during their life but have about half the risk of doing so within two years of infection (known as recent transmission) (8,9). In a majority of the cases, the infection is controlled within 2-10 weeks due to adequate immunity (10) and may force the pathogen into a dormant state (latent TB) for life. These latently infected individuals carry a 2-23% lifetime risk of developing reactivation TB while in immuno-compromised individuals this risk increases to 5-10% per year and is highly age dependent (10).

Because HIV suppresses the body's immune system, HIV-infected individuals are at increased risk of activation of latent infection and rapid progression of recently acquired infection to clinically active tuberculosis (11). The estimates made by Corbett *et al* (1) suggest a lifetime risk of developing TB of 30-40% for co-infected persons in Africa compared to a between 10% and 20% lifetime risk in individuals infected only with TB(8,9).

1.2 Drug- resistant tuberculosis

A further threat to TB control is the emergence of drug resistant TB. Drug resistant TB is defined as the state when *M. tuberculosis* becomes resistant to anti-microbial agents at the levels attainable in blood and tissue. New evidence suggests that drug resistant *M.*

tuberculosis strains are becoming more resistant and unresponsive to current treatment (12). Multi-drug resistant TB (MDR-TB), defined as resistance to at least isoniazid (INH) and rifampin (RIF) is also a major concern as these two drugs are the main drugs in the TB treatment regimen. Such resistance can develop through transmission (primary resistance) or under selective pressure of anti-TB agents (acquired) (13). The primary goals of anti-TB treatment are to kill the tubercle bacilli rapidly and hence prevent emergence of drug resistance, and eliminate persistent bacilli to prevent relapse. To achieve this, combination therapy, using more than one anti-TB drugs must be taken for a sufficiently long time. This will reduce the emergence of drug resistance during treatment as naturally occurring resistant strains would be selected for (selective pressure).

Traditionally, patients with drug resistant TB are classified as having acquired or primary drug resistance on the basis of a history of previous treatment. Primary resistance is defined as the presence of drug resistant *M. tuberculosis* in a patient with no prior or less than one month of treatment (13). This means that the resistance is less likely to develop due to selective pressure of anti-TB agents but more likely to have been transmitted. In general, primary resistance reflects the transmission of drug resistant TB over the years and it gives an indication of the achievements of the TB control program in controlling transmission. However, primary resistance may be grossly underestimated in cases of recurrent tuberculosis due to reinfection with drug resistant strains, since these patients may have had previous treatment (13). Acquired resistance is defined as resistance to anti-tuberculous drugs where a patient has received at least one month of anti-TB drugs (13). Thus acquired resistance can be used to evaluate TB treatment strategies and hence the performances of TB control programs. However this clinical classification is often

misleading, as the term acquired drug resistance includes patients with strains that truly acquired drug resistance during treatment and patients who were initially infected with a drug resistant strain, which was not detected by culture based phenotypic methods or by reinfection with a drug resistant strain (14).

Recent reports from the WHO Global Project on anti-TB drug resistance surveillance indicate that multidrug-resistant-TB (MDR-TB) is present in all regions of the world (5). In addition available data suggests that globally MDR-TB is not a problem of the same magnitude as drug resistance but is at critical levels in specific regions of the world (15). In the 27 high TB burden countries the proportion of MDR-TB among new cases ranged from 1.6 in Ethiopia to 22% in Azerbaijan (5). In the African region, the proportion of MDR-TB among TB patients is generally low with frequencies ranging from 0.5% - 3.9% among new cases and 0.0% to 16.7% among previously treated cases. Though these rates may seem low, the caseload of MDR-TB is high because of the high incidence of TB. There are an estimated 13 000 MDR-TB cases among new cases of TB in South Africa and in addition a growing problem of extensively drug-resistant TB (XDR-TB) (5). This places severe constraints on the health budget as the cost of treating each MDR-TB case is of the order of US$3400(5) excluding hospitalization costs.

1.3 Approaches to tuberculosis control

1.3.1 Classical Epidemiology:

Due to the nature of transmission and acquisition of *M. tuberculosis* infections, persons recently exposed to newly reported cases of TB (contacts) are at high risk of developing TB disease because they have been recently infected (16,17). Two or more cases are

considered linked if there is judged to be sufficient sharing of the environment for transmission to have occurred. Therefore the organism causing disease or infection in the contact is assumed to be the same strain that caused disease in the index cases.

The priorities of TB prevention and control programs include the identification and treatment of all persons with active TB and secondly to find contacts exposed to infectious TB patients and evaluate and treat them for latent TB infection (LTBI) or active TB disease (18,19). The identification of LTBI in contacts involves testing for reactivity to tuberculin by doing a tuberculin skin test (TST) where a positive TST indicates possible exposure to *M. tuberculosis*. Studies have shown that by the time contact investigation is done, some of the contacts exposed and infected by a source may already have progressed to active TB (19,20). However the majority will have latent TB infection (20,21). HIV is a strong risk factor for TB in individuals with latent TB or newly acquired infection.

To reduce morbidity and mortality from TB in people living with HIV infection, WHO now the recommends the provision of antiretroviral therapy (ART) and the Three I's; intensified case finding, isoniazid preventive therapy and infection control (22). It is therefore critical for TB control strategies to not only evaluate contacts in order to find new cases of active TB, but also to identify candidates with LTBI for preventive therapy with isoniazid. Daily administration of isoniazid for 6 months is recommended in people living with HIV who do not have TB including; children, pregnant women, individuals on ART and those who have just completed treatment for TB. (22). Treatment of LTBI

will reduce the latter group's risk of progression to TB, thereby minimizing further transmission of TB.

Studies of contact investigation have focused on TB transmission and identification of active TB among contacts. Close contacts with active TB disease were identified for TB patients with highly sputum smear positive disease and cavitary disease(19), drug resistant TB (22,23) and MDR-TB (24,25). More recently, observations suggest that more casual contact may in some instances be sufficient for the transmission of *M. tuberculosis* (26).

Although contact investigation has been the mainstay of classical epidemiological methods, a clear distinction between routine contact investigation and epidemiological investigations was made by Onorato *et al* (26,27). The former mainly involves the identification and evaluation of persons with either active TB or latent TB, while the latter is more directed toward identifying the dynamics of the disease and its transmission. In most countries, even though active case finding and contact tracing is advocated, infection in the contacts may well have come from a source other than the index case and many of the contacts leading to infection may be casual (28,29). Therefore many of these additional contacts of brief exposure may be entirely missed (16). These links are important in elucidating disease and transmission dynamics in TB and their identification may therefore require deviation from the basic routine classical epidemiological methods.

1.3.2 Molecular epidemiology

During the last decade, genome differences mediated by mycobacterial repetitive elements have been widely exploited in epidemiologic studies of TB (7,30-33). Although the *M. tuberculosis* genome contains several repetitive elements, only a few are polymorphic and have been widely studied. The best characterized repetitive elements include IS*6110* insertions (34), IS1081 (35), polymorphic guanine cytosine–rich repetitive sequences (*PGRS*)(36), direct repeats (*DR*) (37), the oligonucleotide $(GTG)_5$ sequence (38) and mycobacterial interspersed repetitive units (MIRUs) (39).

Molecular epidemiology of TB is most frequently based on the analysis of the distribution of the insertion sequence IS*6110* in different isolates using the internationally standardized IS*6110* restriction fragment length polymorphism (RFLP) methods (40). The transposable element IS*6110* may be repeated up to 25 times per *M. tuberculosis* genome (41,42), although strains with no IS*6110* insertion have been described (42,43). The IS*6110* copies are nearly identical in sequence but the number of insertions and location in the genome vary, thereby yielding distinct RFLP patterns. These are visualized as RFLP banding patterns, generated by restriction digestion of the whole genome with a restriction enzyme *Pvu*II, Southern transfer and hybridization with a labeled IS*6110* probe (40). Strains with six or fewer IS*6110* insertions are classified as low-copy-number strains, whereas those with more than six copies of IS*6110*, as high-copy-number strains (44).

The accuracy of genotyping can also be enhanced by the use of additional probes or secondary typing methods such as spacer oligonucleotide typing (spoligotyping) (45) and the more recently introduced method MIRU-VNTR typing (39) which determines the presence or absence of 43 direct variable repeat (DVR) sequences in the DR region and variable numbers tandem repeats, respectively. Data generated as such is stored in databases of IS*6110*-RFLP or spoligotype patterns of *M. tuberculosis* complex strains such as that at the National Institute of Public Health and Environment (RIVM) in Bilthoven, the Netherlands (46) and spoligotype database (SpolDB3) (47). These databases allow inter-laboratory comparisons to be made.

Sequence analysis of *M. tuberculosis* strains collected from different geographical settings has shown that the frequency of mutation in this organism is relatively low thereby demonstrating that the global population of *M. tuberculosis* is clonal (48). Therefore the transposition of the insertion sequences such as IS*6110* is a fundamental process in generating genomic variation. In addition to the classification using IS*6110*-RFLP within this clonal population of *M. tuberculosis*, strains can be assigned to one of three principal genetic groups according to single nucleotide polymorphisms (SNPs) in the k*atG* and *gyrA* genes (48). These groupings give a broad evolutionary scenario of *M. tuberculosis* strains (48).

The advent of molecular epidemiology using bacterial genetics to track organisms in defined communities and worldwide has enabled researchers to challenge many of the prevailing dogmas about TB epidemiology. IS*6110*-RFLP has been widely used in

community settings to support outbreak investigations (26,49), to document nosocomial transmission of TB (50,50,51), as a tool for studying TB transmission dynamics (31,52,53), for studying the *M. tuberculosis* population structure (32,54,55), as well as other aspects of TB epidemiology. It is envisaged that this data will form the foundation of future TB control strategies.

IS*6110*-RFLPs of *M. tuberculosis* isolates from patients with epidemiologically unrelated TB generally show differences, whereas those from patients infected by a common source are identical. Therefore the basic premise of most molecular epidemiology studies has been that isolates with the identical IS*6110*-RFLPs (clusters) are assumed to be part of the same chain of recent transmission (30,41,42). Conversely, isolates with unique IS*6110*-RFLPs (non-clustered) are assumed to be a result of reactivation of remote (latent) infection or influx from distant communities or failure to identify the source case (30,41,42). The extent of genetic homogeneity among strains in a community reflects the relative contribution of recently transmitted versus reactivated disease. This can have wide-reaching implications both for understanding and controlling TB and therefore must be interpreted carefully. Data from IS*6110*-RFLP however must be interpreted together with epidemiologic information, which is usually obtained by interviewing patients when TB is diagnosed.

2.0 Molecular epidemiology studies in South African communities

Tuberculosis molecular epidemiology studies have been done in a number of different regions of Southern Africa (31-33,52-55). However, in certain studies biases have made it difficult to accurately quantify the extent of recent transmission in different regions.

The amount of clustering seen will depend on the duration of the study, the comprehensiveness of sampling, and the heterogeneity of the *M. tuberculosis* strains in the particular region (56,57). South Africa has by far the most comprehensive molecular epidemiological data in Southern Africa with most of the reports emanating from the Western Cape Province. These studies have defined the population structure of both drug sensitive and drug resistant *M. tuberculosis* strains in communities (33,58,59), estimated the degree of clustering (31,33,52,53), established the risk factors for clustering (31,52,53), and helped document exogenous reinfection (60) and nosocomial transmission of TB (50,61). Furthermore, all these studies have provided information that could potentially be used to guide TB control efforts.

3.0 Mycobacterium tuberculosis strain diversity in South Africa

Initial attempts to characterize *M. tuberculosis* strains present in an epidemiological field site in Cape Town, South Africa, indicated that there was high level of strain diversity (33). The field site occupies an area of 3.4km^2 and has a population of 38 500.

A longitudinal IS*6110*-RFLP database representing clinical isolates from patients presenting with TB in this site within the Cape Town metropolitan was established in July 1992 and data collection is still ongoing. This database is maintained at the University of Stellenbosch's Department of Medical Biochemistry. Data from 1992-1998 shows that 1526 patients were reported with tuberculosis and that 1023 of these had culture positive TB (52,62). Based on IS*6110*-RFLP, approximately 100 different high and low copy number strains were identified per km^2 and cluster analysis grouped these

strains according to similarities. These groupings were assigned arbitrary numbers and are termed strain families. A total of 32 high copy number (≥ 6 IS*6110* bands) strain families, designated F1-F32 and one family designated the low-copy-number clade (strains with 1-6 IS*6110* bands) were identified (table 1). Some isolates were grouped as strain family 0 (2.12%) had unique IS*6110* fingerprints and did not have IS*6110*-RFLP similarities with any of the other strain families.

Among the different strain families identified to date in these communities, strain families designated F11, F28, F29 and the low-copy-number clade represented the groupings that were causing a disproportionately large number of TB cases (Table 1). These strain families have established themselves in this community and clonal variants are being spread from person to person (clonal spread). Interestingly these same four strain families were also responsible for more than 70% of the drug-resistant epidemic in the rural regions of Boland-Overberg and Southern Cape-Karoo of the Western Cape Province in South Africa (58). These findings may imply that within the communities studied, these *M. tuberculosis* strains are more effectively spread or alternatively these strains have been present in the community for longer periods.

Table 1: The frequency of high-and low-copy-number *M. tuberculosis* strain families from the epidemiological study site in Cape Town (data collected from 1993-1998)

Strain family	*Number of IS6110 inserts*	*Cases n (%)*
F 0	1-24	18 (2.3)
F 1	8-13	5 (0.6)
F 2	9-15	26 (3.3)
F 3	9-12	3 (0.4)
F 4	10-13	20 (2.5)
F 5	12-14	2 (0.3)
F 6	14-16	5 (0.6)
F 7	10-15	13 (1.6)
F 8	7-24	2 (0.3)
F 9	8-14	17 (2.1)
F10	14-17	4 (0.5)
F11	10-19	155 (19.4)
F12	2-8	1 (0.1)
F13	11-18	14 (1.8)
F14	9-18	26 (3.3)
F15	13-18	11 (1.4)
F16	10-13	11 (1.4)
F17	14-16	9 (1.1)
F18	13-19	19 (2.4)
F19	9-12	3 (0.4)
F20	18-21	6 (0.8)
F21	5-13	18 (2.3)
F22	9-11	4 (0.5)
F23	9-15	9 (1.1)
F24	9-18	13 (1.6)
F25	13-20	13 (1.6)
F26	11-24	11 (1.4)
F27	15-17	2 (0.3)
F28	8-16	79 (9.9)
F29	16-25	137 (17.2)
F30	8-9	2 (0.3)
F31	15	1 (0.1)
F32	12	1 (0.1)
Low-copy- number clade	1-6	137 (17.2)
Total		**797 (100)**

Marked strain diversity was also found in rural communities of Hlabisa in Durban, South Africa, with IS*6110*-RFLP revealing 175 distinct patterns from 246 patient isolates (53). However no published data on the distribution of the strain families in this study community is available and therefore it was not possible to compare data from the two different field sites.

4.0 Characteristics of dominant *M. tuberculosis* strain families:

4.1 Family 11:

Among the four strain families, F11 is the most predominant strain family in the Cape Town community, being represented by approximately a fifth of the isolates in the database (Table 1). On average the F11 strains have 14 IS*6110* inserts (range 11-19) and a characteristic banding pattern (Fig 1), being represented by 97 different IS*6110* variants in a total of 29 clusters and 52 related unique strains. These F11 strains belong to principal genetic group 2 (63). The spoligotype signature for this strain family is characterized by deletion of DVRs or spacers 9-11, 21-24 and 33-36 and a total of 14 different spoligotypes were identified (63). Family 11 strains have also been reported from other parts of South Africa (unpublished data) and were also shown to have a global distribution (59).

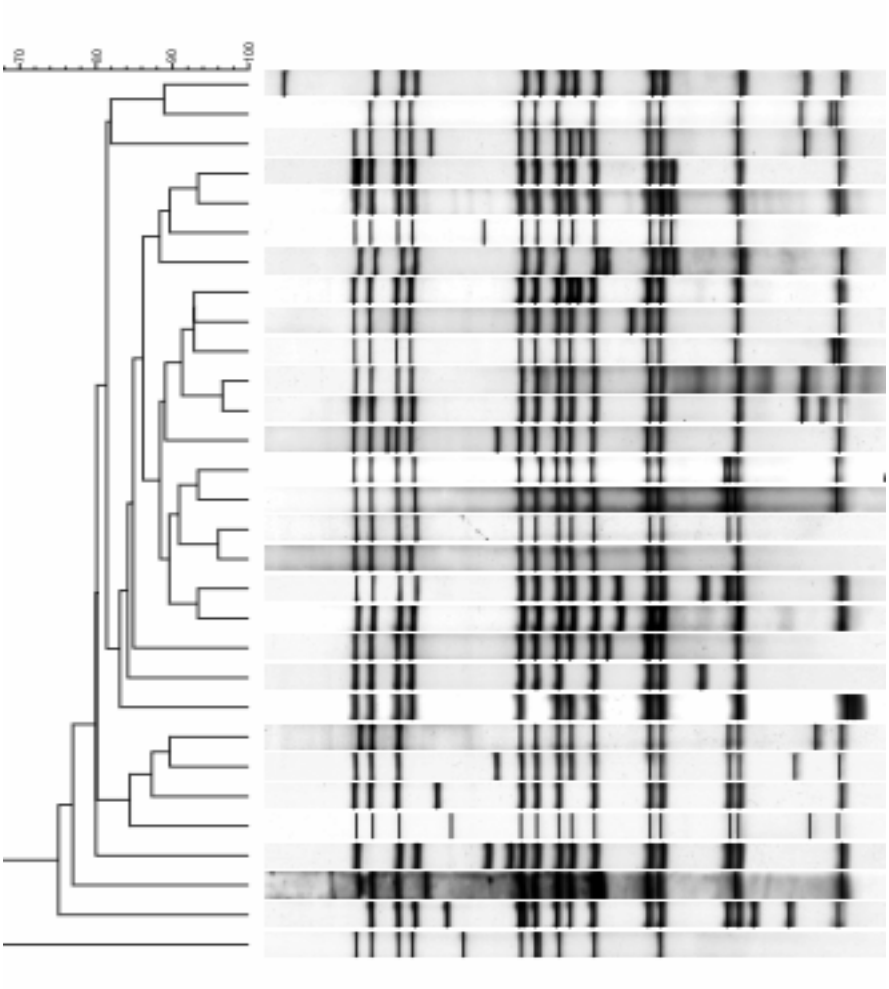


Fig 1: IS*6110* –RFLP patterns of representative isolates of family 11 *Mycobacterium tuberculosis* strains isolated from tuberculosis patients from an epidemiological field site in Cape Town, South Africa (data retrieved from the local database maintained in the Dept of Medical Biochemistry at the University of Stellenbosch). Strain family groupings were defined by isolates having an IS*6110* similarity index of >65%.

4.2 Family 28:

Among the patients with TB from the urban communities in Cape Town, F28 strains cause disease in 9.7% of the cases. Similarly this family is highly clonal demonstrated by high similarity among the IS*6110*-RFLP patterns (Fig 2). Isolates of the Family 28 clade

are represented by 41 characteristic IS*6110*-RFLP patterns (63) with the number of IS*6110*-RFLP hybridizing bands ranging from 8 to13. This strain family also shares a spoligotype signature characterized by deletion of DVRs 9-10 and 33-36 and belongs to principle genetic group 2 (Table 2) and variability in the spoligotype signature is also observed resulting in a total of 14 spoligotypes (63). No reports of this strain family have been made from other parts of South Africa and other regions of Africa. Although F28 appears common in the Western Cape Province, comparison of the IS*6110*-RFLP and the spoligotypes with the international RIVM database suggests that this family may be more widespread globally than previously thought. For instance Nguyen and colleagues (64,65) have identified strains harboring the same spoligotype signature and the same IS*6110* insertion site in low incidence areas of Quebec, Canada but with some of these strains associated with pyrazinamide mono-resistance.

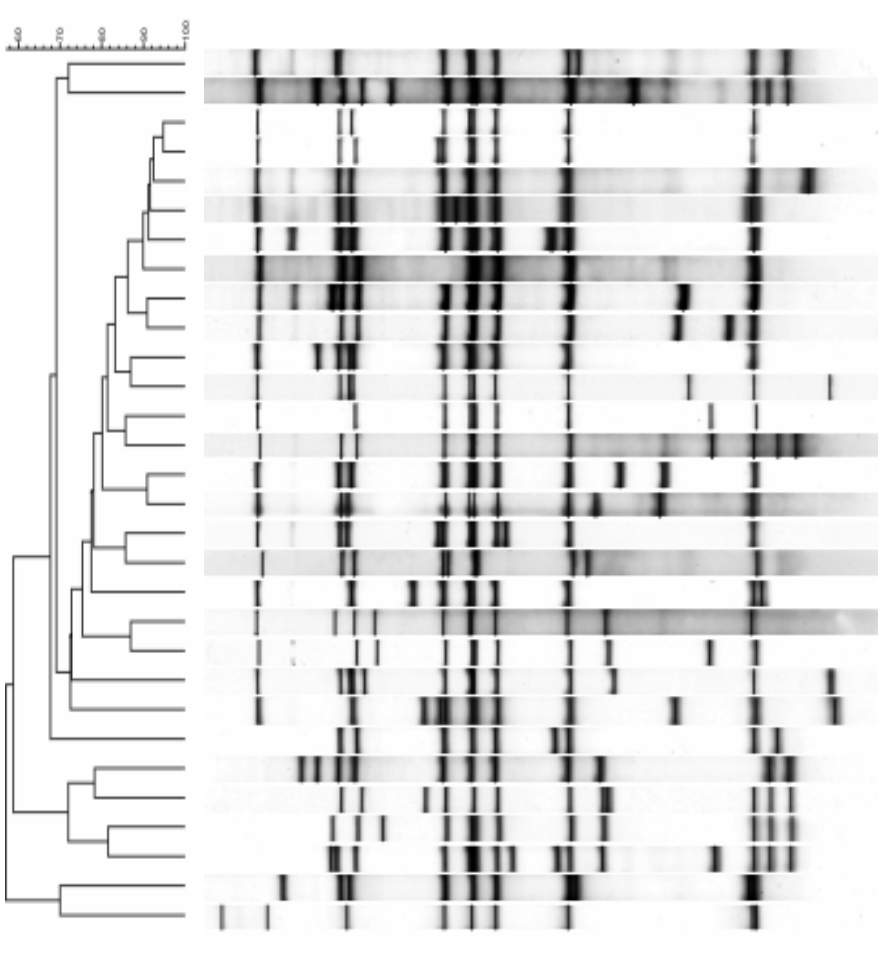


Fig 2: IS*6110*–RFLP patterns of representative isolates of family 28 Mycobacterium *tuberculosis* strains isolated from tuberculosis patients from an epidemiological field site in Cape Town, South Africa (data retrieved from the local database maintained in the Dept of Medical Biochemistry at the University of Stellenbosch). Strain family groupings were defined by isolates having an IS*6110* similarity index of >65%

4.3 Family 29 (Beijing-like clade)

During the period 1992-1998 F29 strains caused disease in approximately 17% of all TB cases in the epidemiological field site in Cape Town (62). Although these strains have a large array of IS*6110*-RFLP patterns characterized by having a high copy number (15-26 bands, average 17) (Fig 3) they are genetically related as they all belong to principal genetic group 1 (63,66) and have the same spoligotype pattern (Table 2) demonstrated by

deletion of the first 34 consecutive spoligotype spacers and presence of spacers 35-43(63). The Beijing strains, often associated with MDR-TB, have been reported to cause outbreaks in institutional settings(67), hospital environments (49) and have been implicated in ongoing community transmission (68-71), contributing to their endemic nature. The prevalence of the Beijing strains is remarkably high in countries in Asia (68,69,70, ,73). Their importance in South African communities has been reported, being associated with an outbreak of MDR-TB in Cape Town(73), and rural communities in the Western Cape Province(58). Only recently have these strains been isolated from other regions in African such as Kenya (74)and Zimbabwe (unpublished data). Although the frequency of Beijing strains in South Africa is not as high as reported elsewhere (68-71), the extensive clustering and high transmission rates suggest that this family of strains is highly transmissible and may represent an emerging clade in the Western Cape Province.

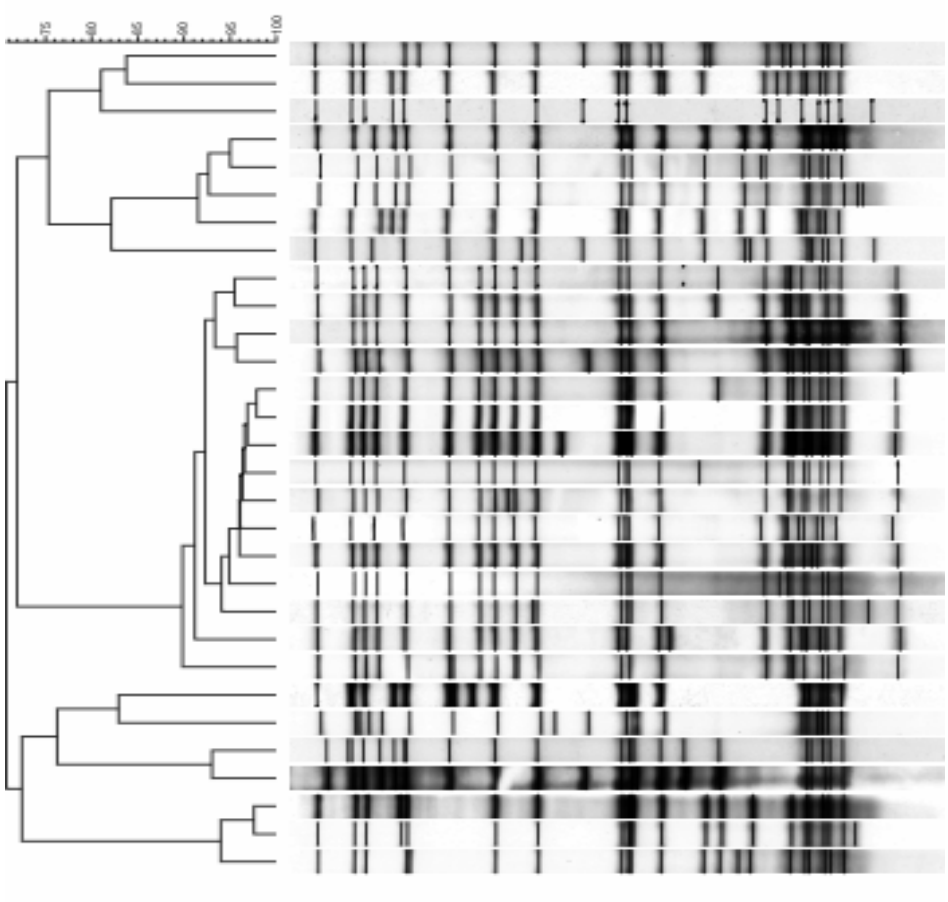


Fig 3: *IS6110* –RFLP patterns of representative isolates of family 29 (Beijing-like clade) *Mycobacterium tuberculosis* strains isolated from tuberculosis patients from an epidemiological field site in Cape Town, South Africa (data retrieved from the local database maintained in the Dept of Medical Biochemistry at the University of Stellenbosch). Strain family groupings were defined by isolates having an IS*6110* similarity index of 65%

4.4 Low-copy-number clade

The low-copy-number clade isolates are characterized by specific IS*6110*-RFLP genotypes and spoligotypes. In the well studied community in Cape Town, South Africa, 18.1% of the patients with TB during 1992-1998 were infected with low-copy-number strains(75). These strains have ≤ 6 IS*6110*-RFLP hybridizing fragments with inherently limited polymorphism (Fig 4). The IS*6110*-RFLP patterns in low copy number strains are

often more stable than their spoligotype pattern resulting in inherently limited IS*6110*-RFLP polymorphism and poor discriminatory power of IS*6110*-RFLP in these strains. This therefore necessitates the use of additional genotyping methods such as principle genetic grouping(48), spoligotyping(45), insertion site mapping (76-78), PGRS(36) and MIRU-VNTR(79), to characterize them.

Evidence from SNPs in the *katG* and *gyrA* genes suggests that *M. tuberculosis* strains of the low copy number clade from Cape Town may belong to either principal genetic group one or 2 representing two independently evolving lineages. Low-copy-number strains belonging to principal genetic group 2 show a deletion of DVRs 33-36, while clones of this group with ≥ 2 and ≤ 6 IS*6110* insertions also showed an additional deletion of DVR 18. Only one low-copy-number isolate belonging to principal genetic group 1 has been identified so far in the epidemiological field site in Cape Town and in this strain DVR 34 is uniquely deleted (Warren, personal communication). Low copy number strains have also been isolated from Central (Free State) and Northern (Gauteng and Mpumalanga) parts of South Africa (unpublished data), neighbouring countries such as Zimbabwe (unpublished data), Europe (76,79) and America (80).

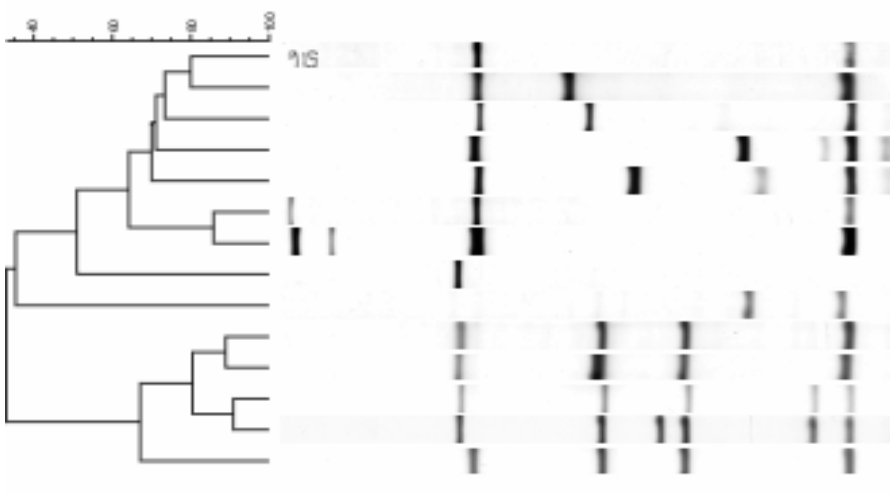


Fig 4: Dendrogram showing the number of IS*6110* bands of *Mycobacterium tuberculosis* isolates belonging to the low copy number clade, isolated from tuberculosis patients from the field site in Cape Town, South Africa. Low-copy-number isolates have 1-6 IS*6110* inserts.

Table 2: Characteristics defining the four predominant strain families isolated from TB patients in an epidemiological field site in Cape Town, South Africa

Strain Family	**IS*6110-3' bands***	**IS*6110-3' variants***	***Principal Genetic group***[a]	***Spoligotype in world wide database***[47]	***No of spoligotypes***
F11	11-19	97	2	33	14
F28	8-13	41	2	34	14
Beijing	15-26	38	1	1	1
Low-copy-number clade	1-6	15	1 or 2	18,53,70,92,119,137, 336, 347,352,489,806	19

[a]Principal genetic group classified according to mutations in the katG and gyrA genes (ref). Principle genetic group 1 isolates have the allele combination of *katG* codon 463 GTG (Leu) and *gyrA* codon 95 ACC (Thr) whereas group 2 isolates have *katG* codon 463 CGG (Arg) and *gyrA* codon 95 ACC (Thr)

5.0 Insights into the transmission of tuberculosis in the communities

Strain typing of TB has not only provided insights into the number and the type of strains circulating in communities but has also been used as a tool to quantify the degree of TB transmission. Depending on the specific epidemiological question being addressed, studies have varied in both design and the sampling methodologies.

5.1 Recent transmission and reactivation

In attempts to determine the number of people with disease due to recent transmission (clustered cases) as opposed to reactivated disease most studies have adopted the "*n* minus one method", based on the assumption that one case per cluster was due to reactivation and this case gave rise to other cases in the cluster, hence allowing one source case per cluster (81,82). The TB epidemic in urban areas of Cape Town has been attributed predominantly to ongoing transmission (52,83). Using conventional and molecular epidemiological methods to group *M. tuberculosis* strains, 58% of the TB cases were attributable to ongoing transmission (52).

This calculation was based on the assumption that the IS*6110* is sufficiently stable to define ongoing transmission. However more recently molecular epidemiology data also shows that IS*6110* may change in a subset of cases over time during the passage from an infected source case to a contact (chains of transmission) giving rise to variants with one or two band differences (84,85). Therefore traditional molecular epidemiology assumptions would exclude any change in IS*6110* banding patterns associated with transmission and such isolates may be incorrectly classified. A study by Warren *et al*(85)

showed that over a period of 6.5 years genotypes are estimated to change by approximately 20% during transmission events. Taking this into consideration van der Spuy *et al* (83) calculated the genetic distance (number of band differences) between strains circulating in the study site. Assuming that strains with a genetic distance of 1-4 represent ongoing transmission the proportion of cases due to ongoing transmission was estimated to be between 66-94% (83).

Not only has ongoing transmission been implicated in TB epidemic urban communities of Cape Town, it has also been implicated in other parts of South Africa, including communal settings with high cure rates of TB (31) and rural settings (53). In these communities where ongoing transmission rates were high (high degree of clustering), smear positivity and being a re-treatment case following default were the main risk factors for clustering (31,52). The apparent lack of epidemiological links in some of the studies in South Africa (31,33,53) and elsewhere (30,86) may point to the possible role of casual contact in transmission of TB or simply that some source cases were missed. The high level of ongoing transmission, clearly points to TB control interventions that are unable to effectively reduce transmission of TB. Whether the DOTS strategy of passive case finding coupled with other elements including observed chemotherapy is adequate in these high incidence communities may be questioned. A need for more emphasis on early presentation, active case finding, rapid diagnosis and subsequent treatment of cases may yield better results.

5.2 Household transmission

The assumption often made is close and prolonged contact and therefore transmission occurs amongst household members. This would imply that if two or more people in one household have TB, *M. tuberculosis* isolates from their cultures would be identical. Consequently a household contact has always been thought to play a major role in the transmission of TB and identifying household contacts and implementation of chemoprophylaxis has been the key in contact investigations. In children, TB is usually due to recently acquired infection. This infection is assumed to have been acquired from an adult within the same household who has active infectious form of TB.

Using IS*6110*-RFLP to classify strains in the epidemiological field site in the Western Cape Province, it was possible to demonstrate that people in the same household who have TB do not always have the same *M. tuberculosis* strain (29) even if the presumed secondary case was a child (87). Classen and colleagues demonstrated that in a majority of the households (55%) at least one patient had a strain different from other household members' strains (29). Similarly, of the 35 children with TB only 19 household contacts were identified and of these only 12 had the same strain as the child. The former study further highlighted the importance of contact in social gatherings which may not necessarily be casual contact (29). These findings imply that in high incidence settings, transmission of *M. tuberculosis* may occur equally in the community and household and these findings may have implications for overall quantification of disease and the traditionally used contact tracing strategies.

5.3 Transmission of drug resistant tuberculosis

Tuberculosis control in high incidence communities may be problematical if the MDR-TB scourge increases. The use of molecular techniques has helped establish the relative contribution of transmission vs. acquisition of drug resistant *M. tuberculosis*. Clustered drug resistant strains would imply recent transmission and genetically diverse and unrelated strains would suggest acquisition of drug resistance provided the possible source cases had been identified.

Using IS*6110*-RFLP and/or spoligotyping, to study the molecular epidemiology of TB in rural and urban communities of the Western Cape (14,33,58) has revealed that transmission is an important mechanism in the spread of drug resistant TB , and particularly MDR-TB. Characterization of isolates by both genotypic and phenotypic methods showed that MDR-TB contributed to the drug resistant TB epidemic (58,88). The extent of the problem is clearly given in the rural regions of Boland-Overberg and Southern Cape Karoo where 40% of the drug resistant cases were MDR-TB (58). However 70% of this drug resistance epidemic was being driven by four strain families only (F11, F28, F29 and *IS6110* low-copy-number clade), the most prevalent being the Beijing/W like and IS*6110 l*ow-copy-number clade (58). The IS*6110*-low-copy number strains are widespread throughout the region studied but a particular group with a characteristic spoligotype (spoligotype 115) was only isolated from one town, possibly implying

an outbreak of MDR-TB (58). The Beijing strains have been associated with drug resistance and MDR-TB in other regions of the Western Cape (62,74) and also worldwide (49,73).

Despite the limited amount of *IS6110*-RFLP data from KwaZulu-Natal Province of South Africa, it was suggested that recent transmission was not contributing significantly to the drug resistant TB problem (89). A substantial increase in the rate of MDR-TB was reported from the same community and *IS6110*-RFLP analysis suggested that between 38-67% of the MDR-TB cases were recently transmitted (88). However there was no evidence of a dominant circulating strain. On the contrary, in the gold mining communities of South Africa, another community with very high rates of TB, MDR-TB isolates were less likely to be clustered than drug susceptible isolates, suggesting that MDR-TB isolates caused fewer secondary cases of TB than drug susceptible ones (31).

Results from the study by van Rie and colleagues (14) further suggested that in an epidemic area, a large proportion of TB drug resistance traditionally classified as acquired may be due to recent transmission (14). However in a setting with high rates of TB transmission and HIV-1 infection, Sonnenberg and colleagues (90)showed that the dominant mechanism of drug resistance while on treatment was acquisition rather than transmission. These findings would therefore imply that the term acquired drug resistance is often misleading, especially when relying only on history of previous treatment. This is because cases classified as having acquired drug resistance, include cases with drug resistance due to true acquisition as well as those due to reinfection with a drug resistant

strain (transmission). This therefore results in the misclassification of the latter group as having acquired drug resistance when in fact they would have recently been reinfected with a drug resistant strain. This misclassification may lead to incorrect treatment regimens being prescribed and these cases may remain infectious for long periods of time, thereby contributing to the pool of infectious cases resulting in increased transmission of drug resistant TB.

5.4 Tuberculosis infection in individual cases

Historically a repeat episode (recurrent) of TB was thought to be due to endogenous reactivation. However, as far back as 1978, exogenous reinfection was suspected to be a contributor to the burden of TB (91). Understanding these mechanisms and their contribution to recurrent disease has extremely important implications for 1) drug trials 2) patient management 3) vaccine development. Exogenous reinfection and relapse are clinically indistinguishable and identification of these causes of recurrence requires the ability to differentiate the strains involved in the first and the subsequent TB episode.

By distinguishing between strains with the same *IS6110*-RFLP fingerprint from those with different fingerprints, molecular techniques using *IS6110* have allowed documentation of exogenous reinfection and now provide us with the tools to determine the mechanisms leading to recurrence (60,92-97). A recurrence of TB is the development of a second episode of active TB in a previously infected individual who had been considered cured. This therefore raises the question whether this represents a recrudescence of the initially infecting organism (endogenous reactivation/relapse) or

infection with a new strain of *M. tuberculosis* (exogenous reinfection) (98). Relapse can occur when TB bacilli persist after treatment despite apparent cure (true relapse) mainly due to inadequate treatment regimen and treatment duration (99). Exogenous reinfection on the other hand is reinfection with a second strain of *M. tuberculosis* after contact with an infectious case and this can occur after curative treatment and/or in individuals who are latently infected (98, 99). The risk of reinfection therefore depends on the background TB incidence in communities and the risk of it breaking down into a second episode of disease (99).

5.5 Exogenous re-infection after curative treatment

Molecular epidemiology has provided dramatic evidence for the mechanisms of recurrence. Anecdotal evidence initially showed that immuno-compromised individuals may be reinfected with different strains after curative treatment (95). Subsequently, data from a high TB incidence epidemiological field site in Cape Town South Africa showed that 75% of retreatment episodes in HIV negative individuals were due to reinfection (60). From this study it was concluded that the initial episode of disease was unable to confer protection against subsequent infections and disease progression. Furthermore, this study highlighted the importance of limiting active transmission which places previously treated individuals at risk of reinfection. However, this observation is not unique to high incidence settings, as more recently, a number of reports have shown that both persistent cases and retreatment cases may result from exogenous reinfection (92,93,95-97)with drug sensitive and resistant strains.

Reinfection has now been equated to a paradigm shift in our understanding of TB pathogenesis. Interestingly the relative proportion of recurrent cases due to reinfection differs in different settings in South Africa. A lower percentage of recurrence due to exogenous reinfection was observed in the gold mines of South Africa (97). By following up 326 South African mineworkers who had successfully completed treatment for pulmonary TB in 1995 and combining classic and molecular epidemiology techniques, it was shown that approximately 20% had a second episode of TB (97). Of these recurrent TB cases, 64% were due to relapse and 36% due to exogenous reinfection. Human immunodeficiency virus was a risk factor for recurrence and HIV-1 was more strongly associated with reinfection but not with relapse (97). Due to immuno-suppression associated with, HIV-1 infection, individuals who are HIV-1 positive may be at high risk of developing infection after exposure to *M tuberculosis*. Exposure to *M tuberculosis* bacilli in turn increases in areas with high TB incidence. Following exposure to *M. tuberculosis*, HIV-1 infected individuals rapidly progress to disease further lowering their immunity to subsequent infection. The high recurrence rate seen in HIV-1 related TB in Africa may be due to increased susceptibility to exogenous reinfection and not treatment failure (94).

5.6 Exogenous reinfection in latently infected individuals

Tuberculosis has always been and still is thought by many to result from a single strain of *M. tuberculosis* and that this infection would protect against subsequent infections. In areas with a high incidence of TB, exogenous reinfection might also be a cause of the first episode of post-primary TB, since the immunity that develops after primary infection followed by a period of latency cannot be expected to confer protection against

subsequent exogenous reinfection. The high TB incidence being experienced in several regions may imply that many individuals may be exposed more than once and therefore exogenous reinfection may be an extremely important mechanism driving secondary TB in latently infected individuals. However this is only based on assumption because it is not possible to genotype the strains in latently infected individuals. Mathematical modelling supports this notion suggesting that as many as 35 % of adults by the age of 35 years will have been infected more than once assuming an annual risk of tuberculosis infection of 3.5 % (P. Uys, personal communication).

Studies have shown that both HIV-negative and HIV-positive individuals can be infected with more than one strain during the same episode (multiple infections) (100,101) in different lesions (102) or during successive TB episodes (reinfection) (60,92,93,95-97). Analysis of lung lesions clearly demonstrated that individuals with TB may have been reinfected due to the high annual risk of infection prior to disease presentation(102). The central dogma that a patient can only be infected with one *M. tuberculosis* strain has been further challenged. Warren and colleagues (103) published the first report of the use of a polymerase chain reaction (PCR)-based method to determine the extent of multiple *M. tuberculosis* infection in sputum specimens collected from both new and retreatment TB cases in a high TB incidence community. They demonstrated that 19% of all the patients were simultaneously infected with a Beijing and non Beijing strain. Multiple strains were demonstrated to occur more frequently in retreatment cases (23%) as compared to new cases (17%).

There is growing evidence which leads us to suspect that some individuals with TB, especially in the high incidence communities, are infected with multiple strains of *M. tuberculosis*, (100-103) which could either be drug sensitive or resistant. In such cases DST is complicated as only one strain may be detectable or different antibiograms in different specimens collected from the same patient, complicating the interpretation and management of patients. These mixed infections may be detected, especially where the bacterial load is high. As with exogenous reinfection it is conceivable that the chances of mixed infection occurring in individuals are directly related to the local TB incidence and rate of infection.

The use of targets that are more specific to other strains may in future help to provide more evidence and the true frequency of mixed infection which has not yet been possible with the currently standardized molecular biology methods. It is imperative to identify these mixed infections in molecular epidemiology because they may influence the interpretation of molecular epidemiology data. Most importantly the study by Warren and colleagues (103), substantiates the argument that prior infection with a pathogenic form of TB does not confer complete protection in this population, questioning the efficacy of the currently used *Mycobacterium bovis* BCG vaccine.

How mixed infections in TB could influence the epidemiology of the disease remains to be determined, although it has been suggested that reinfection may exacerbate disease progression and endogenous reactivation of primary infection (102). This would imply that primary infection is unable to confer protection against subsequent infection, thereby

even further questioning the efficacy of the currently used vaccines. Furthermore these findings will have implications on the interpretation of findings from drug and vaccine trials. It is not clear if patients with mixed infections differ in clinical presentation or response to treatment, which would imply different clinical management of patients. Even though it is conceivable that mixed infections would be low in low incidence communities, their frequency needs to be determined both in high and low incidence communities to allow for accurate inferences to be made from molecular epidemiologic studies.

6.0 Lessons learnt

Strain typing of *M tuberculosis* in our region has revolutionized the epidemiology of TB by providing information on the variety and number of strains circulating in our communities. Within these communities the data has also been used to track the spread of both drug sensitive and drug resistant strains (including MDR strains), to investigate recurrent *M. tuberculosis* infections and to examine the existence of multiple infections. Not only has it provided information on disease dynamics, the data has potential use in determining the evolutionary mechanisms in this species.

A better understanding of these evolutionary mechanisms modeling the genome will provide new insights for the interpretation of molecular epidemiology data. In addition, information from DNA fingerprinting of *M. tuberculosis* strains contributes to the development of new mathematical models used in the study of tuberculosis transmission. This increases the accuracy of the mathematical models in identifying factors that may

have an effect on transmission, which in turn increases the reliability of predictions and design of intervention strategies.

7.0 Hypothesis and aim of the study

We hypothesized that the application of molecular epidemiological tools such as IS6110, spoligotyping and VNTR will help us understand the disease dynamics *M. tuberculosis* in high TB incidence settings in Southern Africa. The general aim of this study is to apply IS6110, spoligotyping and other molecular techniques to study the following aspects of the TB epidemic: the diversity and population structure of drug sensitive and multidrug and extensively resistant tuberculosis, risk factors for disease due to predominant *M. tuberculosis* strains and the evolution of *M. tuberculosis* strains in regions of Southern Africa.

8.0 Acknowledgements

The authors would like thank the South African National Research Foundation (GUN 2054278), IAEA (SAF6008), The Welcome Trust (072402/Z/03/Z and 01680/Z/00/Z) and the NIH (R21 A155800-01) for support.

9.0 References

1. **Corbett, E. L., C. J. Watt, N. Walker, D. Maher, B. G. Williams, M. C. Raviglione, and C. Dye**. 2003. The growing burden of tuberculosis: global trends and interactions with the HIV epidemic. Arch.Intern.Med. **163**:1009-1021.
2. **Dye, C., S. Scheele, P. Dolin, V. Pathania, and M. C. Raviglione**. 1999. Consensus statement. Global burden of tuberculosis: estimated incidence, prevalence, and mortality by country. WHO Global Surveillance and Monitoring Project. JAMA **282**:677-686.

3. **Frieden, T. R., T. R. Sterling, S. S. Munsiff, C. J. Watt, and C. Dye**. 2003. Tuberculosis. Lancet **362**:887-899.

4. **Global Tuberculosis Control,** WHO Report 2002. WHO/CDS/TB/2002.295. http://www.who.int/gtb/publications/globrep02/.

5. **World Health Organization**. Anti-tuberculosis drug resistance in the world. Third Global Report. WHO/IUATLD Global project on anti-tuberculosis drug resistance surveillance 1999-2002. http://www.who.int/gtb/publications/drug resistance/2004/drs_report_1/

6. **UNAIDS Report** on global AIDS epidemic. July 2004. http://www.unaids.org/bangkok2004/gar2004_html/GAR2004_14_en.htm

7. **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 [see comments] [published erratum appears in Lancet 1999 May 15;353(9165):1714]. Lancet **353**:444-449.

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

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

10. **Parrish, N. M., J. D. Dick, and W. R. Bishai**. 1998. Mechanisms of latency in *Mycobacterium tuberculosis*. Trends Microbiol. **6**:107-112.

11. **Murray, J. F.** 2004. A century of tuberculosis. Am.J.Respir.Crit Care Med. **169**:1181-1186.

12. **Jacobs, R. F.** 1994. Multiple-drug-resistant tuberculosis. Clin.Infect.Dis. **19**:1-8.

13. **WHO/IUATLD** Global working group on Antituberculosis Drug Resistance Surveillance.1998. Guidelines for surveillance of drug resistance in tuberculosis. WHO Geneva/IUATLD Paris. Int.J.Tuberc.Lung Dis. **2**:72-89.

14. **van Rie, A., R. Warren, M. Richardson, R. P. Gie, D. A. Enarson, N. Beyers, and P. D. van Helden**. 2000. Classification of drug-resistant tuberculosis in an epidemic area. Lancet **356**:22-25.

15. **Espinal, M. A.** 2003. The global situation of MDR-TB. Tuberculosis.(Edinb.) **83**:44-51.

16. **Rieder, H. L.** 2003. Contacts of tuberculosis patients in high-incidence countries. Int.J.Tuberc.Lung Dis. **7**:S333-S336.

17. **Veen, J.** 1992. Microepidemics of tuberculosis: the stone-in-the-pond principle. Tuber.Lung Dis. **73**:73-76.

18. **Daley, C. L.** 2004. Tuberculosis contact investigations: please don't fail me now. Am.J.Respir.Crit Care Med. **169**:779-781.

19. **Marks, S. M., Z. Taylor, N. L. Qualls, R. J. Shrestha-Kuwahara, M. A. Wilce, and C. H. Nguyen**. 2000. Outcomes of contact investigations of infectious tuberculosis patients. Am.J.Respir.Crit Care Med. **162**:2033-2038.

20. **Dasgupta, K., K. Schwartzman, R. Marchand, T. N. Tennenbaum, P. Brassard, and D. Menzies**. 2000. Comparison of cost-effectiveness of tuberculosis screening of close contacts and foreign-born populations. Am.J.Respir.Crit Care Med. **162**:2079-2086.

21. **Rothenberg, R. B., P. D. McElroy, M. A. Wilce, and S. Q. Muth**. 2003. Contact tracing: comparing the approaches for sexually transmitted diseases and tuberculosis. Int.J.Tuberc.Lung Dis. **7**:S342-S348.

22. **WHO, HIV/TB**. Guidelines for intensified tuberculosis case-finding and isoniazid preventive therapy for people living with HIV in resource constrained settings. http://whqlibdoc.who.int/publications/2011/9789241500708_eng.pdf Accessed 13 January 2011

23. **Mehta, J., C. Diaz, B. Guha, L. Harvill, and G. Krishnaswamy**. 1995. Appearance of drug-resistant tuberculosis in rural Tennessee. South.Med.J. **88**:60-64.

24. **Bayona, J., A. M. Chavez-Pachas, E. Palacios, K. Llaro, R. Sapag, and M. C. Becerra**. 2003. Contact investigations as a means of detection and timely treatment of persons with infectious multidrug-resistant tuberculosis. Int.J.Tuberc.Lung Dis. **7**:S501-S509.

25. **Teixeira, L., M. D. Perkins, J. L. Johnson, R. Keller, M. Palaci, V. D. do, V, L. M. Canedo Rocha, S. Debanne, E. Talbot, and R. Dietze**. 2001. Infection and disease among household contacts of patients with multidrug-resistant tuberculosis. Int.J.Tuberc.Lung Dis. **5**:321-328.

26. **Valway, S. E., M. P. Sanchez, T. F. Shinnick, I. Orme, T. Agerton, D. Hoy, J. S. Jones, H. Westmoreland, and I. M. Onorato**. 1998. An outbreak involving extensive transmission of a virulent strain of *Mycobacterium tuberculosis* [see comments] [published erratum appears in N Engl J Med 1998 Jun 11;338(24):1783]. N.Engl.J.Med. **338**:633-639.

27. **Onorato, I. M.** 2000. Tuberculosis outbreaks in the United States. Int.J.Tuberc.Lung Dis. **4**:S121-S126.

28. **Behr, M. A., P. C. Hopewell, E. A. Paz, L. M. Kawamura, G. F. Schecter, and P. M. Small**. 1998. Predictive value of contact investigation for identifying recent transmission of *Mycobacterium tuberculosis*. Am.J.Respir.Crit Care Med. **158**:465-469.

29. **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.

30. **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 [see comments]. N.Engl.J.Med. **330**:1710-1716.

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

32. **Lockman, S., J. D. Sheppard, C. R. Braden, M. J. Mwasekaga, C. L. Woodley, T. A. Kenyon, N. J. Binkin, M. Steinman, F. Montsho, M. Kesupile-Reed, C. Hirschfeldt, M. Notha, T. Moeti, and J. W. Tappero**. 2001. Molecular and conventional epidemiology of *Mycobacterium tuberculosis* in Botswana: a

population-based prospective study of 301 pulmonary tuberculosis patients. J.Clin.Microbiol. **39**:1042-1047.

33. **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 [see comments]. S.Afr.Med.J. **86**:45-49.

34. **Cave, M. D., K. D. Eisenach, P. F. McDermott, J. H. Bates, and J. T. Crawford**. 1991. IS6110: conservation of sequence in the *Mycobacterium tuberculosis* complex and its utilization in DNA fingerprinting. Mol.Cell Probes **5**:73-80.

35. **van Soolingen, D., P. W. Hermans, P. E. de Haas, and J. D. van Embden**. 1992. Insertion element IS1081-associated restriction fragment length polymorphisms in *Mycobacterium tuberculosis* complex species: a reliable tool for recognizing Mycobacterium bovis BCG. J.Clin.Microbiol. **30**:1772-1777.

36. **van Soolingen, D., P. E. de Haas, P. W. Hermans, P. M. Groenen, and J. D. van Embden**. 1993. Comparison of various repetitive DNA elements as genetic markers for strain differentiation and epidemiology of *Mycobacterium tuberculosis*. J.Clin.Microbiol. **31**:1987-1995.

37. **Hermans, P. W., D. van Soolingen, E. M. Bik, P. E. de Haas, J. W. Dale, and J. D. van Embden**. 1991. 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. **59**:2695-2705.

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

39. **Mazars, E., S. Lesjean, A. L. Banuls, M. Gilbert, V. Vincent, B. Gicquel, M. Tibayrenc, C. Locht, and P. Supply**. 2001. High-resolution minisatellite-based

typing as a portable approach to global analysis of *Mycobacterium tuberculosis* molecular epidemiology. Proc.Natl.Acad.Sci.U.S.A **98**:1901-1906.

40. **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.

41. **Warren, R., M. Richardson, S. G. van der, 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.

42. **Yuen, L. K., B. C. Ross, K. M. Jackson, and B. Dwyer**. 1993. Characterization of *Mycobacterium tuberculosis* strains from Vietnamese patients by Southern blot hybridization. J.Clin.Microbiol. **31**:1615-1618.

43. **Agasino, C. B., d. L. Ponce, R. M. Jasmer, and P. M. Small**. 1998. Epidemiology of *Mycobacterium tuberculosis* strains in San Francisco that do not contain IS6110. Int.J.Tuberc.Lung Dis. **2**:518-520.

44. **Fomukong, N., M. Beggs, H. el Hajj, G. Templeton, K. Eisenach, and M. D. Cave**. 1997. Differences in the prevalence of IS6110 insertion sites in *Mycobacterium tuberculosis* strains: low and high copy number of IS6110. Tuber.Lung Dis. **78**:109-116.

45. **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.

46. **Douglas, J. T., L. Qian, J. C. Montoya, J. M. Musser, J. D. van Embden, D. van Soolingen, and K. Kremer**. 2003. Characterization of the Manila family of *Mycobacterium tuberculosis*. J.Clin.Microbiol. **41**:2723-2726.

47. **Brudley K, Driscoll JR, Leen Rigouts, Prodinger WM, Gori A, Hajoj SA, Allix C, Aristimuno L, *et al*.** *Mycobacterium, tuberculosis* complex genetic diversity: mining the fourth international spoligotyping database (SpolDB4) for classification, population genetics and epidemiology. *BMC Microbiology* 2006;6:23

48. **Sreevatsan, S., X. Pan, K. E. Stockbauer, N. D. Connell, B. N. Kreiswirth, T. S. Whittam, and J. M. Musser**. 1997. Restricted structural gene polymorphism in the *Mycobacterium tuberculosis* complex indicates evolutionarily recent global dissemination. Proc.Natl.Acad.Sci.U.S.A **94**:9869-9874.

49. **Agerton, T., S. Valway, B. Gore, C. Pozsik, B. Plikaytis, C. Woodley, and I. Onorato**. 1997. Transmission of a highly drug-resistant strain (strain W1) of *Mycobacterium tuberculosis*. Community outbreak and nosocomial transmission via a contaminated bronchoscope [see comments]. JAMA **278**:1073-1077.

50. **Wilkinson, D., J. Crump, M. Pillay, and A. W. Sturm**. 1997. Nosocomial transmission of tuberculosis in Africa documented by restriction fragment length polymorphism. Trans.R.Soc.Trop.Med.Hyg. **91**:318.

51. **Beck-Sague, C., S. W. Dooley, M. D. Hutton, J. Otten, A. Breeden, J. T. Crawford, A. E. Pitchenik, C. Woodley, G. Cauthen, and W. R. Jarvis**. 1992. Hospital outbreak of multidrug-resistant *Mycobacterium tuberculosis* infections. Factors in transmission to staff and HIV-infected patients. JAMA **268**:1280-1286.

52. **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.

53. **Wilkinson, D., M. Pillay, J. Crump, C. Lombard, G. R. Davies, and A. W. Sturm**. 1997. Molecular epidemiology and transmission dynamics of *Mycobacterium tuberculosis* in rural Africa. Trop.Med.Int.Health **2**:747-753.

54. **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 [see comments]. J.Infect.Dis. **171**:1504-1513.

55. **Heyderman, R. S., M. Goyal, P. Roberts, S. Ushewokunze, S. Zizhou, B. G. Marshall, R. Makombe, J. D. van Embden, P. R. Mason, and R. J. Shaw**. 1998. Pulmonary tuberculosis in Harare, Zimbabwe: analysis by spoligotyping [see comments]. Thorax **53**:346-350.

56. **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.

57. **Godfrey-Faussett, P.** 1999. Interpretation of cluster studies of tuberculosis [comment]. Lancet **353**:427-428.

58. **Streicher, E. M., R. M. Warren, C. Kewley, J. Simpson, N. Rastogi, C. Sola, G. D. van der Spuy, P. D. van Helden, and T. C. Victor**. 2004. Genotypic and phenotypic characterization of drug-resistant *Mycobacterium tuberculosis* isolates from rural districts of the Western Cape Province of South Africa. J.Clin.Microbiol. **42**:891-894.

59. **Victor, T. C., P. E. de Haas, A. M. Jordaan, G. D. van der Spuy, M. Richardson, D. van Soolingen, P. D. van Helden, and R. Warren**. 2004. Molecular

characteristics and global spread of *Mycobacterium tuberculosis* with a western cape F11 genotype. J.Clin.Microbiol. **42**:769-772.

60. **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.

61. **Pillay, T., D. G. Pillay, M. Pillay, M. Adhikari, and A. W. Sturm**. 1999. Are specific strains of *mycobacterium tuberculosis* responsible for disease in newborns? Pediatr.Infect.Dis.J. **18**:844-845.

62. **Richardson, M., S. W. van Lill, G. D. van der Spuy, Z. Munch, C. N. Booysen, N. Beyers, P. D. van Helden, and R. M. Warren**. 2002. Historic and recent events contribute to the disease dynamics of Beijing-like *Mycobacterium tuberculosis* isolates in a high incidence region. Int.J.Tuberc.Lung Dis. **6**:1001-1011.

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

64. **Nguyen, D., P. Brassard, J. Westley, L. Thibert, M. Proulx, K. Henry, K. Schwartzman, D. Menzies, and M. A. Behr**. 2003. Widespread pyrazinamide-resistant *Mycobacterium tuberculosis* family in a low-incidence setting. J.Clin.Microbiol. **41**:2878-2883.

65. **Nguyen, D., P. Brassard, D. Menzies, L. Thibert, R. Warren, S. Mostowy, and M. Behr**. 2004. Genomic characterization of an endemic *Mycobacterium tuberculosis* strain: evolutionary and epidemiologic implications. J.Clin.Microbiol. **42**:2573-2580.

66. **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.

67. **Toungoussova, O. S., A. Mariandyshev, G. Bjune, P. Sandven, and D. A. Caugant**. 2003. Molecular epidemiology and drug resistance of *Mycobacterium tuberculosis* isolates in the Archangel prison in Russia: predominance of the W-Beijing clone family. Clin.Infect.Dis. **37**:665-672.

68. **Anh, D. D., M. W. Borgdorff, L. N. Van, N. T. Lan, T. van Gorkom, K. Kremer, and D. van Soolingen**. 2000. *Mycobacterium tuberculosis* Beijing genotype emerging in Vietnam. Emerg.Infect.Dis. **6**:302-305.

69. **Chan, M. Y., M. Borgdorff, C. W. Yip, P. E. de Haas, W. S. Wong, K. M. Kam, and D. van Soolingen**. 2001. Seventy percent of the *Mycobacterium tuberculosis* isolates in Hong Kong represent the Beijing genotype. Epidemiol.Infect. **127**:169-171.

70. **Moss, A. R., D. Alland, E. Telzak, D. Hewlett, Jr., V. Sharp, P. Chiliade, V. LaBombardi, D. Kabus, B. Hanna, L. Palumbo, K. Brudney, A. Weltman, K. Stoeckle, K. Chirgwin, M. Simberkoff, S. Moghazeh, W. Eisner, M. Lutfey, and B. Kreiswirth**. 1997. A city-wide outbreak of a multiple-drug-resistant strain of *Mycobacterium tuberculosis* in New York. Int.J.Tuberc.Lung Dis. **1**:115-121.

71. **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.

72. **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.

73. **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.

74. **Githui, W. A., A. M. Jordaan, E. S. Juma, P. Kinyanjui, F. G. Karimi, J. Kimwomi, H. Meme, P. Mumbi, E. M. Streicher, R. Warren, P. D. van Helden, and T. C. Victor**. 2004. Identification of MDR-TB Beijing/W and other *Mycobacterium tuberculosis* genotypes in Nairobi, Kenya. Int.J.Tuberc.Lung Dis. **8**:352-360.

75. **Warren , R.M., Victor T., Streicher, E. M., Richardson M., van der Spuy G.D., Johnson R, Chihota V.N., Locht C., SUpply, P. and van Helden, P.D.** 2004. Clonal expansion of a globally disseminated lineage of *Mycobacterium tuberculosis* with low IS6110 copy numbers. J *Clin Microbiol* 42: 5774-5782

76. **Dale, J. W., H. Al Ghusein, S. Al Hashmi, P. Butcher, A. L. Dickens, F. Drobniewski, K. J. Forbes, S. H. Gillespie, D. Lamprecht, T. D. McHugh, R. Pitman, N. Rastogi, A. T. Smith, C. Sola, and H. Yesilkaya**. 2003. Evolutionary relationships among strains of *Mycobacterium tuberculosis* with few copies of IS6110. J.Bacteriol. **185**:2555-2562.

77. **Sampson, S. L., R. M. Warren, M. Richardson, G. D. van der Spuy, and P. D. van Helden**. 1999. Disruption of coding regions by IS6110 insertion in *Mycobacterium tuberculosis*. Tuber.Lung Dis. **79**:349-359.

78. **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.

79. **Cowan, L. S., L. Mosher, L. Diem, J. P. Massey, and J. T. Crawford**. 2002. Variable-number tandem repeat typing of *Mycobacterium tuberculosis* isolates with low copy numbers of IS6110 by using mycobacterial interspersed repetitive units. J.Clin.Microbiol. **40**:1592-1602.

80. **Bauer, J., A. B. Andersen, K. Kremer, and H. Miorner**. 1999. Usefulness of spoligotyping To discriminate IS6110 low-copy-number *Mycobacterium tuberculosis* complex strains cultured in Denmark. J.Clin.Microbiol. **37**:2602-2606.

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

82. **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 [see comments]. N.Engl.J.Med. **330**:1703-1709.

83. **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.

84. **Niemann, S., S. Rusch-Gerdes, E. Richter, H. Thielen, H. Heykes-Uden, and R. Diel**. 2000. Stability of IS6110 Restriction Fragment Length Polymorphism Patterns of *Mycobacterium tuberculosis* Strains in Actual Chains of Transmission. J.Clin.Microbiol. **38**:2563-2567.

85. **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 IS6110-based restriction fragment length polymorphism pattern during the transmission of *Mycobacterium tuberculosis*. Journal of Clinical Microbiology **40**:1277-1282.

86. **Braden, C. R., G. L. Templeton, M. D. Cave, S. Valway, I. M. Onorato, K. G. Castro, D. Moers, Z. Yang, W. W. Stead, and J. H. Bates**. 1997. Interpretation of restriction fragment length polymorphism analysis of *Mycobacterium tuberculosis* isolates from a state with a large rural population. J.Infect.Dis. **175**:1446-1452.

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

88. **Davies, G. R., M. Pillay, A. W. Sturm, and D. Wilkinson**. 1999. Emergence of multidrug-resistant tuberculosis in a community-based directly observed treatment programme in rural South Africa. Int.J.Tuberc.Lung Dis. **3**:799-804.

89. **Wilkinson, D., M. Pillay, G. R. Davies, and A. W. Sturm**. 1996. Resistance to antituberculosis drugs in rural South Africa: rates, patterns, risks, and transmission dynamics. Trans.R.Soc.Trop.Med.Hyg. **90**:692-695.

90. **Sonnenberg, P., J. Murray, S. Shearer, J. R. Glynn, B. Kambashi, and P. Godfrey-Faussett**. 2000. Tuberculosis treatment failure and drug resistance--same strain or reinfection? Trans.R.Soc.Trop.Med.Hyg. **94**:603-607.

91. **Styblo, K.** 1978. [Current status of the problem. I. Epidemiology of tuberculosis]. Bull.Int.Union Tuberc. **53**:153-166.

92. **Bandera, A., A. Gori, L. Catozzi, E. A. degli, G. Marchetti, C. Molteni, G. Ferrario, L. Codecasa, V. Penati, A. Matteelli, and F. Franzetti**. 2001. Molecular

epidemiology study of exogenous reinfection in an area with a low incidence of tuberculosis. J.Clin.Microbiol. **39**:2213-2218.

93. **Caminero, J. A., M. J. Pena, M. I. Campos-Herrero, J. C. Rodriguez, O. Afonso, C. Martin, J. M. Pavon, M. J. Torres, M. Burgos, P. Cabrera, P. M. Small, and D. A. Enarson**. 2001. Exogenous reinfection with tuberculosis on a European island with a moderate incidence of disease. Am.J.Respir.Crit Care Med. **163**:717-720.

94. **Godfrey-Faussett, P., W. Githui, B. Batchelor, R. Brindle, J. Paul, M. Hawken, S. Gathua, J. Odhiambo, S. Ojoo, P. Nunn, and .** 1994. Recurrence of HIV-related tuberculosis in an endemic area may be due to relapse or reinfection. Tuber.Lung Dis. **75**:199-202.

95. **Nardell, E., B. McInnis, B. Thomas, and S. Weidhaas**. 1986. Exogenous reinfection with tuberculosis in a shelter for the homeless. N.Engl.J.Med. **315**:1570-1575.

96. **Small, P. M., R. W. Shafer, P. C. Hopewell, S. P. Singh, M. J. Murphy, E. Desmond, M. F. Sierra, and G. K. Schoolnik**. 1993. Exogenous reinfection with multidrug-resistant *Mycobacterium tuberculosis* in patients with advanced HIV infection. N.Engl.J.Med. **328**:1137-1144.

97. **Sonnenberg, P., J. Murray, J. R. Glynn, S. Shearer, B. Kambashi, and P. Godfrey-Faussett**. 2001. HIV-1 and recurrence, relapse, and reinfection of tuberculosis after cure: a cohort study in South African mineworkers. Lancet **358**:1687-1693.

98. **Fine, P. E. and P. M. Small**. 1999. Exogenous reinfection in tuberculosis. N.Engl.J.Med. **341**:1226-1227.

99. **Lambert, M. L., E. Hasker, A. Van Deun, D. Roberfroid, M. Boelaert, and S. P. Van der**. 2003. Recurrence in tuberculosis: relapse or reinfection? Lancet Infect.Dis. **3**:282-287.

100. **Braden, C. R., G. P. Morlock, C. L. Woodley, K. R. Johnson, A. C. Colombel, M. D. Cave, Z. H. Yang, S. E. Valway, I. M. Onorato, and J. T. Crawford**. 2001. Simultaneous infection with multiple strains of *Mycobacterium tuberculosis*. *Clinical Infectious Diseases* **33**:E42-E47.

101. **Richardson, M., N. M. Carroll, E. Engelke, G. D. van der Spuy, F. Salker, Z. Munch, R. P. Gie, R. M. Warren, N. Beyers, and P. D. van Helden**. 2002. Multiple *Mycobacterium tuberculosis* Strains in Early Cultures from Patients in a High-Incidence Community Setting. J.Clin.Microbiol. **40**:2750-2754.

102. **du Plessis, D. G., R. Warren, M. Richardson, J. J. Joubert, and P. D. van Helden**. 2001. Demonstration of reinfection and reactivation in HIV-negative autopsied cases of secondary tuberculosis: multilesional genotyping of *Mycobacterium tuberculosis* utilizing IS6110 and other repetitive element-based DNA fingerprinting. Tuberculosis **81**:211-220.

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

Predominance of a single genotype of *Mycobacterium tuberculosis* in regions of Southern Africa[2]

Violet Chihota, Ludwig Apers, Stanley Mungofa, Webster Kasongo, Irene M. Nyoni, Ruth Tembwe, Grace Mbulo, Mathias Tembo, Elizabeth M. Streicher Gian D. van der Spuy, Thomas C. Victor, Paul D. van Helden and Robin M. Warren

[2] Published as an original article: **Chihota V**, Apers L, Mungofa S, Kasongo W, Streicher E.M, van der Spuy GD, Victor T, van Helden P.D and Warren RM. Predominance of the SAF1 genotype of *Mycobacterium tuberculosis* in regions of Southern Africa. *I JTLD* 2007; Mar 11: 311-318

Summary

***Setting*:** Zimbabwe and Zambia

Objective: To determine the genetic diversity of *Mycobacterium tuberculosis* strains isolated from tuberculosis (TB) patients in Zimbabwe and Zambia.

***Design*:** *M. tuberculosis* isolates cultured from TB patients presenting at referral hospitals in Zimbabwe and healthcare-clinics in Zambia were characterized by IS*6110*-genotyping and/or spoligotyping using internationally standardized methods. Genotypic data was compared to that from Cape Town and SpolDB3.0 database.

Results: A predominant group of strains could be identified among 116 of 246 (47.2%) Zimbabwean isolates, by their characteristic IS*6110*-banding pattern and unique spoligotype- signature, where spacers 21-24, 27-30 and 33-36 were deleted. Comparison with strains from Cape Town showed that they were closely related to a family of strains present in 2.3% of Cape Town patients. Comparison of the spoligotypes with those obtained from 114 isolates from Zambia showed that 74 (65%) of these isolates had the same spoligotype-signature. Spoligotypes on the SpolDB3.0 database showed that this group of strains was rarely isolated from other parts of the world but commonly isolated in Southern-Africa.

***Conclusion*:** A predominant group of strains infecting approximately half of the patients in the study are major contributors to the TB epidemic in this region. We have designated this group of strains the Southern Africa 1"SAF1" family.

Introduction

In terms of its historical and current disease burden *Mycobacterium tuberculosis,* the etiological agent for tuberculosis (TB), is undeniably the most successful human pathogen. Approximately one third of the world's population is infected by *M. tuberculosis*; however, less than 10% of these develop active TB during their lifetime. Sub-Saharan Africa has the highest annual TB incidence rates since the emergence of the human immunodeficiency virus (HIV) epidemic and most countries in this region belong to the group of 22 high burden countries that collectively account for 80% of cases worldwide (1).

In Zimbabwe, the incidence of TB is estimated at 604 new cases per 100 000 population (2). The situation is exacerbated by co-infection with the HIV with approximately 60% of cases presenting with tuberculosis, also being co-infected with HIV (2). Sputum smear microscopy is the mainstay for diagnosis of pulmonary TB where culture is not available. However it is well established that smear negative cases may be disproportionately higher in HIV positive than in HIV negative individuals(3;4) and hence the true denominator of total cases is largely unknown. Although countries in Southern Africa have by far the largest TB disease burden, very little is known about the disease dynamics and characteristics of the bacterial populations circulating in these communities.

Approaches to understanding the disease dynamics of TB have included the use of molecular methods together with conventional TB epidemiology (5-7). Molecular epidemiology techniques have enabled differentiation of clinical isolates of *M.*

tuberculosis into distinct genotypes which may be important in understanding bacterial pathogenesis. The most commonly used methods being IS*6110*-genotyping (8). The use of additional markers such as spoligotyping (9) and mycobacterial interspersed repetitive units(MIRUs) (10, 11) has greatly enhanced the accuracy in differentiating *M. tuberculosis* strains.

The lack of comprehensive molecular epidemiological data from most countries in Africa, such as Zimbabwe and Zambia, has limited the understanding of TB disease dynamics in these areas. A molecular epidemiology study conducted in Harare reported a high level of strain clustering (12). However no comprehensive studies have been done in Zimbabwe to describe the degree of genetic heterogeneity of *M. tuberculosis* isolates circulating in this region. Earlier attempts to characterize *M. tuberculosis* strains from Zimbabwe using spoligotyping, were limited by small sample size, use of one marker, and study period (13). In the absence of such information it is possible that the molecular epidemiology data may have been misinterpreted.

In this study we have used the internationally standardized IS*6110*-genotyping method together with spoligotyping to determine the genetic diversity of *M. tuberculosis* strains isolated from TB patients in Harare and Gweru, Zimbabwe. Analysis of this data identified a predominant strain family contributing to the high TB incidence rates. In order to gain an insight into the global significance of this strain family, genotypic data from Zimbabwe was compared to that from other countries in the Southern Africa region, thereby allowing an assessment of how molecular epidemiology could be used to

determine the dynamics of the current TB epidemic, such as transmission of the disease between communities, regions and countries.

MATERIALS AND METHODS

Recruitment of patients

The *M. tuberculosis* strains were cultured from patients presenting with pulmonary TB in two cities in Zimbabwe. During the period October to December 2001 sputum samples were collected from a total of 120 consecutive smear-positive pulmonary tuberculosis patients presenting at the Beatrice Road Infectious Diseases Hospital (BRIDH), (the main referral centre for infectious diseases in Harare, Zimbabwe). During the period September 2000 to September 2001 sputum samples were collected from 300 consecutive smear-positive or smear- negative pulmonary TB patients presenting at the Gweru Provincial Hospital (GPH).

The Zambian isolates were collected as part of a national drug resistance surveillance study and were obtained from new smear positive patients presenting at various medical centres in different provinces of Zambia.

This study was approved by the Medical Research Council of Zimbabwe (MRCZ/A/895, MRCZ/A/967) and the Stellenbosch University Faculty of Health Sciences (2003/022/N).

Bacterial strains

The sputum samples from BRIDH were sent to Biomedical Research and Training Institute (BRTI) in Harare, for sputum microscopy and subsequent culture on Lowenstein

Jensen (LJ) slants using conventional methods. Sputum samples from GPH were also sent to BRTI for sputum microscopy and culture on MGIT (14). All positive cultures were sub-cultured onto LJ slants and allowed to grow at 37ºC until confluent growth was observed to ensure adequate DNA for subsequent IS*6110*-genotyping. Only one culture from each patient was used for IS*6110*-genotyping and interpretation of the data.

The samples from Zambia were cultured on LJ slants at the Chest Diseases Laboratory (CDL) in Zambia using conventional culture methods. All culture positive isolates confirmed as MTB were sent to the Tropical Disease Research Centre (TDRC) for DNA extraction. A total of 114 isolates had DNA that could be used for genotyping.

IS*6110*-genotyping

DNA was extracted from *M. tuberculosis* cultures from Zimbabwe and Cape Town as previously described (15). IS*6110*–genotyping was performed using the international standard typing method for *M. tuberculosis (*8). Briefly Genomic DNA was digested with the restriction endonuclease *Pvu*II and subjected to electrophoresis in a 0.8% agarose gel. Each lane on the gel included an internal marker (Marker X; Boehringer Mannheim, Germany) to enable normalization between lanes, while each gel included two external marker lanes (MTB 14323) to ensure inter-gel comparisons. The DNA fragments on the gel were Southern-blotted to Hybond N + (Amersham Pharmacia-Biotec) and IS*6110*-containing fragment were visualized after hybridization with Enhanced Chemiluminescence (ECL)-labelled probes IS*6110*-3' and Marker X DNA (Boehringer Mannheim, Germany). Each probe was stripped from the membrane by denaturation

before hybridization with the subsequent probe. Each DNA genotype was normalised and analysed using Gelcompar (Version 4.0, Applied Maths, BVBA, Kortrijk, Belgium) and entered into a database. The IS*6110*-genotype patterns were aligned by use of the GelCompar program with tolerance parameters allowing a 6% shift in each pattern as a whole and a 0.4% variance in individual band positions. Cluster analysis was done using the Unweighted Pair Group Method with Arithmetic mean (UPGMA), based on Dice Coefficient. The isolates were classified into strain families according to the similarity of the IS*6110*-genotype patterns. Strain families were defined according to a similarity index of ≥ 70% (16). Isolates within strain families were classified into clusters of identical strains (strains sharing an identical IS*6110*-genotype pattern) or uniques (strains with a unique IS*6110*-genotype pattern). The IS*6110*-genotype database for Zimbabwe was compared to the IS*6110*-genotype database of isolates from the epidemiological study site in Cape Town (maintained at Stellenbosch University) (15, 17, 18) to identify similarities in IS*6110*-genotype patterns.

IS*6110*-genotyping could not be done for the isolates from Zambia as there was insufficient DNA.

Spoligotyping

Isolates from Zimbabwe, Cape Town and Zambia were spoligotyped according to a standardized protocol (9) to detect DNA polymorphism in the direct repeat (DR) region. Spoligotyping was done at the Stellenbosch University, Department of Medical Biochemistry and the resultant autoradiographs were analysed manually and information

entered into a Microsoft Excel database to allow comparisons between isolates from the different regions.

To determine the global distribution of the spoligotypes identified in this study, the spoligotypes were compared with the spoligotypes deposited in the SpolDB-3.0 worldwide database (19).

RESULTS

IS*6110*-genotype patterns of *M. tuberculosis* strains isolated in Zimbabwe and Cape Town.

Sputum cultures were available for genotyping from 246 different patients resident in Harare (n = 62) and Gweru (n = 184). Among these isolates, 220 (89.4%) harboured more than six IS*6110* copies (high copy number strains), 24 (9.8%) had six or fewer IS*6110* copies (low copy number strains) and the remaining two isolates had poor IS*6110*-genotype patterns due to poor quality DNA. Strains lacking IS*6110* were not identified in either of the study sites.

Cluster analysis based on the UPGMA and Dice coefficient grouped the isolates into a number of defined strain families (according to a similarity index of ›70%). A total of 15 strain families were identified among the Zimbabwean isolates (Figure 1). However 116 of the 246 isolates (47.2%) belonged to a group of strains sharing a similarity of ›80%. These 116 isolates were characterised by 8-17 IS*6110* hybridizing fragments (Figure 2). The majority of these isolates shared six IS*6110*-PvuII fragments of 0.9, 1.67, 2.43, 2.60,

2.7, 4.16 and 4.6 Kb (Figure 2). The IS*6110*-genotyping analysis revealed 73 different patterns among these 116 isolates. Sixty-three isolates belonged to 20 clusters each containing between 2-8 isolates. The remaining 53 IS*6110*-genotype patterns were unique. Smear microscopy data was available for 75 isolates from this strain family and 51 of these (68%) were smear positive whereas 24 (32%) were smear negative.

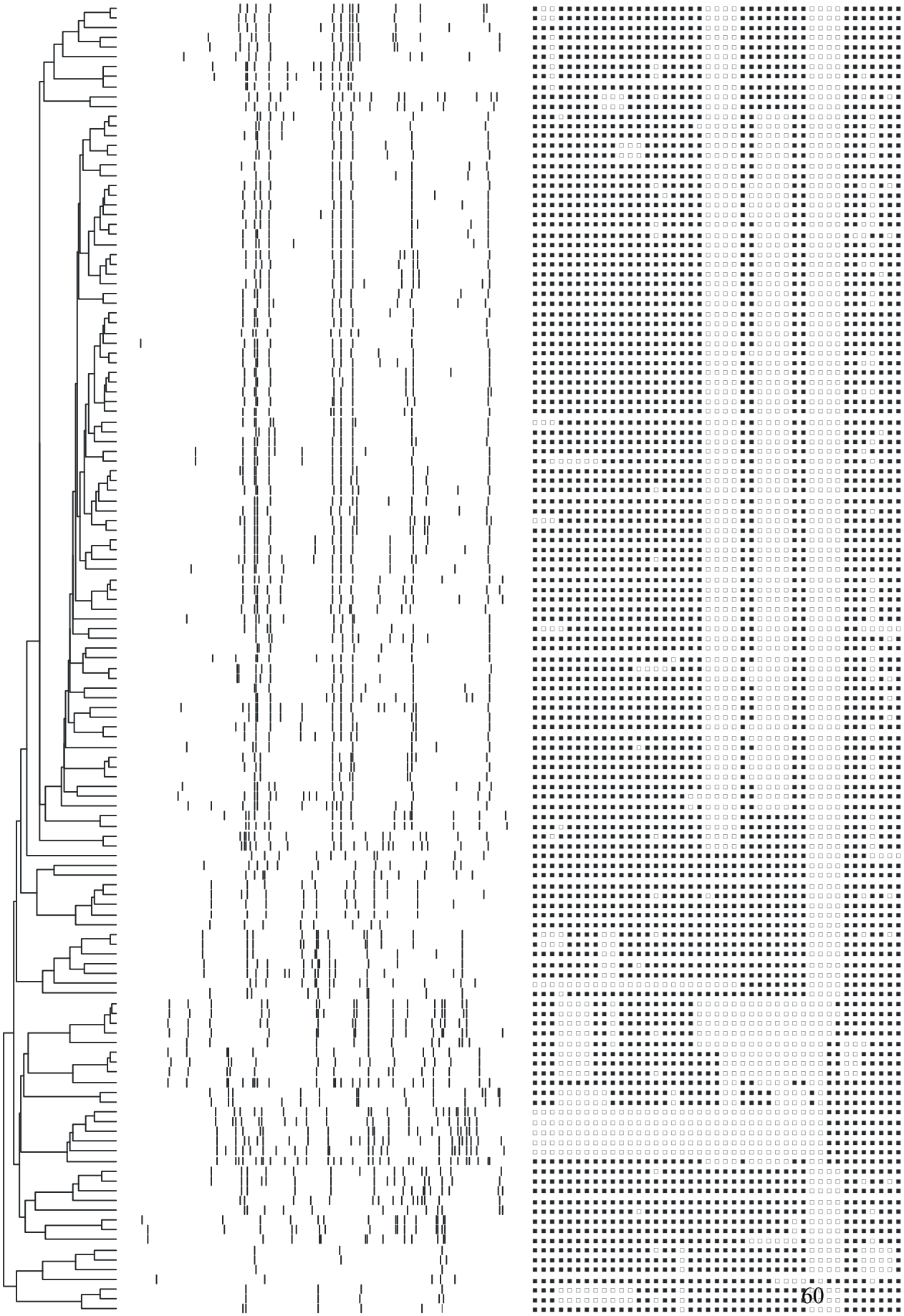


Figure 1: **IS*6110*-genotype and spoligotype patterns of the *M. tuberculosis* isolates from Zimbabwe.** Cluster analysis of *M. tuberculosis* isolates from Harare and Gweru in Zimbabwe. The analysis was based on UPGMA and Dice Coefficient. The resulting dendrogram shows the IS*6110* similarity index on the left and the IS*6110*-genotype banding patterns in the middle. Strain families were assigned according to IS*6110* similarity index of ≥ 70% (16). On the extreme right are the corresponding DNA polymorphisms in the DR region (spoligotype patterns).

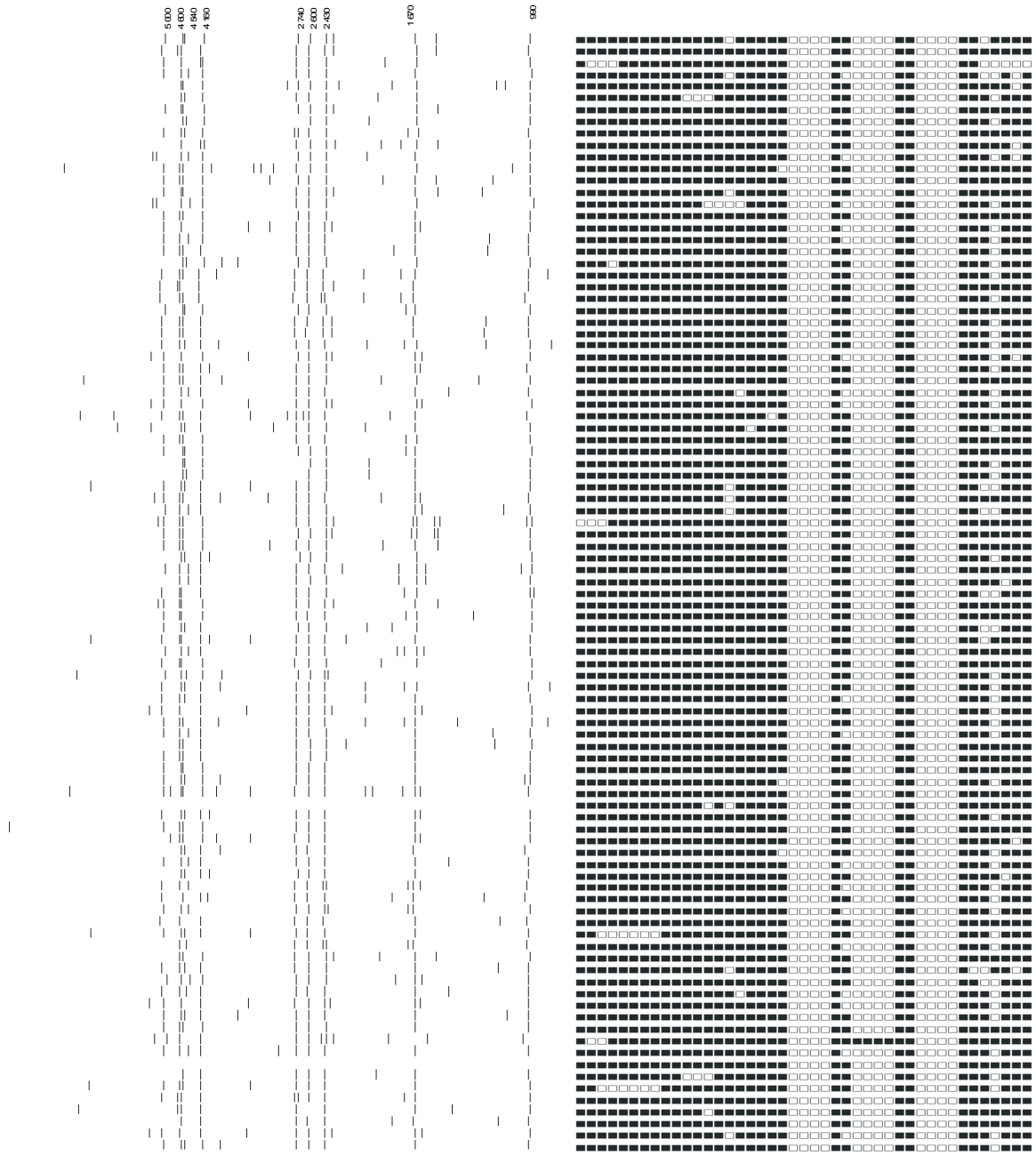


Figure 2: IS*6110*-genotype and spoligotype patterns of the "SAF1" Family isolates from Zimbabwe. IS*6110*-genotype banding patterns of isolates from Zimbabwe showed a similarity of ≥70%. All isolates shared a common banding pattern of six IS*6110*-*PvuII* fragments, 0.9, 1.67, 2.43, 2.60, 2.7, 4.16 and 4.6 Kb, respectively. In addition, a spoligotype "signature" lacking spacers 21-24, 27-30 and 33-36 was observed among the isolates. The two isolates with no IS*6110*-RFLP due to poor DNA also had this characteristic spoligotype pattern.

Analysis of *M. tuberculosis* strains with spoligotyping.

To confirm that the isolates belonged to a closely related group, representing clonal expansion, the 116 isolates were further characterized by analysis of the DR region using spoligotyping. One hundred and fifteen isolates shared a characteristic spoligotype signature lacking spacers 21-24, 27-30 and 33-36(Figure 2). One isolate only lacked spacers 21-24 and 33-36, but had the characteristic IS*6110*-genotype pattern possibly representing a more ancestral strain. A total of 27 distinct spoligotypes were identified among these 116 isolates. Twelve of the isolates each had a unique spoligotype pattern and the remaining 104 (89.7%) belonged to 15 clusters (Figure 3).

Comparison of isolates belonging to the major family grouping with other databases.

Comparison of the IS*6110*-genotype patterns from the Zimbabwean isolates with those from the Cape Town IS*6110*-genotype database showed ›70% similarity to19 Cape Town isolates (representing 2.3% of TB patients in that study setting)(16). Four different IS*6110*-genotype patterns were shared between isolates from Cape Town and from Zimbabwe. Eighteen of the Cape Town isolates had the characteristic spoligotype signature, lacking spacers 21-24, 27-30 and 33-36 and one isolate had the ancestral spoligotype (Figure 3). Three spoligotype patterns (811, 815 and 59) were shared among isolates from Zimbabwe and Cape Town (Figure 3).

We compared the characteristic spoligotype patterns with spoligotype patterns from 114 Zambian isolates. Seventy-two of these isolates (63.2%) showed the characteristic spoligotype pattern; while a further 2 isolates showed the ancestral spoligotype (Figure

3). Restricted strain diversity was observed among the Zambian isolates with a total of 17 spoligotype patterns being identified among the 74 isolates. Of the 17 different patterns, 4 patterns were shared among 61 isolates and 13 were unique (Figure 3). Five spoligotype patterns were shared among isolates from Zimbabwe and Zambia, while 3 spoligotypes were shared between the Zambian and Cape Town isolates. Spoligotype pattern 59, formed the largest clusters in Zambia (44 isolates), Zimbabwe (36 isolates) and Cape Town (6 isolates).

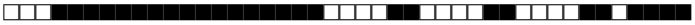
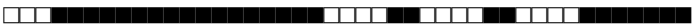
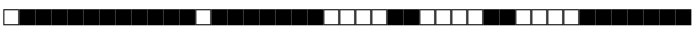
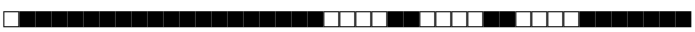
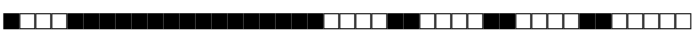
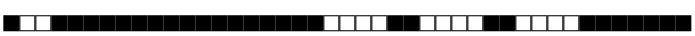
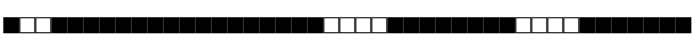
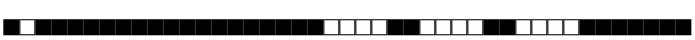
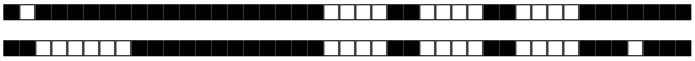
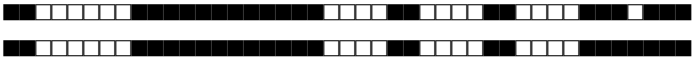
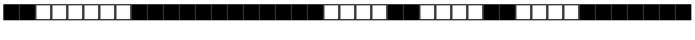
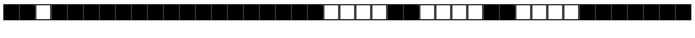
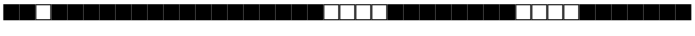
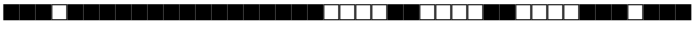
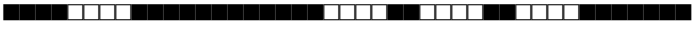
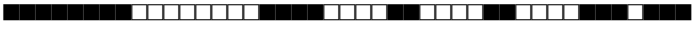
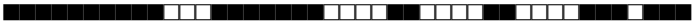
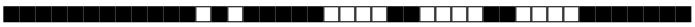
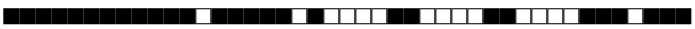
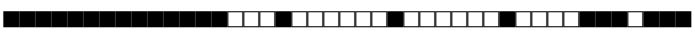
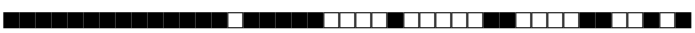
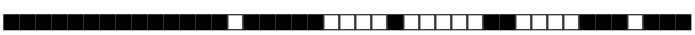
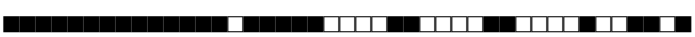
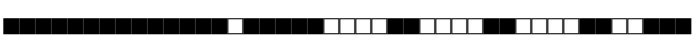
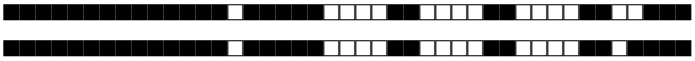
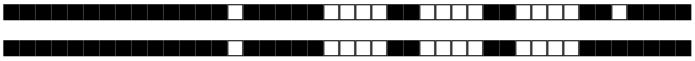
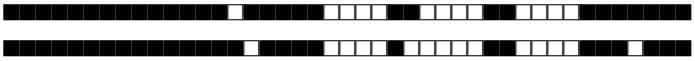
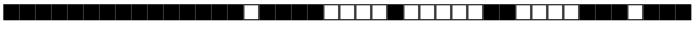
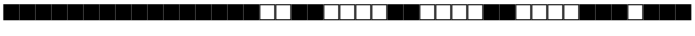
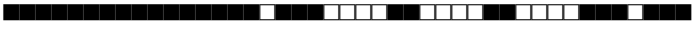
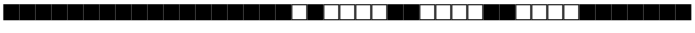
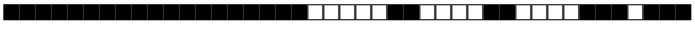
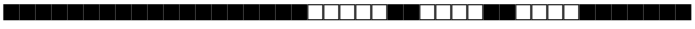
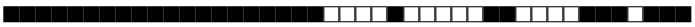
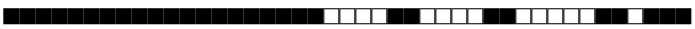
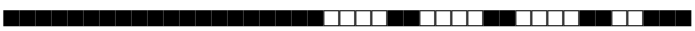
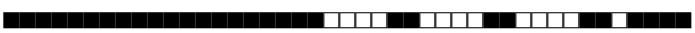
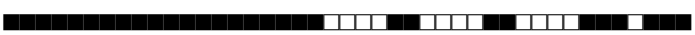
Spoligo pattern	Spoligotype	Zimbabwe (n)	Zambia (n)	Cape Town (n)
	orphan	0	2	0
	816	3	0	0
	orphan	0	1	0
	orphan	0	3	0
	412	1	0	0
	orphan	0	0	1
	753	1	0	0
	orphan	0	0	1
	orphan	2	0	0
	orphan	0	1	0
	orphan	0	1	3
	20	0	1	0
	orphan	1	0	0
	orphan	0	1	0
	orphan	0	1	0
	orphan	2	0	0
	orphan	1	0	0
	orphan	1	1	0
	orphan	0	0	1
	orphan	0	1	0
	orphan	2	0	0
	orphan	1	0	0
	orphan	1	0	0
	orphan	2	0	0
	orphan	1	0	0
	orphan	4	0	0
	84	1	0	0
	orphan	0	0	3
	orphan	2	0	0
	orphan	1	0	0
	orphan	5	0	0
	orphan	1	0	0
	orphan	2	0	0
	811	14	0	2
	orphan	0	1	0
	813	3	1	0
	814	1	1	0
	815	20	12	1
	orphan	2	1	0
	orphan	5	0	0
	59	36	44	6
	42	0	1	1
Total	42	116	74	19

Figure 3: Schematic representation of spoligotype patterns obtained from Cape Town Zambia and Zimbabwe. The spoligotype nomenclature is according to the SpolDB3.0 database (19). Orphan patterns

did not have matches in the SpolDB3.0 database. Spoligotypes 811, 813, 814 and 815 have been previously reported only from Zimbabwe (12;19).

Dissemination of *M. tuberculosis* isolates of the identified major strain family grouping in other countries.

The updated SpolDB3.0 database (19) was searched for isolates with the identical spoligotype signature as those found in Zimbabwe, Zambia and Cape Town. Strains with similar spoligotype signatures have only been rarely isolated from other regions such as the Americas, Europe and Madagascar (19). With the exception of the isolates with spoligotype pattern 20, 42, 59, 412 and 753 this family of strains appears to be predominant in central regions of Southern Africa (Zimbabwe and Zambia) but not in Cape Town. Based on these findings we have now called this strain family Southern Africa Family 1"SAF1".

Discussion

This study represents the first genetic study on *M. tuberculosis*, using IS*6110*-genotyping in combination with spoligotyping in Zimbabwe. This investigation has led to the identification of a predominant group of strains infecting approximately half of the patients in the study, and thereby are major contributors to the TB epidemic in this region. The strains are characterised by a distinct IS*6110*-genotype pattern and a characteristic spoligotype signature suggesting clonal expansion from a common progenitor. This characteristic spoligotype signature was found not only among isolates from Zimbabwe but also predominantly from isolates from Zambia and in a minority of isolates from Cape Town, South Africa. We have named this strain family "SAF1". Only

rarely were "SAF1" spoligotypes identified in other regions of the world (19). Previous comparative genomic studies have shown that the "SAF1" genotype is a member of the LAM family20, which includes F11, F13, F14, F15 and F26 from Cape Town (21). IS*6110* insertion site mapping suggests that these strain families form part of a group of strains that have originated from a common progenitor (21). This would therefore imply that the SAF1 family is a branch within this super family of genotypes, where different branches appear to have different frequencies in different settings. The strain family F11 occurs as the highest frequency strain family in Cape Town, South Africa, being isolated in 21.4% of the patients (22).

We acknowledge that this study has certain limitations. Firstly, the isolates were collected from different settings and therefore the epidemiological relationships according to clustering were not appropriate to define transmission event. Our data set could only be applied to demonstrate the broad distribution of closely related genotypes. To gain further insight into the disease dynamics in Southern African future studies will need to comprehensively analyse isolates of *M. tuberculosis* over an extended period. Such studies must be able to establish whether the *M. tuberculosis* population structure is influenced by HIV co-infection or whether there is an association between clinical presentation and strain genotype. Secondly, the restricted time interval over which the isolates were collected has limited the application of this study to only determine the population structure of *M. tuberculosis.* However, we do not believe that this has introduced a significant bias given that specimens were cultured from patients from diverse geographical settings. Lastly, it is possible that the *M. tuberculosis* strain

population structure may differ between smear positive and smear negative cases. The isolation of Beijing strains in very low proportions may be because these strains have been recently introduction in this population.

Despite these limitations, our results provide evidence that the burden of disease of tuberculosis in the central Southern African regions is largely due to a relatively small subset of actively circulating strains. The success of particular predominant clones related to high incidence of TB was also reported in earlier studies (18, 23-25). The identification of a predominant strain genotype in central parts of Southern Africa may reflect that this strain has either gained a higher degree of virulence, or that the strains in these high incidence countries may have originated from a few clones that expanded rapidly in the recent past. The reasons for the selection and dissemination of these strains are unknown. A recent study on the global genetic population structure for *M. tuberculosis* has suggested that the organism has adapted to specific human populations in different geographical settings (26). As postulated for other predominant strain families, the success of these strains may be a result of increased transmissibility or from expression of an as yet unidentified virulence factor (27). The reason for high frequency of Beijing strains has been related to virulence and mutator phenotypes and this may explain a better adaptability to stress conditions and hostile intracellular conditions (28).

This study may have important implications for the interpretation of molecular epidemiology data. In molecular epidemiological analysis the predominance of closely related strains implies restricted genetic diversity, which may be interpreted as transmission. This may hinder the quantification of the contribution of reactivation and

recent transmission to the epidemic. Furthermore, it may be difficult to define the mechanism leading to recurrent tuberculosis. However, the interpretation of genetic diversity will generally depend on the method of genotyping used, underscoring the need to develop molecular epidemiological tools appropriate to settings with low *M. tuberculosis* IS*6110* genotypic diversity.

Acknowledgements

The authors would like to thank the Wellcome Trust (Research Leave Fellowship Grant 01680/Z/00/Z to Violet Chihota), the South African National Research Foundation (GUN 2054278 to Tommie Victor) and the IAEA (RAF 6025 grant to Tommie Victor). We would also like to thank The Biomedical Research and Training Institute (Harare, Zimbabwe) for the laboratory culture of specimens.

REFERENCES

1. **Dye C, Scheele S, Dolin P et al**. Consensus statement. Global burden of tuberculosis: estimated incidence, prevalence, and mortality by country. WHO Global Surveillance and Monitoring Project. *JAMA* 1999; 282(7):677-686.

2. **Corbett EL, Watt CJ, Walker N et al**. The growing burden of tuberculosis: global trends and interactions with the HIV epidemic. *Arch Intern Med* 2003; 163(9):1009-1021.

3. **Elliott AM, Namaambo K, Allen BW et al**. Negative sputum smear results in HIV-positive patients with pulmonary tuberculosis in Lusaka, Zambia. *Tuber Lung Dis* 1993; 74(3):191-194.

4. **Johnson JL, Vjecha MJ, Okwera A et al.** Impact of human immunodeficiency virus type-1 infection on the initial bacteriologic and radiographic manifestations of pulmonary tuberculosis in Uganda. Makerere University-Case Western Reserve University Research Collaboration. I*nt J Tuberc Lung Dis* 1998; 2(5):397-404.

5. **Alland D, Kalkut GE, Moss AR et al**. Transmission of tuberculosis in New York City. An analysis by DNA fingerprinting and conventional epidemiologic methods. *N Engl J Med* 1994; 330(24):1710-1716.

6. **Godfrey-Faussett P, Sonnenberg P, Shearer SC et al**. Tuberculosis control and molecular epidemiology in a South African gold-mining community. *Lancet* 2000; 356(9235):1066-1071.

7. **Small PM, Hopewell PC, Singh SP et al**. The epidemiology of tuberculosis in San Francisco. A population-based study using conventional and molecular methods. *N Engl J Med 1994*; 330(24):1703-1709.

8. **van Embden JD, Cave MD, Crawford JT et al**. Strain identification of *Mycobacterium tuberculosis* by DNA fingerprinting: recommendations for a standardized methodology. *J Clin Microbiol* 1993; 31(2):406-409.

9. **Kamerbeek J, Schouls L, Kolk A et al.** Simultaneous detection and strain differentiation of *Mycobacterium tuberculosis* for diagnosis and epidemiology. *J Clin Microbiol* 1997; 35(4):907-914.

10. **Supply P, Mazars E, Lesjean S et al**. Variable human minisatellite-like regions in the *Mycobacterium tuberculosis* genome. *Mol Microbiol* 2000; 36(3):762-771.

11. **Mazars E, Lesjean S, Banuls AL et al**. High-resolution minisatellite-based typing as a portable approach to global analysis of *Mycobacterium tuberculosis* molecular epidemiology. *Proc Natl Acad Sci* U S A 2001; 98(4):1901-1906.

12. **Easterbrook PJ, Gibson A, Murad S et al**. High rates of clustering of strains causing tuberculosis in Harare, Zimbabwe: a molecular epidemiological study. *J Clin Microbiol* 2004; 42(10):4536-4544.

13. **Heyderman RS, Goyal M, Roberts P et al**. Pulmonary tuberculosis in Harare, Zimbabwe: analysis by spoligotyping. *Thorax* 1998; 53(5):346-350.

14. **Apers L, Mutsvangwa J, Magwenzi J et al.** A comparison of direct microscopy, the concentration method and the Mycobacteria Growth Indicator Tube for the examination of sputum for acid-fast bacilli. I*nt J Tuberc Lung Dis* 2003; 7(4):376-381.

15. **Warren R, Hauman J, Beyers N et al.** Unexpectedly high strain diversity of *Mycobacterium tuberculosis* in a high-incidence community. *S Afr Med J* 1996; 86(1):45-49.

16. **Richardson M, van Lill SW, van der Spuy GD et al.** Historic and recent events contribute to the disease dynamics of Beijing-like *Mycobacterium tuberculosis* isolates in a high incidence region. *Int J Tuberc Lung Dis* 2002; 6(11):1001-1011.

17. **Verver S, Warren RM, Munch Z et al**. Transmission of tuberculosis in a high incidence urban community in South Africa. *Int J Epidemiol* 2004; 33(2):351-357.

18. **Warren R, Richardson M, van der SG et al**. DNA fingerprinting and molecular epidemiology of tuberculosis: use and interpretation in an epidemic setting. *Electrophoresis* 1999; 20(8):1807-1812.

19. **Filliol I, Driscoll JR, van Soolingen D et al**. Snapshot of moving and expanding clones of *Mycobacterium tuberculosis* and their global distribution assessed by spoligotyping in an international study. *J Clin Microbiol* 2003; 41(5):1963-1970.

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

21. **Warren RM, Sampson SL, Richardson M et al.** Mapping of IS*6110* flanking regions in clinical isolates of *M. tuberculosis* demonstrates genome plasticity. *Mol Microbiol* 2000; 37(6):1405-1416.

22. **Victor TC, de Haas PE, Jordaan AM et al**. Molecular characteristics and global spread of *Mycobacterium tuberculosis* with a western cape F11 genotype. *J Clin Microbiol* 2004; 42(2):769-772.

23. **Haas WH, Engelmann G, Amthor B et al**. Transmission dynamics of tuberculosis in a high-incidence country: prospective analysis by PCR DNA fingerprinting. *J Clin Microbiol* 1999; 37(12):3975-3979.

24. **Hermans PW, Messadi F, Guebrexabher H et al**. 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(6):1504-1513.

25. **Niobe-Eyangoh SN, Kuaban C, Sorlin P et al**. Genetic biodiversity of *Mycobacterium tuberculosis* complex strains from patients with pulmonary tuberculosis in Cameroon. *J Clin Microbiol* 2003; 41(6):2547-2553.

26. **Gagneux S, DeRiemer K, Van T et al**. Variable host-pathogen compatibility in *Mycobacterium tuberculosis*. *Proc Natl Acad Sci* U S A 2006; 103:2869-2873.

27. **Bifani PJ, Mathema B, Liu Z et al**. Identification of a W variant outbreak of *Mycobacterium tuberculosi*s via population-based molecular epidemiology. *JAMA* 1999; 282(24):2321-2327.

28. **Rad ME, Bifani P, Martin C et al.** Mutations in putative mutator genes of *Mycobacterium tuberculosis* strains of the W-Beijing family. *Emerg Infect Dis* 2003; 9(7):838-845.

Geographical distribution of *Mycobacterium tuberculosis* strains and pulmonary tuberculosis cases notified during one year in Gweru, Zimbabwe[3]

Apers L, Chimusoro A, **Chihota V.**

[3] Part of these findings were published in *The Central African Journal of Medicine* 2005; 51: 62-64. Apers L, Chimusoro A, Chihota V. Geographical distribution of pulmonary tuberculosis cases notified during one year in Gweru, Zimbabwe.

SUMMARY

We analyzed the geographical distribution of adult pulmonary tuberculosis (PTB) cases that occurred in a well defined area in Gweru, Zimbabwe during the period September 2000 to September 2001. Using data obtained though routine registration of TB patients the number of TB cases was linked to the population density of the respective districts where the cases originated. A subset of *M. tuberculosis* genotypes was plotted on a map using a geographical information system (GIS) to determine the distribution of the SAF1/RD^{Rio} strains genotypes in this city. The SAF1/RD^{Rio} strains were spread in the districts with high incidence of TB only, further supporting the earlier findings that these strains may be highly transmissible and that they play an important role in the TB epidemic in Zimbabwe.

Introduction

Tuberculosis remains an important cause of premature death and efforts to control TB are undermined by the impact of HIV-associated TB. Among new adult TB cases in Zimbabwe, 67% are estimated to be co-infected with HIV (1). A steep rise in notification rates was observed from 1997 in cities in Zimbabwe suggesting that transmission was playing a major role in the epidemic.

Molecular typing, particularly IS*6110*-restriction fragment length polymorphism and spoligotyping, have allowed strain differentiation of *Mycobacterium tuberculosis* and this has helped in understanding strain diversity, transmission patterns and distribution of strains in specific areas to be studied (2, 3, 4). Ever since Lord Snow discovered the association between water supply and cholera, epidemiologists have used maps when analyzing associations between location, environment and disease (5). Since the 1970s Geographic information systems (GIS) have added a powerful tool for the analysis of these associations because of its spatial analysis and display capabilities (6). A recent application of this is the mapping of *M. tuberculosis* DNA fingerprints. This has led to a better understanding of the transmission dynamics of tuberculosis (2, 3, 7, 8).

Gweru is an industrial mining town with 141 210 inhabitants (2002 population census) and is situated on the central plateau of Zimbabwe and is in the Midlands province. The Gweru provincial hospital is the referral centre for the population of the Midlands province but also acts as a district hospital for Gweru District. The population for the district was 303 938 in 2001 as extrapolated from the 2002 census and that of Gweru City

is 141 210. The TB notification rate for the province was 412 per 100 000 population in 2001, 66.7 % were pulmonary TB (PTB), and of these 51.5 % had a positive smear (9). We set out to determine the geographic distribution of PTB cases notified over a period of one year in Gweru and to describe the distribution of *M. tuberculosis* genotypes, which were later described in a another study. (10).

Methodology

Setting

Patients were recruited at Gweru Provincial Hospital, the referral centre for the population of Midlands Province. Gweru City has a business centre, low density (district 1-3 and 5) and high density living areas (district 4, 6-9) and an extensive heavy industrial area (Figure 1). Besides the government provincial hospital, health services are provided in 11 health facilities (clinics) governed by the health department of the city council, and also in one private hospital. Patients suspected of TB in private facilities and clinics are referred to, and eventually diagnosed and registered in the provincial hospital.

Participants

Consecutive patients diagnosed with PTB and started on treatment at Gweru Provincial Hospital from September 2000 until September 2001. For this analysis only patients that were residing for at least one year in Gweru City, were included. Demographic data (age, gender) and residential addressed were abstracted for the government TB registers.

***M. tuberculosis* isolates**

For the genotypic analyses a convenient subset of 90 *M. tuberculosis* isolates with genotyping data were selected.

Geographical database

All available geographical data on communal services, public places, roads and rivers, open and industrial areas in Gweru were imported in a geographical information system (GIS) using ArcView 3.2 (Environmental Systems Research Institute, Redlands, USA). Coordinates of residences of TB-cases were measured using a hand held GPS-receiver or were introduced in the GIS on the basis of the patient's address.

To compare the spatial distribution of SAF1/RD^{Rio} strains, their distribution was compared to that of the other genotypes. TB cases in Gweru, the TB cases identified in the different districts of the city were also plotted to show the distribution of cases. This was based on the residential address of the patient as indicated in the TB notification register.

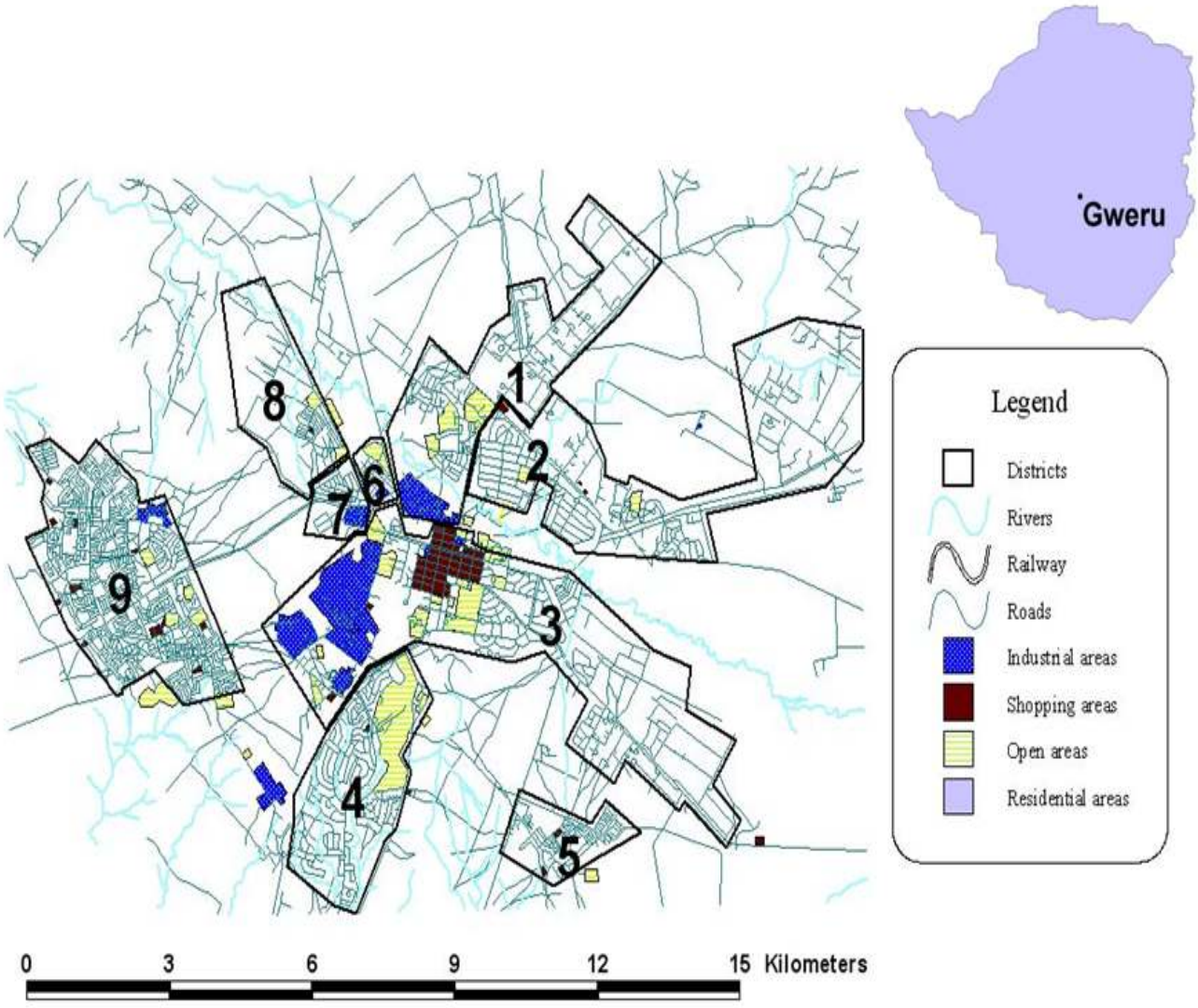


Fig 1: Geographical map showing the city of Gweru showing the nine residential areas in the city (also referred to as districts).

Ethical considerations

This study was approved by the Medical Research Council of Zimbabwe (MRCZ/A/895, MRCZ/A/967), the Provincial Medical Director of the Midlands Province, the Medical Superintendent of Gweru Provincial Hospital and the Stellenbosch University Faculty of Health Sciences (2003/022/N).

Results

***M. tuberculosis* genotypes identified from a subset of isolates**

The 90 *M. tuberculosis* isolates were grouped into 15 IS*6110*-RFLP families based on classification by Richardson and colleagues (11). Of the 90 isolates 35 (38.8%) belonged to the SAF1/RD^{Rio} strains and shared a characteristic IS*6110*-RFLP pattern and spoligotype signature as described previously (11). The remaining 55 isolates belonged to 14 genotypes including other Latin American Mediterranean (LAM;n=18; 20.0%), X (n=7; 7.8%), unknown types (n=7; 7.8%), S (n=6; 6.7%), Beijing (n=5; 5.6%), Central Asian (CAS;n=5; 5.5%), T (n=4; 4.4%), and Haarlem (H) genotypes (n=3; 3.3%).

Geographical distribution of PTB cases

Figure 2 depicts the geographical distribution of the 204 PTB patients that were notified from September 2000-September 2001. There was a concentration of cases in district 6 and 9 and these are densely populated townships at the western and north-western side of

the heavy industrial area. Incidence rates for the districts are shown in table 1 below.

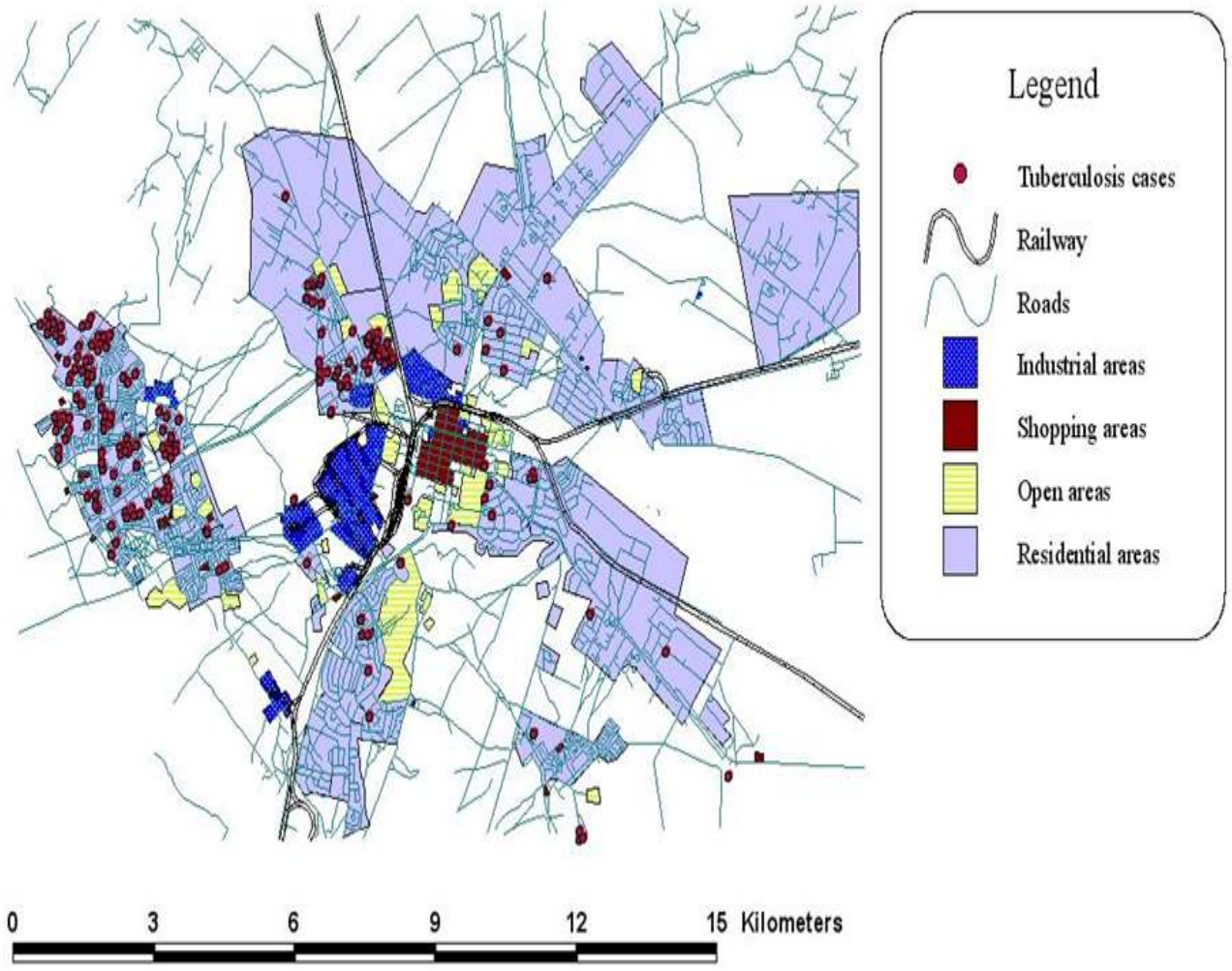


Figure 2: Geographical distribution of PTB cases identified in Gweru City over a period of one year. Each red dot represents one TB case

Table 1: The pulmonary TB incidence in each of the nine districts in Gweru during September 2000 –September 2001

District	Population	No of new pulmonary TB cases	Population density Inhabitants/km²	Incidence (n/1000)
1	8230	2	805	0.24
2	11357	11	914	0.97
3	6547	16	487	2.44
4	8605	9	1629	1.04
5	6525	9	4083	1.38
6	7357	19	15792	2.58
7	6992	16	5614	2.29
8	6813	10	1554	1.47
9	78784	112	8311	1.42
Total	141210	204		1.44

Geographical distribution of *M. tuberculosis* strains

Figure 3 shows the distribution of *M. tuberculosis* strains in Gweru City. As observed with the number of cases of PTB (Fig 2), the SAF1/RD^{Rio} strains were identified mainly in district 6, 7, 8 and 9. These districts also had a high number of TB cases notified. The SAF1/RD^{Rio} strains were predominantly isolated in district 9 (23/35; 65.7%) and were widely spread within this district. Though districts 6, 7 and 8 had high number of PTB cases the SAF1/RD^{Rio} strains were identified in lower proportions.

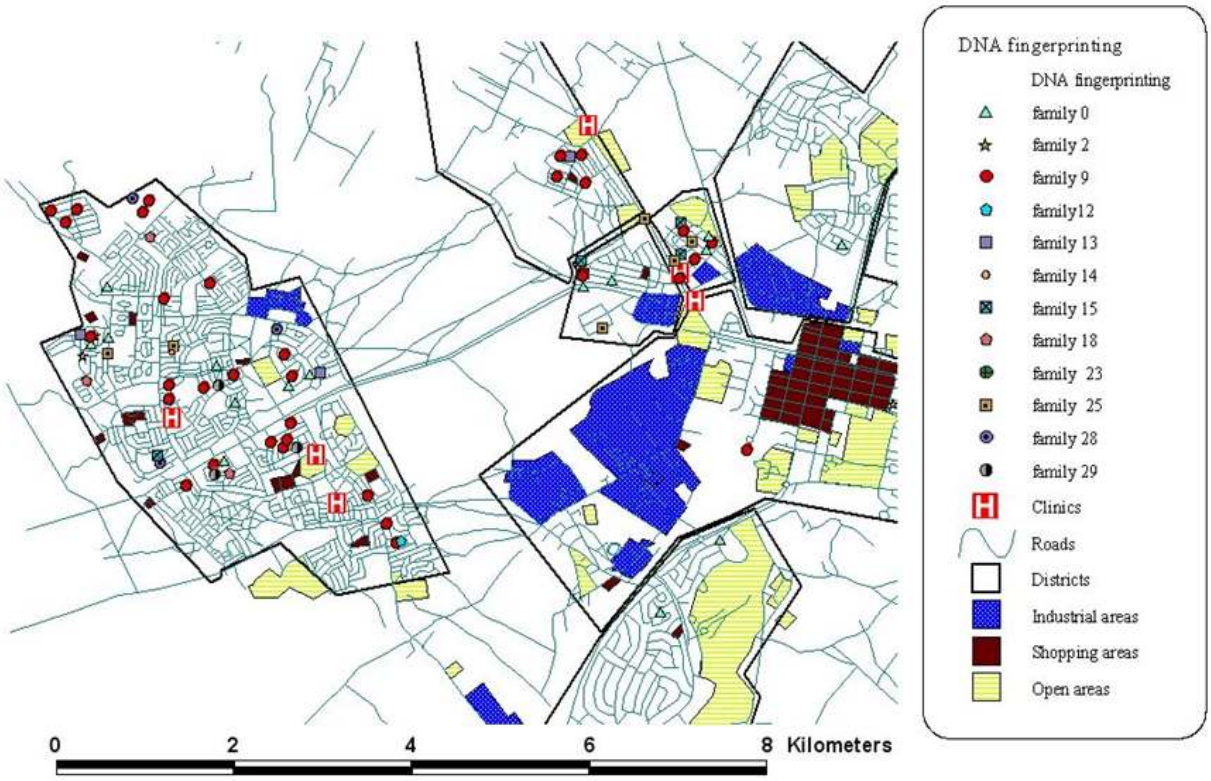


Figure 3: Geographical distribution of *Mycobacterium tuberculosis* genotypes (IS*6110*-DNA fingerprint families assigned according to Richardson and colleagues [11]). Different coloured shapes represent different strain families and IS*6110* DNA fingerprint family 9= SAF1/RD^{Rio}. The abbreviation H in red represents clinics.

Discussion

TB is a disease that is closely associated with poverty and transmission occurs in settings where infected persons are in close contact with others. Using data collected over a year we looked at the distribution of TB cases in Gweru city and we then went further to look at the distribution of *M. tuberculosis* genotype. By plotting the distribution of *M. tuberculosis* genotypes in this high incidence setting, SAF1/RD^{Rio} strains were found to be spread out in high incidence districts in Gweru. This findings confirms those of an earlier study (10), showing that the SAF1/RD^{Rio} strains play an important role in the TB epidemic in Zimbabwe.

The study had several limitations. Gweru residents that were diagnosed outside the City, and were not transferred to the provincial hospital for further follow-up, were missed and the numbers were unknown. Therefore the number of TB cases included in this analysis may be lower than expected and we may underestimate the number of circulating strains. A majority of cases were from the high population density suburbs (district 6 and 9) but there were a substantial number of cases from the low population density districts (district 1-3 and 5). Not all genotypes from 184 isolates analysed as part of the previous study (chapter 2) were included in this analysis but only a subset with available genotyping data at the time of this analysis. Therefore the proportions of strains may have been under-represented.

The findings of the RFLP analyses showed low strain diversity and SAF1/RD^{Rio} predominated. The SAF1/RD^{Rio} strains were more likely to be identified in high TB

incidence districts such as district 6, 8 and 9 and because of the small sample size this could have occurred by chance. Transmission as defined by identical *M. tuberculosis* strains, in this case SAF1/RD^{Rio} strains seemed to occur more commonly in these high TB incidence settings. However none were identified in district 3, another high TB incidence setting, and this could possibly be explained by the low number if isolates included in this analysis. Earlier studies in Harare also found a trend in association towards clustering and certain districts in Harare and that the incidence of TB was influenced by a small subset of actively circulating strains (12). Similarly, in high incidence settings in Cape Town, South Africa, some extremely large clusters could be identified that represented a disproportionate number of TB cases (13). It therefore seems reasonable that the SAF1/RD^{Rio} isolates identified in this study would be spread in high incidence settings which may imply that they play an important role in driving the TB epidemic.

In conclusion, the distribution of SAF1/RD^{Rio} strains in Gweru is widespread and the majority of cases were found in a densely populated township. Though conclusions cannot be drawn from this cross sectional study, there was no evidence of transmission within households. Further studies will need to be done to confirm or refute these initial observations. The *M. tuberculosis* SAF1/RD^{Rio} strains were not clustered within the districts but were widespread in the city, implying that they play an important role in the TB epidemic in Gweru.

References

1. **Corbett EL, Watt CJ, Walker N et al**. The growing burden of tuberculosis: global trends and interactions with the HIV epidemic. *Arch Intern Med* 2003; 163(9):1009-1021.

2. **Hayward AC, Goss S, Drobniewski F, Saunders N, Shaw RJ, Goyal M, Swan A, Uttely A, Pozniak, Grace-Parekr J and Watson JM**. The molecular epidemiology of tuberculosis in inner London. *Epidemiol Infect* 2002; 128:175-184

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

4. **Verver S, Warren R, M Munch Z, Richardson M, van der Spuy GD, Enarson DA, Borgdorff W, Behr MA and Beyers *N.*** Transmission of tuberculosis in a high incidence urban community in South Africa*.* *Int J Epidemiol* 2004; 33: 351-357.

5. **Gesler, W**. The uses of spatial analysis in medical geography: a review. *Soc.Sci.Med* 1991; *198:* 963-973.

6. **Clarke KC, McLafferty SL, Tempalski BJ.** On epidemiology and geographic information systems: a review and discussion of future direction. *Emerg Infect Dis* 1996; 2: 85-92.

7. **Barnes, P. F., Z. Yang, S. Preston-Martin, J. M. Pogoda, B. E. Jones, M. Otaya, K. D.Eisenach, L. Knowles, S. Harvey, and M. D. Cave.** Patterns of tuberculosis transmission in Central Los Angeles. *JAMA* 1997; 278: 1159-1163.

8. **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**. The epidemiology of tuberculosis in San Francisco. A population-based study using conventional and molecular methods. *N Engl.J Med* 1994; 330:1703-1709.

9. **Apers, L., J. Mutsvangwa, J. Magwenzi, N. Chigara, A. Butterworth, P. Mason, and S. P. Van der Stuft**. A comparison of direct microscopy, the concentration method and the Mycobacteria Growth Indicator Tube for the examination of sputum for acid-fast bacilli. *Int J Tuberc.Lung Dis* 2003; 7:376-381

10. **Chihota, V., L. Apers, S. Mungofa, W. Kasongo, I. M. Nyoni, R. Tembwe, G. Mbulo, M. Tembo, E. M. Streicher, G. D. van der Spuy, T. C. Victor, P. van Helden, and R. M. Warren**. Predominance of a single genotype of *Mycobacterium tuberculosis* in regions of Southern Africa. *Int J Tuberc.Lung Dis* 2007; 11:311-318.

11. **Richardson M, van Lill SW, van der Spuy GD et al**. Historic and recent events contribute to the disease dynamics of Beijing-like *Mycobacterium tuberculosis* isolates in a high incidence region. *Int J Tuberc Lung Dis* 2002; 6 :1001-1011.

12. **Easterbrook PJ, Gibson A, Murad S, Lamprecht D, Ives N, Ferguson A, Lowe O, Mason P, Ndudzo A, Taziwa A, Makombe R, Mbengeranwa L, Sola C, Rastogi N, Drobniewski F**. High rates of clustering of strains causing tuberculosis in Harare, Zimbabwe: a Molecular epidemiological study. J Clin Microbiol 2004; 42: 4536-4544.

13. **Warren R, Richardson M, van der Spuy G, Victor T, Sampson S, Beyers N, van Helden PD**. DNA fingerprinting and molecular epidemiology of tuberculosis: use and interpretation in an epidemic setting. *Electrophoresis* 1999; 20(8):1807-1812.

Tuberculosis due to *Mycobacterium tuberculosis* "SAF1/RD^{Rio}" strains is associated with smoking, but not with cavitary lung disease in an HIV co-infected population[4]

Violet Chihota, Shungu S. Munyati, Gian D. van der Spuy, Tsitsi Bandason, Beauty Makamure, Paul D. van Helden, Robin M. Warren, Elizabeth L. Corbett

[4] Manuscript in preparation and to be submitted to a peer reviewed journal

Abstract

Background

Whether the clinical spectrum of disease is dependent on *Mycobacterium tuberculosis* strains and whether specific factors predispose to infection with specific strains is unclear.

Objectives

To characterize *M. tuberculosis* strains from tuberculosis (TB) patients in Harare, Zimbabwe and to analyze the relationship between genotype, demographic characteristics and clinical consequences.

Methods

M. tuberculosis isolates from patients presenting with chronic cough, were genotyped using IS*6110*-restriction fragment length polymorphism (RFLP) and spacer-oligonucleotide typing (spoligotyping). Demographic and clinical information collected at enrolment, were used to identify potential risk factors for disease and clinical consequences associated with disease due to the predominant genotype.

Results

Southern Africa Family 1/RD$^{\mathbf{Rio}}$ (SAF1/RD$^{\mathbf{Rio}}$) strains were identified from 69/134 (51.5%) isolates. Among HIV co-infected TB patients smoking was associated with disease due to SAF1/RD$^{\mathbf{Rio}}$ strains (adjusted OR (aOR) 3.50, 95% CI 1.20-10.20; p=0.02), but not with cavitary disease. However considering all TB cases, regardless of HIV status, smoking was associated with disease due to SAF1/RD$^{\mathrm{Rio}}$(aOR 3.33, 95% CI 1.25-8.83; p=0.02) and so was cavitary disease (aOR 2.11, 95% CI 1.00-4.51; p=0.05).

Though not significant, HIV negative patients were more likely to present with cavitary disease.

Conclusion

The SAF1/RD^{Rio} *M. tuberculosis* strains predominate in this population in Harare, Zimbabwe. Smoking may increase the chances of getting disease due to SAF1/RD^{Rio}. The observation that cavitary disease was common among HIV negative TB patients presenting with disease due to SAF1/RD^{Rio}, suggests a more severe disease that is easily transmissible in this population. Early diagnosis and treatment are critical in curbing transmission of TB in this community.

Background

Tuberculosis (TB) an infectious disease caused by *Mycobacterium tuberculosis* remains one of the world's leading cause of mortality. The World Health Organization estimates that one-third of the world's population is currently infected with *M. tuberculosis* of which 5-10% will become sick or infectious at sometime in their life (1). The risk of developing disease increases to 10% per year in those infected with both *M. tuberculosis* and HIV (1).

The chances of exposure to infection with *M. tuberculosis* is largely dependent on prevailing risk factors which may include the number of incident infectious cases in a community, duration of infectiousness, nature of interactions between a case and a susceptible contact, age and gender (2). The HIV pandemic is one of the most important challenges facing TB control. Immune suppression increases the risk of reactivation of latent TB infection and rapid progression to active TB disease and TB also speeds progression of HIV/AIDS (3, 4, 5). The incidence of TB remains high in Zimbabwe increasing from 604 to 782 per 100 000 population in 2000 and 2007 respectively (6, 7).

Molecular typing based on genetic markers allows the rapid identification of mycobacteria within the *M. tuberculosis* complex as well as providing tools for examining the transmission and evolution of these microorganisms. DNA fingerprinting has enabled the accurate classification of *M. tuberculosis* strains allowing them to be classified according to genotypic similarity (8). Isolates can be classified into identical clones (strains) and also into anciently related major global genotypes (lineages/clades)

and each of these classifications address different questions thereby providing insights into TB disease dynamics in host populations (9, 10, 11, 12). Efforts to characterize the strains present within populations have led to an increasing understanding of their transmission and global distribution. These studies have shown that in certain areas a small number of strains cause a disproportionate number of TB cases (13, 14, 15, 16). This may be a result of some strains being more inherently transmissible than others either because; firstly, they cause sputum smear positive disease that is more likely to be transmitted; secondly, infection with these strains results in delayed onset of symptoms in patients, making them more infectious for longer; or thirdly, these strains are more virulent than others.

The Beijing genotype, the best characterized strain family has been confirmed as an emerging pathogen in several areas and a predominant endemic strain in others (13); it is frequently associated with young age suggesting an increased tendency to spread in East Asian countries (17,18,19) and also with MDR in Estonia, United States and Russia(20,21,22). Studies in Cape Town suggest that the rapid emergence of a predominant Beijing genotype demonstrated adaptations to conditions within the study community and is likely to reflect enhanced pathogenicity rather than transmissibility (12). Recently Gagneaux *et al* reported on the phylogenetic structure of *M. tuberculosis* genotypes and provided data that supported the hypothesis that certain genotypes are preferentially adapted to particular human populations (23).

In an earlier study we described a predominant group of strains, the Southern Africa Family 1 (SAF1/RD$^{\mathbf{Rio}}$), identified in 47.2% of the *M. tuberculosis* isolates from TB patients in Harare and Gweru in Zimbabwe (15). The SAF1/RD$^{\mathbf{Rio}}$ isolates were characterized by 8-17 IS*6110* hybridizing fragments and the majority of the isolates shared six IS*6110*-*Pvu*II fragments of 0.9; 1.67; 1.43; 2.60; 2.7; 4.16 and 4.6 Kb. Spoligotype analysis revealed a characteristic spoligotype signature lacking spacers 21-24; 27-30 and 33-36. By comparing the spoligotype patterns with what is available in the SpolDB4 database it was established that the SAF1/RD$^{\mathbf{Rio}}$ strains belonged to the Latin American-Mediterranean (LAM) genotype and were found to fall into the South African F9 family and also similar to the RD$^{\mathrm{Rio}}$ which have been found to be the a significant cause of TB worldwide (24, 25).

The LAM genotype is the most prevalent *M. tuberculosis* genotype globally accounting for approximately 15% of the global TB burden. These strains were implicated in earlier studies in Zimbabwe as causing high rates of clustering (26,27). *M. tuberculosis* strains of the SAF1/RD$^{\mathrm{Rio}}$ genotype belong to the LAM family which is found within the principal genetic group 2 (PGG2), in phylogenetic cluster VI (28, 29), have a deletion in the region of difference (RD) TbD1 and a genotype specific large sequence polymorphism (LSP) (23).

This genotype has been associated with disease outbreaks and MDR-TB including a recent description of the LAM4 subfamily, a leading cause of extreme drug resistant (XDR) TB in KwaZulu Natal in South Africa (30). SAF1/RD$^{\mathrm{Rio}}$ strains have been

reported in similar proportions in Rio de Janeiro and Belo Horizonte in Brazil contributing 30% and 37% of total TB burden, respectively (16, 31). Despite its large contribution to TB, much less is known about clinical and demographic factors associated with infection with the LAM family strains.. Associations were found with haemoptysis, weight loss, higher bacillary load (16) and cavitary disease (31). In Harare, where these similar strains have been found to predominate, it is unknown if tuberculosis caused by these strains will present with a differential clinical spectrum.

Despite the contribution of the SAF1/RD^{Rio} genotypes to TB burden in Zimbabwe, very little is known about its association with clinical and demographic factors. We hypothesize that specific host and pathogen characteristics exist in this population that may be associated with disease due to *M. tuberculosis* SAF1/RD$^{\mathbf{Rio}}$ strains.

As part of a study investigating the causes of chronic cough in Harare (32) we undertook firstly; to genotype *M. tuberculosis* isolates from TB patients identified in the study, secondly; to analyze potential risk factors if any associated with these SAF1/RD$^{\mathbf{Rio}}$ strains and thirdly to look at clinical consequences of disease due to SAF1/RD$^{\mathbf{Rio}}$ strains that are disproportionately represented in this community.

Methods

Study design

A prospective cohort of ambulatory patients was recruited from 2 primary health care clinics in Mbare, a high density suburb of Harare between April-September 2003. A

systematic approach was used where recruitment was limited to weekdays and to the first five patients with chronic cough for each clinic per day. This study was censored to include only patients enrolled in the first 5 months of the study because of inflation related costs of preparation and shipping of cultures.

Study setting and population

Mbare is a high-density, southern suburb of Harare with an estimated population of 300 000 people. Patients presenting with symptoms suggestive of pulmonary TB (cough ≥3 weeks) are treated after diagnosis, according to the guidelines of the National TB control Programme (NTP). Though NTP activities have been fully integrated within primary health care, diagnostic facilities are only provided at the referral hospital Beatrice Road Infectious diseases hospital (BRIDH). In this study participants were investigated at primary health care clinics.

Participants were enrolled to the study if they had had a cough for ≥3 weeks (regardless of other symptoms) and were ≥16 years. Only those individuals with culture confirmed tuberculosis were included in the study. Potential participants were excluded if they were receiving treatment for TB or required immediate admission to hospital, were unwilling to undergo confidential HIV testing or did not reside in Mbare suburb.

Investigations

All participants completed a questionnaire and provided a specimen for confidential HIV testing with written informed consent. Diagnostic evaluation followed a preset algorithm

of first- and second-line management and physical examination (32). Following examination 3 sputum specimens were collected for TB microscopy and culture and send to the Biomedical Research and Training Institute (BRTI) in Harare for microscopy, culture and organism identification. Radiographs were obtained on the first day and if abnormal again on day 7 and were graded by 2 independent readers (33).

Case definitions and clinical data

During the study the diagnosis of TB was confirmed if two or more positive results of sputum smear or culture plus compatible symptoms were identified. Diagnosis of probable TB was made if one positive result of sputum smear and/or positive result of culture with compatible clinical or radiological illness, failure to respond to broad spectrum antibiotics and response to TB treatment by 1 month (32). Patients with a positive diagnosis were referred to the city health department for clinical management and follow up. Clinical data was collected using a questionnaire and this included gender, age, duration of cough, signs and symptoms, history of previous treatment for TB, household TB contact, smoking and exposure to heavy dust or other irritants.

Laboratory methods

HIV testing

Confidential HIV testing was performed using the Determine test (Abbott). All positive results and 1 in 10 negative results were confirmed with Unigold (Trinity Biotech).

Bacterial strains

Concentrated sputum specimens were stained with auramine and examined under fluorescent light and positive slides were confirmed with Ziehl-Nieelsen (ZN). Mycobacterial culture was performed with Lowenstein-Jensen (LJ) slopes and Kirchner's media after decontamination with 4% sodium hydroxide. All positive LJ slants were shipped to Stellenbosch University. A fresh LJ was re-inoculated at Stellenbosch LJ cultures and were incubated at 37°C until there was confluent growth to ensure adequate DNA for subsequent IS*6110* genotyping. Only one isolate per patient was included in the analysis. Cultures were also tested for sensitivity to isoniazid (H) and rifampicin (R), using BACTEC MGIT 460 at critical drug concentrations of 0.1 mg/ml for H and 2.0 mg/ml for R respectively. Drug susceptibility testing for streptomycin (S) and ethambutol (E) at critical drug concentrations of 2.0 mg/ml and 2.5 mg/ml, respectively, were only done if there was any resistance to H or R.

DNA extraction

DNA was extracted from *M. tuberculosis* cultures using previously described methods (34). IS*6110*–genotyping was performed using the international standard typing method for *M. tuberculosis* (35).

DNA fingerprints were analyzed using Gelcompar (Version 4.0, Applied Maths, BVBA, Kortrijk, Belgium) and entered into a database. Cluster analysis was done using the Unweighted Pair Group Method with Arithmetic mean (UPGMA), based on Dice Coefficient. The isolates were classified into strain families according to the similarity of

the IS*6110*-genotype patterns. Isolates with an IS*6110* similarity index of ≥70% were grouped into the same strain family (35). Isolates within strain families were classified into clusters of identical strains (strains sharing an identical IS*6110* genotype pattern therefore considered as clustered) or unique (strains with unique genotype patterns therefore assumed to be unique cases).

To allow classification of the strains into known family groupings the Harare dataset for the Mbare community was compared initially to the IS*6110* genotype database of isolates from the epidemiological study site in Cape Town and the previous database from Harare and Gweru, Zimbabwe (both databases are maintained at Stellenbosch University) (12, 15), to identify similarities in IS*6110* genotype patterns .

Spoligotyping

M. tuberculosis isolates were spoligotyped by amplifying the Direct Repeat (DR) of the *M. tuberculosis* genome using the internationally standardized PCR protocol (36). To determine the reproducibility of the spoligotyping method, *M. tuberculosis* H37Rv DNA was amplified and included on each blot as a reference strain. The amplified products were hybridized to a set of 43 immobilized oligonucleotides each corresponding to one of the spacer DNA sequences in the DR locus. Resultant autoradiographs were analyzed manually and the binary pattern was entered into Microsoft Excel database (Microsoft Corporation, Redmond WA USA). The spoligotypes were compared with the spoligotypes deposited in the SpolDB4 worldwide database of the Pasteur Institute of Guadeloupe (available at http://www.pasteurguadeloupe.fr:8081/SITVITDemo).

Statistical analysis

Data management

Statistical analyses were performed using STATA 10.0 (Statacorp, College Station, Texas USA). An initial analysis was done by merging the clinical database with the IS*6110* and spoligotype datasets.

Study outcomes

For the primary outcome *M. tuberculosis* strains were categorized into two major groups, the SAF1/RD^{Rio} strains versus all other strains families.

The clinical outcome of cavitary disease was considered together with secondary outcomes of smear status and presence of three or more TB symptoms. Patients were coded as having cavitary disease if this was detected on chest X-ray. Disease severity scored as 0-6 based on the American thoracic society system (33) were further categorized as 0=none, less severe (a disease score of 2-3) and more severe (a disease score of 4-6).

Smear positive isolates were categorized as positive and all scanty and negative smear isolates categorized as negative. As all participants enrolled to the study had a cough for ≥3 weeks they were asked if they had any other symptoms including fever, night sweats and weight loss in addition to the cough. In this analysis the number of symptoms present was used as a proxy to severity of disease.

Clustering was also considered as a binary outcome variable with strains categorized as clustered if more than one isolate had the same IS*6110* pattern and unique if the IS*6110* pattern was not shared with any other isolate.

Exposure variables considered

To establish if specific exposures predisposed to infection with the predominant strain family, the following exposures were considered: gender, age, HIV status, prior history of TB, household contact, smoking and exposure to heavy dust or irritants. Participants reporting having shared a home with someone on TB treatment were considered to have had a household contact with someone with TB.

In univariable analyses each exposure variable was cross tabulated with strain family for all cases. Unadjusted odds ratios (OR) and their 95% confidence intervals (CIs) were estimated using logistic regression. Stratification of the analysis by HIV status was limited by power due to the limited numbers in the HIV negative group. Therefore we present here an analysis for all TB patients and a sub-analysis for the HIV positive group.

For the clinical outcome, cavitary tuberculosis and other secondary outcomes, strain family was considered as an exposure variable alongside exposure variables above. Each variable was cross tabulated with the secondary outcome (smear status or TB symptoms reported) and odds ratios and 95% confidence intervals estimated using logistic regression.

All variables with evidence of an association with the outcome (p-value<0.05) in the univariable analysis were included in the multivariable model and adjusted odds ratios (aOR) and their 95% CIs estimated. The likelihood ratio test (LRT) was used to assess the overall associations.

Ethical consideration

Written informed consent was provided by all participants. Approval for the study was granted by the Ethics committees of the Biomedical Research Institute, Harare Zimbabwe, The Medical Research Council of Zimbabwe, the London School of Hygiene and Tropical Medicine and the Ethics Committee (Institutional Review Board) of the Faculty of Health Sciences Stellenbosch University. Voluntary counseling and testing was provided to all participants wanting to know their status, with onward referral for HIV-positive participants.

Results

In the 5 month period between 2 April 2003-1 September 2003, a total of 336 patients presenting with a cough ≥3 weeks (regardless of other symptoms) and culture positive for TB were enrolled to the study. Of these 149 (44.3%) were confirmed as having TB (two or more positive results of sputum smears or culture plus compatible illness). Culture was positive for 134 of 149 (89.9%) TB patients and these were included in this study.

Of the 134 TB patients included in the study 119 (88.8%) were HIV positive and 15 (11.2%) were HIV negative. The majority of the participants were males (56.0%), median age of 30 years (inter-quartile range 27-38 years), 61.9% reported having a cough for

more than 5 weeks and 58.2% were smear positive (Table 1). 16.4% of the TB patients who reported having been previously on TB treatment were all HIV positive. Cavitary TB was reported in 31.3% of the participants (Table 1).

Table 1: Characteristics of TB patients with genotyped *M. tuberculosis* isolates

Characteristics*	All (n=134)	HIV positive (n=119)	HIV negative (n=15)
	n (column %)	n (column %)	n (column %)
Age (years)	**30(16-73)	**30(16-56)	**30 (19-73)
18-24	16 (12.0)	13 (10.9)	3 (21.4)
25-29	41 (30.8)	38 (31.9)	3 (21.4)
30-34	30 (22.6)	28 (23.5)	2 (14.3)
35-39	16 (12.0)	13 (10.9	3 (21.4)
≥40	30 (22.6)	27 (22.8)	3 (21.4)
Gender			
Female	59 (44.0)	54 (45.38)	5 (33.3)
Male	75 (56.0)	65 (54.6)	10 (66.7)
Smear status			
Positive	78 (58.2)	69 (58.0)	9 (60.0)
Negative	56 (41.8)	50 (42.0)	6 (40.0)
Duration of cough (weeks)			
3-4	51 (38.1)	45 (37.8)	6 (40.0)
≥5	83 (61.9)	74 (62.2)	9 (60.0)
Past TB treatment			
Yes	22 (16.4)	22 (18.5)	0 (0.0)
No	112 (83.6)	97 (81.5)	15 (100.0)
Symptoms and signs			
Purulent sputum	31 (23.1)	27 (23.1)	4 (26.7)
Haemoptysis	27 (20.1)	22 (18.6)	5 (33.3)
Felt Feverish	129 (96.3)	117 (98.3)	12 (80.0)
Night sweats	123 (91.2)	112 (94.1)	11 (73.3)
Loss of weight	271 (71.7)	117 (98.3)	13 (86.7)
Household TB contact			
Yes	94 (70.1)	83 (69.8)	11 (73.3)
No	40 (29.9)	36 (30.2)	4 (26.7)
Exposure to occupation heavy dust			
Yes	24(18.0)	22(18.6)	2(13.3)
No	109(82.0)	96(81.4)	13(86.7)
Smoking status			
Never smoked	93 (69.0)	82 (73.9)	11 (73.3)
Current smoker	13 (9.7)	11 (9.9)	2 (13.3)
Former smoker	20 (14.9)	18 (16.2)	2 (13.3)
Cavitary tuberculosis			
Yes	42 (31.3)	34 (28.8)	8 (53.3)
No	92 (68.7)	84 (71.2)	7 (46.7)
Cavitation			
None	92 (68.7)	84 (71.2)	7 (46.7)
mild	16 (11.9)	13 (11.0)	3 (20.0)
severe	26 (19.4)	21 (17.8)	5 (33.3)
***Diseased lung zones			
None	45 (33.6)	41 (34.8)	4 (26.7)
Less extensive	51 (38.1)	43 (36.1)	7 (46.7)
More extensive	38(28.4)	34 (28.8)	4 (26.7)

Data missing for age (n=1), HIV status (n=2), duration of cough (n=1), symptoms and signs (n=5), household contact (n=2), exposure to occupational heavy dust (n=1), smoking status (n=8), **Median (Range) years; ***Diseased lung zones: none=score 0; less extensive=score≤3; more extensive=score≥4

IS*6110* restriction fragment length polymorphism patterns

A total of 134 isolates had both IS*6110* RFLP and spoligotype pattern. Strains with more than six IS*6110* bands (high copy number strains) were identified in 125 (93.3%). Using the classification by Richardson *et al* (35), a total of 18 IS*6110* high copy number strain families could be identified and the most common were F9 (n=70), F13 (n=10), F11 (n=6), F14 (n=6), F28 (n=5) and F29 (n=8). Low copy number strains were also identified from 9 (6.7%) patients.

Seventy-nine (59.0%) isolates were clustered into 28 clusters with sizes ranging from 2-7 isolates. The remaining 55 (41.0%) isolates had IS*6110* RFLP patterns not shared with any other isolate in this cohort (unique).

Genotypes identified by spoligotyping

By comparing the spoligotype patterns with those deposited on the SpolDB4 database, 13 genotypes were identified from the 134 isolates. Ten patterns from 20 isolates did not matched any of the 1939 shared types (STs) representing 39 295 strains posted on the SpolDB4 database.

The five most prevalent genotypes were the LAM (54.5%), T genotypes (9.7%), the IS*6110* low banding family X genotypes (6.7%), the Beijing genotypes (6.0%) and the Central Asian (CAS) genotypes (5.2%) (Figure 1), which included both CAS1_Dehli and CAS1_Kili genotypes. A large proportion of the isolates (14.9%) had spoligotype patterns that were not matched to any type on the SpolDB4 database and are referred to

as orphan strains in this study. One isolate was identified as F9 using IS*6110* and as belonging to the T genotype using spoligotyping possibly pointing to a mixed infection.

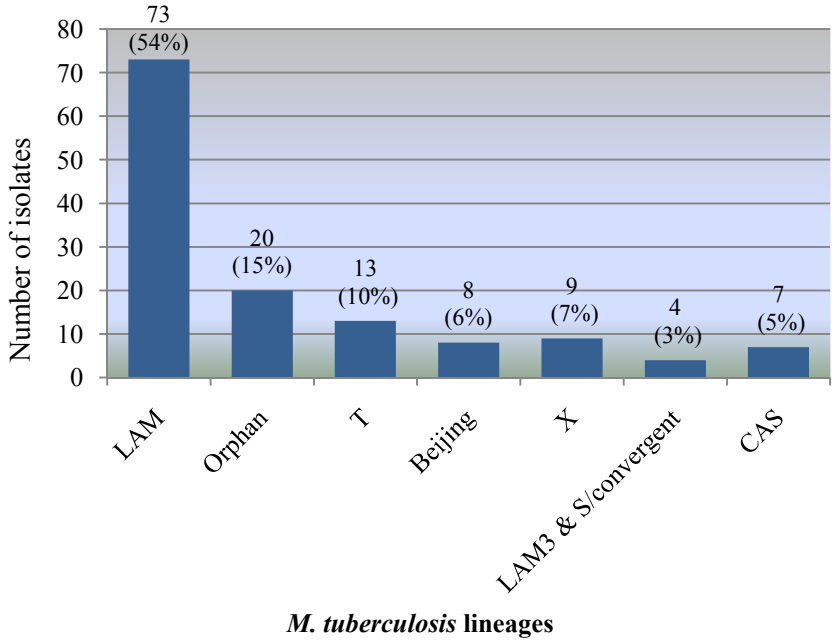


Figure 1: *Mycobacterium tuberculosis* spoligotypes identified from patients presenting with tuberculosis in Mbare Harare

The SAF1/RD^{rio} genotype

The SAF1/RD^{Rio} strains were identified among *M. tuberculosis* isolates from 69 (51.5%) patients, excluding the isolate with more than one genotype (mixed), 58(84.0%) of whom were HIV positive. Among the SAF1/RD^{Rio} strains 15 different spoligotype patterns were identified, 9 of which were not matched to any STs in the SpolDB4 database (Figure 2). Using IS*6110* RFLP a total of 46 IS*6110* RFLP patterns were observed among the 69 isolates. Of these 37 (53.6%) isolates were clustered into 14 different IS*6110* RFLP patterns with cluster size ranging from 2-7. The remaining 32 (46.4%) isolates were unique. On univariable analysis clustering was not associated with disease due to SAF1/RD^{Rio} strains (OR 0.78, 95% CI 0.39-1.56; p=0.49).

Genotype	SpoliDB4 code	Number of isolates(n)	Spoligotype pattern
LAM11_ZWE	807	1	
LAM11_ZWE	84	1	
LAM11_ZWE	812	3	
LAM11_ZWE	813	3	
LAM11_ZWE	815	5	
LAM11_ZWE	59	24	
LAM4	811	18	
*#N/A	#N/A	3	
#N/A	#N/A	2	
#N/A	#N/A	2	
#N/A	#N/A	2	
#N/A	#N/A	1	
#N/A	#N/A	1	
#N/A	#N/A	1	
#N/A	#N/A	1	
#N/A	#N/A	1	
Total		69	

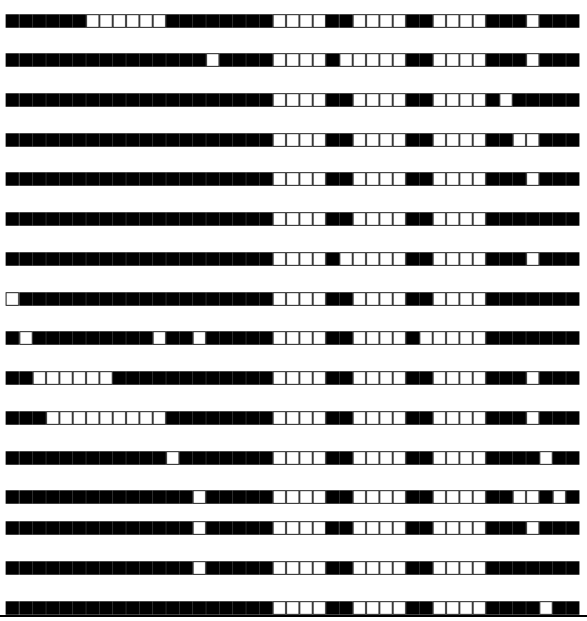


Figure 2: Spoligotype patterns and genotypes of the SAF1/RD^{Rio} strains (n=69) as defined from SpolDB4 database

*#N/A: These are spoligotype patterns that were not matched to any of those in the SpolDB4 database

Risk factors for infection with SAF1/RD^{Rio} strains

Among all TB cases, though not significant, there was a trend towards an association between those who were currently smoking or ever smoked with disease due to SAF1/RD^{Rio} strains (OR 2.01, 95% CI 0.94-4.29; p=0.06) and so was HIV status (OR 2.89, 95% CI 0.87-9.59; p=0.06), with HIV negative TB patients more likely to present with disease due to SAF1/RD^{Rio} strains (Supplemental table 1). These variables were included in a multivariable analysis together with gender and age as they are likely confounders of smoking. After controlling for both gender and age, smoking was

associated with disease due to SAF1/RD^{Rio} strains (aOR 3.33, 95% CI 1.25-8.83; p=0.02) (Table 2).

Table 2: Factors associated with disease due to *Mycobacterium tuberculosis* SAF1/RD^{Rio} strains among culture positive TB patients-multivariable analysis

	Outcome: *Mycobacterium tuberculosis* SAF1/RD^{Rio} strains				
	*Unadjusted OR (95% CI)	**P value	***Adjusted OR (95% CI)	**** value	P
All patients					
Age (years)					
18-24	1	0.24	1	0.25	
25-29	0.71 (0.22-2.26)		0.65 (0.19-2.21)		
30-34	1.0 (0.30-3.37)		1.00 (0.28-3.58)		
35-39	3.0 (0.67-13.4)		2.93 (0.62-13.80)		
≥40	1.14 (0.34-3.84)		0.98 (0.27-3.55)		
Gender					
Male	1	0.89	1	0.13	
Female	0. 95 (0.48-1.88)		1.99 (0.81-4.88)		
TB treatment history					
No	1	0.38			
Yes	0.93 (0.37-2.32)				
HIV status					
Positive	1	0.06	1	0.08	
Negative	2.89 (0.87-9.59)		3.06 (0.87-10.76)		
Currently/ever smoked					
No	1	0.06	1	0.02	
Yes	2.01 (0.94-4.29)		3.33 (1.25-8.83)		

*Overall P value for univariable association

Adjusted Odds ratio from logistic regression representing the summary relative odds of having disease due to *M. tuberculo*sis Southern Africa 1 (SAF1)/RD^{Rio} strains compared with other strains, adjusting for all factors listed; CI confidence interval; * Overall P value for multivariable association

In the unadjusted analysis among HIV co-infected TB patients smoking was associated with disease due to SAF1/RD^{Rio} strains (OR 2.21, 95% CI 0.99-4.88; p=0.04) (Supplemental table 2). TB patients also co-infected with HIV, reporting having had contact with someone within their household who had TB were more likely to present

with disease due to SAF1/RD^{Rio} strains, though this was not significant (OR 2.09, 95% CI 0.94-4.69; p=0.07) similarly those exposed to heavy dust or irritants (OR 0.41, 95%CI 0.15-1.10, p=0.07). These were included in a multivariable analysis together with gender and sex as they are likely confounders of smoking. Smoking remained strongly associated with disease due to SAF1/RD^{Rio} strains (aOR 3.50, 95% CI 1.20-10.20; p=0.02) (Table 3).

Table 3: Factors associated with disease due to *Mycobacterium tuberculosis* SAF1/RD^{Rio} strains, among HIV positive TB patients with culture positive disease-multivariable analysis

	Outcome: *Mycobacterium tuberculosis* SAF1/RD^{Rio} strains			
	Unadjusted OR (95% CI)	*P value	**Adjusted OR (95%CI)	***P value
HIV positive TB patients				
Age (years)				
18-24	1	0.56	1	0.70
25-29	0.85 (0.24-3.01)		0.94 (0.25-3.55)	
30-34	1.17 (0.31-4.35)		1.36 (0.34-5.48)	
35-39	2.63 (0.53-13.07)		2.48 (0.46-13.06)	
≥40	1.08 (0.29-4.08)		1.15 (0.28-4.75)	
Gender				
Male	1	0.90	1	0.23
Female	0.96 (0.47-1.97)		1.82 (0.68-4.89)	
Household TB contact				
No				
Yes	1	0.07	1	0.15
	2.09 (0.94-4.69)		1.87 (0.80-4.38)	
Currently/ever smoked				
No	1	0.04	1	0.02
Yes	2.21 (0.99-4.88)		3.50 (1.20-10.20)	
Exposure to dust or irritants				
No	1	0.07	1	0.06
Yes	0.41 (0.15-1.10)		0.38 (0.14-1.05)	

*Overall P value for univariable association

**Adjusted odds ratio from logistic regression representing the summary relative odds of having disease due to *M. tuberculosis* Southern Africa 1 (SAF1) /RD^{Rio} strains compared with other *M. tuberculosis* strains, adjusting for all factors listed; ***Overall P value for multivariable association; CI: confidence interval

Clinical consequences of TB disease due to SAF1/RD^{Rio} strains

Cavitary tuberculosis was reported among 42/134 (31.3%) TB patients, being more common among HIV negative (53.3%) compared to HIV positive (28.8%) patients (Table 1). In an unadjusted analysis, disease due to *M. tuberculosis* SAF1/RD^{Rio} was associated with a higher odds of cavitary tuberculosis (OR 2.20, 95% CI 1.03-4.67;

p=0.04) (Supplemental table 3). In the multivariable analysis this remained associated with cavitary tuberculosis (Table 4).

Table 4: Factors associated with cavitary tuberculosis disease, among culture positive TB patients –multivariable analysis

	Unadjusted OR (95% CI)	*P value	**Adjusted OR (95% CI)	***P value
All patients				
Age (years)				
18-24	1	0.09	1	0.09
25-29	0.20 (0.58-0.69)		0.23 (0.06-0.81)	
30-34	0.26 (0.07-0.92)		0.27 (0.07-1.00)	
35-39	0.20 (0.04-0.91)		0.15 (0.03-0.75)	
≥40	0.22 (0.06-0.80)		0.22 (0.06-0.85)	
Strain family				
Other	1	0.04	1	0.05
SAF1/RD^{Rio}	2.20 (1.03-4.67)		2.29 (1.00-5.27)	
Currently/ever smoked				
No	1	0.50	1	0.64
Yes	0.76 (0.33-1.71)		0.81 (0.33-1.96)	
HIV status				
Positive	1	0.06	1	0.25
Negative	2.82 (0.95-8.40)		2.03 (0.61-6.76)	

* Overall P value for univariable association;**Adjusted odds ratio from logistic regression representing the summary relative odds of having cavitary tuberculosis among culture positive TB patients, adjusting for all factors listed; ***Adjusted odds ratio; ****Overall P value for multivariable association; CI confidence interval

Though not significant, HIV negative patients were more likely to present with cavitary tuberculosis than HIV positive patients (Supplemental table 3). Among TB patients co-infected with HIV, cavitary tuberculosis was not associated with any of the factors explored (Supplemental table 4).

Smear positive disease was relatively common being identified in 78/134 (58.2%) culture positive TB patients (Table 1). In an unadjusted analysis, though not significant, TB patients with more extensive diseased lung zones were more likely to have smear positive disease (OR 1.64, 95% CI 0.65-4.11; p=0.06) (Table 5). Though not significant, TB patients with disease due to *M. tuberculosis* SAF1/RD^{Rio} strains were more likely to be smear positive (OR 1.81 95% CI 0.91-3.64; p=0.08) (Table 5), and similarly patients with cavitary tuberculosis were more likely to be smear positive (OR 1.99 95% CI 0.92-4.32; p=0.07) (Table 5). After adjusting for age, strain family and smoking, no trend in association between diseased lung zones and smear status was observed.

Among 119 HIV positive patients 69 (58%) patients had smear positive disease (Table 1). Within this group disease due to *M. tuberculosis* SAF1/RD^{Rio} strains was not associated with being smear positive, however, HIV positive patients with cavitary disease were more likely to be smear positive than those without (OR 2.65, 95% 1.11-6.34; p=0.02). After adjusting for age the association was no longer significant.

Table 5: Factors associated with smear positive tuberculosis among culture positive TB patients

		Outcome: Smear positive				
	Total (All cases)	All TB cases (row %)	*Unadjusted OR (95% CI)	**P value	***Adjusted OR (95% C1)	****P value
All patients	n=134	n=78				
Age (years)						
18-24	16	10 (62.5)	1	0.89	1	0.84
32-29	41	25 (61.0)	0.93 (0.28-3.08)		1.34(0.36-4.97)	
30-34	30	15 (50.0)	0.60 (0.17-2.07)		0.81 (0.21-3.05)	
35-39	16	9 (56.3)	0.78 (0.18-3.17)		0.76 (0.17-3.54)	
≥40	30	18 (60.0)	0.90 (0.26-3.13)		1.13 (0.29-4.40)	
Gender						
Male	75	46 (61.3)	1	0.40		
Female	59	32 (54.2)	0.75 (0.37-1.49)			
Strain family						
Other	65	33 (50.8)	1	0.08	1	0.15
SAF1/RD^{Rio}	69	45 (65.2)	1.81 (0.91-3.64)		1.73 (0.82-3.65)	
HIV status						
Positive	119	69 (58.0)	1	0.88		
Negative	15	9 (60.0)	1.08 (0.36-3.24)			
Cavitary TB						
No	91	48 (52.8)	1	0.07	1.73 (0.68-4.37)	0.25
Yes	42	29 (69.1)	1.99 (0.92-4.32)			
*****Diseased lung zones						
None	45	27 (60.0)	1	0.06	1	0.09
Less extensive	50	23 (46.0)	0.57 (0.25-1.28)		0.67 (0.19-1.13)	
More extensive	38	27 (71.0)	1.64 (0.65-4.12)		1.23 (0.42-3.63)	

*Unadjusted Odds ratio from logistic regression representing the summary relative odds of presenting with three or more symptoms compared with less among TB patients; CI confidence interval;**P value for univariable association

Adjusted Odds Ratio; CI confidence interval;*P value for multivariable association

***** Diseased lung zones: none=score 0; less extensive=score≤3; more extensive=score≥4

At the time of presentation a majority of the TB patients (121/134; 90.3%) had three or more symptoms of TB in addition to a cough ≥3 weeks. Disease due to *M. tuberculosis* SAF1/RD^{Rio} strains was not associated with the presence of three or more symptoms. In the unadjusted analysis older age groups, male gender, HIV positive TB patients and patients without cavitary TB were more likely to present with three or more symptoms of TB (Table 6). Gender and HIV status remained associated with presence of three or more symptoms in the adjusted analysis (Table 6).

Table 6: Factors associated with presence of three or more symptoms, among TB patients with culture positive disease

		Outcome: Presence of three or more symptoms				
	Total (All cases)	All TB cases (row %)	*Unadjusted OR (95% CI)	**P value	***Adjusted OR (95% C1)	****P value
All patients	n=134	n=134				
Age (years)						
18-24	16	13 (81.3)	1	0.04	1	0.08
25-29	41	38 (92.7)	2.92 (0.52-16.3)		0.80 (0.08-7.71)	
30-34	30	24 (80.0)	0.92 (0.19-4.31)		0.18 (0.02-1.75)	
35+	46	45 (97.8)	10.4 (0.99-108.42)		14.1 (0.67-296.9)	
Gender						
Male	75	71 (94.7)	1	0.05	1	0.05
Female	59	50 (84.8)	0.31 (0.09-1.07)		0.18 (0.01-0.99)	
Strain family						
Other	65	59 (90.8)	1	0.85		
SAF1/RD^{Rio}	69	62 (89.9)	0.90 (0.29-2.83)			
HIV status						
Positive	119	112 (94.1)	1	<0.001	1	<0.001
Negative	15	9 (60.0)	0.09 (0.03-0.34)		0.02 (0.002-0.16)	
Cavitary TB						
No	91	86 (94.5)	1	0.02	1	0.08
Yes	42	34 (81.0)	0.25 (0.08-0.81)		0.25 (0.06-1.18)	
*****Diseased lung zones						
None	45	44 (97.8)	1	0.07		
Less extensive	50	43 (86.0)	0.14 (0.02-1.18)			
More extensive	38	33 (86.8)	0.15 (0.02-1.35)			

*Unadjusted Odds ratio from logistic regression representing the summary relative odds of presenting with three or more symptoms compared with less among TB patients; CI confidence interval;**P value for univariable association; ***Adjusted Odds ratio; CI confidence interval; ****P value for multivariable association ***** Diseased lung zones: none=score 0; less extensive=score≤3; more extensive=score≥4

Drug susceptibility testing

Of the 104 *M. tuberculosis* isolates with drug susceptibility test results, 2(1.9%) were resistant to H and only one (0.96%) to R. Due to this low level of resistance the drug susceptibility testing results were not considered as a potential risk factor in the univariable analysis.

Discussion

This study confirms that the SAF1/RD^{Rio} *M. tuberculosis* strains which belong to the LAM genotype continue to cause disproportionate number of cases of TB in Harare, Zimbabwe. To begin to understand host factors in disease due to SAF1/RD^{Rio} strains the present study looked at the strain diversity, demographic and clinical consequences among TB patients presenting at peripheral clinics and not only at a referral hospital therefore more likely to be representative of the general community in this high density suburb. The SAF1/RD^{Rio} strains characterized as strains of the LAM genotypes were the cause of TB in 51.5% of the TB patients. Other genotypes were identified, including the globally emerging Beijing, CAS, well characterized X and the ill defined T genotypes, these but were identified in much lower proportions.

Prevalent among the SAF1/RD^{Rio} genotype were the LAM11_ZWE strains which have been reported as predominant from this community in Zimbabwe over 13 years (26, 27). The reasons for the predominance of the SAF1/RD^{Rio} genotype and how this *M. tuberculosis* population structure affects disease dynamic also remain unclear. The SAF1/RD^{Rio} strains displayed a rich diversity in IS*6110* DNA fingerprint (79 patterns)

and spoligotype patterns (15 patterns). This rich diversity may point to endemic strains that have been circulating in this community for a long time and due to the several transmission events may have either lost or gained IS*6110* bands hence generating the more fitter/successful strains that are well adapted to transmit and cause secondary cases in this specific population. This finding is indicative but not full proof of host-pathogen associations. Additional support for this host pathogen association was provided in findings from other studies (16, 23, 37). Strains belonging to the LAM genotype, including these SAF1/RD^{Rio} strains have also been recognized as playing a role in the TB epidemic, accounting for approximately 15% of the strains submitted to the SpolDB4 database (25) and have been associated with multidrug resistant (MDR)(38) and extremely drug resistant (XDR)TB (30). Studies by Gagneaux *et al* (23) looking at lineage specific secondary case rate were more indicative of *M. tuberculosis* genotypes that are adapted to specific populations.

Our findings show that culture positive HIV negative, TB patients with disease due to SAF1/RD^{Rio} strains are twice more likely to have cavitary disease and that they were more likely to be smear positive. However risk factor analyses could not be done in this subgroup because the number of such individuals was very small. By looking at all the TB patients regardless of HIV status an association between cavitary disease and strain family was observed. Not surprising, in the sub group analysis that included only the HIV co-infected TB patients, there was no association of disease due to SAF1 and cavitary disease.

In a study in Belo Horizonte in Brazil where TB cases with clinical suspicion of multidrug resistance were studied, patients with disease due to RD^{Rio} were eight times more likely to have cavitary lung disease (31). However in this study 70/87 (80.5%) TB patients tested for HIV were HIV negative unlike the population in our study where a large proportion was HIV positive. However by looking at the TB patients, regardless of HIV status, both studies show an association with cavitary disease. The lack of association between positive smear status and cavitary disease was surprising and contradicted earlier studies which found an association between cavitary TB and high bacillary burden and transmissibility (39). Similar associations between other predominant genotypes and cavitary disease were found in Western India. However associations between cavitary pulmonary disease and smear status were not established in this (40).

The finding of an association between *M. tuberculosis* genotype and cavitary disease highlights the importance of pathogen characteristics in the epidemic. This together with earlier findings of faster rate of growth in the LAM genotypes (41), may imply a fitness advantage.

This study is the first to show an association between TB due to SAF1/RD^{Rio} and smoking. Smoking was reported in 31% of the participants and an association was observed with *M. tuberculosis* SAF1/RD^{Rio} genotype. Smoking has long been associated with TB and interacts with this disease by increasing disease susceptibility and worsening outcomes. Lung damage caused by smoking could enhance the risk of progression to

disease. A TB disease dynamics model using data from Zimbabwe showed that smoking enhanced transmission, progression to active disease and the number of active TB cases increased as number of smokers increased (42). It is not apparent, however, from our study how much of the transmission of SAF1/RD^{Rio} strains is attributable to smoking.

Despite the lack of a strong association between smear status and main outcome measures, the proportion of smear positive disease was high (58.2%). In addition clustering was high (58% by IS*6110* RFLP), clearly indicating that the epidemic is driven by transmission, possibly as a consequence of diagnostic delay and similarly cavitary disease. In earlier studies in the same population with high HIV prevalence, a substantial burden of undiagnosed smear positive TB disease (0.4%) was reported (43). Our findings underscore the need for early diagnosis and treatment of all TB cases in particular the infectious cases in this population with high prevalence of HIV.

Limitations include a potential inclusion bias of patients with disease due to SAF1/RD^{Rio} strains. The recruitment was restricted to week days, over 5 months and only the first 5 participants with chronic cough at each clinic were included. Stratification of the analysis by HIV status was limited by power, because the numbers of TB patients that were HIV negative was very small. The finding of SAF1/RD^{Rio} as predominant strains was reported in a previous study though in current study the proportion of disease due to these strains is much higher.

In conclusion SAF1/RD^{Rio} *M. tuberculosis* strains are the most prevalent genotype causing TB in Harare Zimbabwe where 88% of the population were HIV positive. These strains are well adapted to this population. Specific host factors such as smoking and pathogen characteristics such as ability to cause smear positive and cavitary TB may play a significant role in the transmission, acquisition and outcome of disease. Early detection of cases and treatment of cases are fundamental in controlling the epidemic.

Acknowledgements

We would like to thank the Harare City Health Department staff, Mavis Mudzana and Vongai Madamombe for running the Project clinics. We thank Beauty Makamure, Julitta Magwenzi, Tafadzwa Mavhudzi and Mrs Mandisodza for their input in the study. We thank Beauty Makamure for arranging shipment of the isolates to Stellenbosch, Erica Engelke and Marian de Kok for assistance with laboratory sub culture and drug susceptibility testing.

Financial support:

This study was funded by Rockefeller Foundation; E L. Corbett is supported by a Career Development Fellowship from the Wellcome Trust. The molecular epidemiology component was funded through the Wellcome Trust Research Leave Fellowship Grant (01680/2/00/2 to V N. Chihota).

References

1. **World Health Organization (WHO).** Global tuberculosis control: Short update to the 2009 report. WHO/HTM/TB/2009.426: WHO 2009 Geneva. Available from http://www.who.int/tb/publications/global_report/2009/update/en/index.html accessed 19 January 2010

2. **Reider HL.** Epidemiologic basis of tuberculosis control. First Edition 1999. International Union Against tuberculosis and Lung disease.

3. **Aaron L, Saadoun D, Calatroni I, Launay O, Mémain N, Vincent V, Marchal G, Dupont B, Bouchaud O, Valeyne D and Lortholary O**. Tuberculosis in HIV-infected patients: a comprehensive review. *Clin Microbiol Infect* 2004; 10: 388-398

4. **Daley CL, Small PM, Schecter GF, Schoolnik GK, McAdam RA, Jacobs WR Jr and Hopewell PC.** An outbreak of tuberculosis with accelerated progression among persons infected with human immunodeficiency virus. An analysis using restriction fragment length polymorphism. *N Engl J Med* 1992; 326:231-5

5. **Shafer RW, Singh S, Larkin C and Small PM.** Exogenous reinfection with multidrug resistant *Mycobacterium tuberculosis* in immunocompetent patients. *Tubercle Lung Dis* 1995; 76:573-577

6. **Corbett EL, Watt CJ, Walker N, Maher D, Williams BG, Raviglione MC and Dye C**. The growing burden of tuberculosis: Global trends and interactions with the HIV epidemic. *Arch Intern Med* 2003;163: 1009-1021

7. **WHO report 2009**. Global tuberculosis control http://apps.who.int/globalatlas/predefinedReports/TB/PDF_Files/zwe.pd. Accessed 13th February 2010

8. **Hermans PW, Messadi F, Guerbrexabher H, van Soolingen D, de Haas PE, Heersma H, de Neelin H, Ayoub A, Portaels F and 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; 3:1055-1060.

9. **Small PM, Hopewell PC, Singh SP, Paz A, Parsonnet J, Ruston DC ,Scheter GF, Daley CL and Schoolink 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

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

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

12. **Van de Spuy GD, Kremer K, Ndabambi SL, Beyers N, Dunbar R, Marais BJ, van Helden PD and Warren RM.** Changing *Mycobacterium tuberculosis* population highlights clade specific pathogenic characteristics. *Tuberculosis* 2009; 89: 120-125

13. **European concerted Action** on New Generation Genetic Markers and Techniques for the Epidemiology and control of tuberculosis. Beijing /W Genotype *Mycobacterium tuberculosis* and drug resistance. *Emerg Infect Dis* 2006;12:736-742

14. **Glynn JR, Crampin AC, Traore H, Chaguluka S, Mwafulirwa DT, Alghamdi S, Ngwira B.M, Yates M.D, Drobniweski FD and Fine PE**. Determinants of cluster size in large population-based molecular epidemiology of tuberculosis Northern Malawi. *Emerg Infect Dis*; 2008: 14:1060-1066.

15. **Chihota V, Apers L, Mungofa S, Kasongo W, Nyoni IW, Tembwe R, Mbulo G, Tembo M., Streicher EM, van der Spuy GD, Victor T, van Helden P and Warren RM.** Predominance of a single genotype of *Mycobacterium tuberculosis* in regions of Southern Africa. *Int J Tuberc Lung* Dis; 11:311-318

16. **Lazzarrini LC, Huard RC, Boechat NL, Gomes HM, Oelemann MC, Kurepina N, Shashkina E, Mello F.C.Q, Gibson A., Virginio MJ, Marsico AG, Butler R, Kreiswirth B.N., Suffys PN, Silva JRL and Ho JL.** Discovery of a novel *Mycobacterium tuberculosis* lineage that is a major cause of tuberculosis in Rio de Janeiro, Brazil. *J Clin Microbiol* 2007; 3891-3902

17. **Anh DD, Borgdorff MW, Van LT, van Gorkom T, Kremer K and van Soolingen D**. *Mycobacterium tuberculosis* Beijing genotype emerging in Vietnam. *Emerg Infect Dis* 2000; 6:302-305.

18. **Jou R, Chiang CY, and Huang WL**. Distribution of the Beijing family genotypes of *Mycobacterium tuberculosis* in Taiwan. *J Clin Microbiol* 2005; 43:95-100.

19. **Buu TN, Huyen MN, Lan NTN, Quy HT, Hen NV, Zignol M, Borgdorff MW, Cobelens FGJ and van Soolingen D**. The Beijing genotype is associated with young age and multidrug-resistant tuberculosis in rural Vietnam. *Int J Tuberc Lung Dis* 2009; 13:900-906

20. **KrüüuerA, Hoffner SE, Sillastu H, Danilovits M, Levina K, Svenson SB, Ghebremicahel S, Koivula T and Kallenius G**. Spread of drug resistant pulmonary tuberculosis in Estonia. *J. Clin Microbiol* 2001; 39:3339-3345

21. **Bifani PJ, Plitkaytis BB, Kapur V, Stockbauer K, Pan X, Lutfey SL, Moghazeh W, Eisner W, Daniel TM, Kaplan MH, Crawford JT, Musser JM and Kreiswirth BN**. Origin and interstate spread of a New York City multidrug resistant *M. tuberculosis* clone family. *JAMA* 1996;275:452-457

22. **Baranov AA, Mariandyshev AO, Mannsaker T, Dahle UR and Bjune GA**. Molecular epidemiology and drug resistance of widespread genotypes of *Mycobacterium tuberculosis* in northwestern Russia. *Int J Tuberc Lung Dis* 2009;13:1288-93

23. **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 and Small P**. Variable host-pathogen compatibility in *Mycobacterium tuberculosis*. *Proc Natl Acad Sci U S A* 2006, 103(8):2869-2873.

24. **Gibson AL, Huard RC, Gey van Pittius NC, Lazzarini LCO, Driscoll J, Kurepina N, Zozio T, Sola C, Spindola SM, Kritski AL, Fitzgerald D, Kremer K, Mardassi H, Chitale P, Brinkworth J, Garcia de Viedma D, Gicquel B, Pape JW, van Soolingen D, Kreiswirth BN, Warren RM, van Helden PD, Rastogi N, Suffys PN, Lapa e Silva J, and Ho JL**. Application of sensitive and specific molecular methods to uncover global dissemination of the major RD^{Rio} sublineage of the Latin American Meditteranean *Mycobacterium tuberculosis* spoligotype family. *J Clin Microbiol* 2008; 46: 1259-1267.

25. **Brudey K, Driscoll JR, Rigouts L, Prodinger WM, Gori A, Al-Hajoj SA, Allix C, Aristimuño L, Arora J, Baumanis V, Binder L, Cafrune P, Cataldi A, Cheong S, Diel R, Ellermeier C, Evans JT, Fauville-Dufaux M, Ferdinand S, Garcia de Viedma D, Garzelli C, Gazzola L, Gomes HM, Guttierez MC, Hawkey PM, van Helden PD, Kadival GV, Kreiswirth BN, Kremer K, Kubin M, Kulkarni SP, Liens B, Lillebaek T, Ho ML, Martin C, Martin C, Mokrousov I, Narvskaïa O, Ngeow YF, Naumann L, Niemann S, Parwati I, Rahim Z, Rasolofo-Razanamparany V, Rasolonavalona T, Rossetti ML, Rüsch-Gerdes S, Sajduda A, Samper S, Shemyakin IG, Singh UB, Somoskovi A, Skuce RA, van Soolingen D, Streicher EM, Suffys PN, Tortoli E, Tracevska T, Vincent V, Victor TC, Warren RM, Yap SF, Zaman K, Portaels F, Rastogi N, Sola C.** *Mycobacterium, tuberculosis* complex genetic diversity: mining the fourth international spoligotyping database (SpolDB4) for classification, population genetics and epidemiology. *BMC Microbiology* 2006;6:23

26. **Easterbrook PJ, Gibson A, Murad S, Lamprecht D, Ives N, Ferguson A, Lowe O, Mason P, Ndudzo, Taziwa A, Makombe R, Mbengeranwa L, Sola ., Rastogi N and Drobnisewski F**. High rates of clustering of strains causing tuberculosis in

Harare Zimbabwe: A molecular epidemiological study. *J. Clin Microbiol* 2004; 42:4536-4544

27. **Heydermann RS, Goyal M, Roberts P, Ushewokunze S, Zizhou S, Marshall BG, Makombe R, Vann Embden JDA, Mason PR and Shaw R.J**. Pulmonary tuberculosis in Harare Zimbabwe: analysis by spoligotyping. *Thorax* 1998, 53:346-350

28. **Gutacker MM, Mathema B, Soini H, SHashkina E, Kreiswirth BN, Graviss EA and Musser JM**. Single nucleotide polymorphism based population genetic analysis of *Mycobacterium tuberculosis* strains from 4 geographic sites. *J Infect Dis* 2006; 193:121-128

29. **Gutacker MM, Smoot JC, Migliaccio CA, Ricklefs SM, Hua S, Cousins DV, Graviss EA, Shashkina E, Kreiswirth BN and Musser JM**. Genome-wide analysis of synonymous single nucleotide polymorphisms in *Mycobacterium tuberculosis* complex organisms: resolution of genetic relationships among closely related microbial strains. *Genetics* 2002; 162:1533-1543.

30. **Pillay M and Sturm AW.** Evolution of the extensively drug resistant F15/LAM4/KZN strain of *Mycobacterium tuberculosis* in KwaZulu Natal, South Africa. *Clin Infect Dis* 2007;45:1409–1414

31. **Lazzarini LCO, Spindola SM, Bang H, Gibson AL, Weisenberg S, da Silva Carvalho W, Augusto CJ, Huard RC, Kritski AL and Ho JL**. RD^{rio} *Mycobacterium tuberculosis* infection is associated with higher frequency of cavitary pulmonary disease. *J Clin Microbiol* 2008; 46:2175-2183

32. **Munyati SS, Dhoba T, Makanza ED, Mungofa S, Wellington M, Mutsvangwa J, Gwanzura L, Hakim J, Nyakabau M, Mason PR, Robertson V, Rusakaniko S, Butterworth AE and Corbett EL**. Chronic cough in primary health care attendees, Harare, Zimbabwe: Diagnosis and impact of HIV infection. *CIin Infect Dis* 2008; 40:1818-1827.

33. **American Thoracic Society**. Diagnosis and treatment of tuberculosis. *Am J Resp Crit Care Med* 1990; 142:725-735

34. **Richardson M, van Lill SW, van der Spuy GD, Munch Z, Booysen CN, Beyers N, van Helden PD and Warren RM**. Historic and recent events contribute to the disease dynamics of Beijing-like *Mycobacterium tuberculosis* isolates in a high incidence region. *Int J Tuberc Lung Dis* 2002; 6(11):1001-1011.

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

36. **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 Apr; 35(4):907-14.

37. **Baker L, Brown T, Maiden MC and Drobniewski F**. Silent nucleotide polymorphism and a phylogeny of *Mycobacterium tuberculosis. Emerg Infect Dis* 2004;10:1568-1577

38. **Ignatova A, Dubiley S, Stepanshina V and Shemyakin I**. Predominance of multidrug-resistant LAM and Beijing family strains among *Mycobacterium tuberculosis* isolates recovered from prison inmates in Tula Region Russia. *J. Med Microbiol*. 2006; 55:1413-1418

39. **Driver CR, Macaraig W, McElroy PD, Clark C, Munsiff SS, Kreiswirth BN, Driscoll J and Zhao N**. Which patients' factors predict the rate of growth of *Mycobacterium tuberculosis* clusters in urban community? *Am J Epidemiol* 2006; 164: 21-31

40. **Chatterjee A, D'Souza D, Vira T, Bamne A, Ambe GT, Nicol MP, Wilkinson RJ and Mistry N**. Strains of *Mycobacterium tuberculosis* from Western Maharshtra, India, exhibit a high degree of diversity and strain-specific associations with drug resistance, cavitary disease and treatment failure. *J. Clin Microbiol* 2010; 48:3593-3599

41. **Von Groll A, Martin A, Felix C, Prata PFS, Honscha G, Portaels F, Van Dame P, Almelda da Silva PE and Palomino JC**. Fitness study of the RDrio lin*eage and Latin American-Me*diterranean family of *Mycobacterium tuberculosis* in the city of Rio Grande, Brazil. FEMS *Immunl Med Microbiol* 2010;58: 119-127

42. **Bhunu CP, Mushayabasa S and Tchuenche JM**. A theoretical assessment of the effects of smoking on transmission dynamics of Tuberculo*sis. Bull Math Biol 2010; Aug 20. [Epub ahe*ad of print]

43. **Corbett EL, Bandason T, Cheung YB, Makamure B, Dauya E, Munyati SS, Churchyard GJ, Williams BG, Butterworth AE, Mungofa S, Hayes RH and Mason P**. Prevalent infectious tuberculosis in Harare, Zimbabwe: Burden, risk factors and implications for control. *Int J Tuberc Lung Dis* 2009;13:1231-1237

Chapter 4

Clonal expansion of a globally disseminated lineage of *Mycobacterium tuberculosis* with low IS*6110* copy numbers[5].

R.M Warren, Victor T, Streicher EM, Richardson M, GD van der Spuy, Johnson R,

Chihota VN, Locht C, Supply P and PD van Helden

[4]Published as an original article: R.M Warren, Victor T, Streicher EM, Richardson M, GD van der Spuy, Johnson R, **Chihota VN**, Locht C, Supply P and PD van Helden (2004). Clonal expansion of a globally disseminated lineage of *Mycobacterium tuberculosis* with low IS*6110* copy numbers. *Journal of Clinical Microbiology* 42 (12); 5774-5782.

Abstract

Knowledge of the clonal expansion of *Mycobacterium tuberculosis* and the accurate identification of predominant evolutionary lineages remain limited in this species, especially with regard to low-IS*6110*-copy-number strains. In this study, 170 *M. tuberculosis* isolates with ≤ 6 IS*6110* insertions, identified in Cape Town, South Africa, were characterized by principle genetic grouping, restriction fragment length polymorphism analysis, spoligotyping, IS*6110* insertion site mapping and variable number tandem repeat (VNTR) typing. This analysis indicates that all but one of the isolates analyzed were members of principal genetic group 2 and of the same low-IS*6110*-copy-number lineage. The remaining isolate was a member of principal genetic group 1 and a different low-IS*6110*-copy-number lineage. Phylogenetic reconstruction suggests clonal expansion through sequential acquisition of additional IS*6110* copies, expansion and contraction of VNTR sequences and the deletion of specific direct variable repeat sequences. Furthermore, comparison of the genotypic data of 91 representative low-IS*6110*-copy-number isolates from Cape Town, other Southern African regions, Europe and America suggests that certain low-IS*6110*-copy-number strain spoligotypes and IS*6110* fingerprints have been acquired in a distant past. These clones have subsequently become widely disseminated and now play an important role in the global tuberculosis epidemic.

Introduction

Sequence analysis of *Mycobacterium tuberculosis* strains collected from different geographical settings has shown that the frequency of mutation in this species is extremely low (31). Single nucleotide polymorphism (SNP) and Variable Number Tandem Repeat-(VNTR) based analyses have consistently indicated that the global population of *M. tuberculosis* is highly clonal (1,15,34). Within this clonal structure, strains can be assigned to one of three principal genetic groups according to SNP's in the *katG* and *gyrA* genes (31). Using an expanded array of synonymous single nucleotide polymorphisms (sSNP), these principal genetic groups have been further divided into 8 main clusters to depict the evolution of *M. tuberculosis* (15). Simulations have suggested that the branch points in the SNP derived trees are accurate (1). However, subsequent clonal expansion occurring at each of these branch points remains largely unknown due to the lack of resolution of SNP-based analyses (1).

Various combinations of more variable markers, primarily used for molecular epidemiological studies, have been utilized to identify genotype families which fall into the different clusters defined by SNP's (15). The most commonly used marker is IS*6110*, a transposable element used as a probe in restriction fragment length polymorphism (RFLP) analysis of clinical isolates (35). Genetic relationships between strains have been inferred according to an IS*6110*-RFLP dice similarity index of > 65 % and the inheritance of other specific polymorphisms. This data has been applied to depict clonal expansion in high-copy-number strains with > 6 IS*6110* copies (5,37,40,42) and for their epidemiologic analysis on a global scale (4).

However, only a limited amount of evolutionary data exists for low-IS*6110*-copy-number strains with ≤ 6 IS*6110* hybridizing bands, due to their intrinsically limited IS*6110*-RFLP polymorphism. One report has presented evidence to show evolution of the IS*6110* banding pattern in the progeny of a low-IS*6110*-copy-number strain (22). More recently a high degree of congruence was shown to exist between IS*6110* banding patterns and other markers in strains with few IS*6110* copies collected in London, United Kingdom (7). Accordingly, these strains have been classified into three different groups (7). One of these groups (Group II) includes the principal genetic group 2 clusters IV and V defined by sSNP analysis, representing strains with 1-3 and 4-6 IS*6110* copies, respectively (15). A second distinct group includes strains from principal genetic group 1 (cluster I (15)), which have been associated with patients from East-Africa and Asia (6,30). However, the process of clonal expansion within these groups of low-IS*6110*-copy-number strains remains largely unresolved. Moreover, the genetic relationships between strains within these groups from different geographical regions are poorly understood.

In this study we have used RFLP (42), principal genetic grouping (31), IS*6110* insertion site mapping (6), spoligotyping (19), and PCR analysis of Variable Number of Tandem Repeats (VNTRs) interspersed in multiple loci (14,23,24,33) to determine the genetic relationship between low-IS*6110*-copy-number strains collected in Cape Town, South Africa. This data has been compared to available genotypic data from low-IS*6110*-copy-number strains isolated from other geographical areas, in order to better define the

evolution of low-IS*6110*-copy-number strains and the impact of such evolution on the interpretation of molecular epidemiological data.

Materials and methods

Study Setting

Between January 1992 and December 1998, *Mycobacterium tuberculosis* isolates were obtained from patients resident in two suburbs in Cape Town, South Africa (3) as well as a subset of patients resident in the adjoining suburbs. The two suburbs have a population of 38500 residents within an area of 3.4 km^2 and have two healthcare clinics. In this setting the average annual incidence of new bacteriologically confirmed tuberculosis cases (culture and/or smear-positive) was 313/100,000 (38).

DNA fingerprinting

Genomic DNA from each isolate of *M. tuberculosis* was digested with either *Pvu*II or *Hin*fI, electrophoretically fractionated and Southern transferred to Hybond N+ (Amersham, United Kingdom). The blots containing the *Pvu*II-digested DNA were sequentially hybridized with ECL-labelled probes complementary to the 3' domain of the IS*6110* element (IS-3') (35), the 5' domain of the IS*6110* element (IS-5') (42), direct repeat (DR) (42) and Marker X (Roche, Germany). Each probe was stripped by denaturation before the next probe was applied. The *Hin*fI Southern blots were hybridized with the ^{32}P-labeled MTB484(1) probe complementary to the polymorphic G+C-rich repeat sequences (PGRS) (41). The autoradiographs were normalized and the

IS-3', IS-5' and DR bands were assigned using GelCompar II software. Cluster analysis was done using the UPGMA (unweighted pair group method with arithmetic mean) and Dice coefficient (17). Mutations in the IS*6110*-flanking domains were determined as previously described (43). The band corresponding to the IS*6110* insertion in the DR region was identified by aligning the DR and IS-3' autoradiographs (42). The blots probed by MTB484(1) were visually analyzed by two independent persons (42).

PCR amplification was used to determine the presence or absence of an IS*6110* insertion in the genes Rv0403c, Rv1758, or Rv3018c according to the previously described method (6). IS*6110* insertion in the gene Rv2787c was determined using the primer set 5'-TTCAACCATCGCCGCCTCTAC-3' and 5'-GGCCAAATCCAGCACGGTGAAC-3'.

Mutation analysis

The *M. tuberculosis* isolates were classified into three principle genetic groups according to polymorphisms in the *katG* and *gyrA* genes (31), using the dot blot hybridization method (39).

SpoligotypingDNA polymorphism in the DR locus was detected in isolates with ≤ 6 IS*6110* insertion elements by spoligotyping according to a standardized protocol (19).

MIRU-VNTR typing

M. tuberculosis isolates were genotyped by PCR amplification of the 12 loci containing VNTRs of elements called Mycobacterial Interspersed Repetitive Units (MIRUs) (33) and 9 loci containing VNTRs of other interspersed sequences (14,21,24,Supply et al., in preparation) using both manual (33) and automated techniques (32). The primers against the MIRU-VNTR flanking regions were the same as previously described (33), except that hex labelling was replaced by Vic labelling. The primers against the other loci are described in Table 1. The samples were subjected to electrophoresis using a 96-well ABI 377 automated sequencer as previously described (32). Sizing of the PCR fragments and assignment of the various VNTR alleles were done using the GeneScan and Genotyper software packages (PE Applied Biosystem), as previously described (32) and based on the data described in Table 1. Tables used for VNTR allele scoring are available at http://www.ibl.fr/mirus/mirus.html. Allele assignments were identical between the manual and automated methods.

Table 1: Conditions for multiplex PCRs of nine VNTR loci

Multiplex	Conventional designation[a]	VNTR length (bp)	$MgCl_2$ (mM)	PCR primer pairs (5' to 3', with labeling indicated)
Mix E	VNTR 2347	57	1.5	GCCAGCCGCCGTGCATAAACCT (FAM) AGCCACCCGGTGTGCCTTGTATGAC
	VNTR 2461	57		ATGGCCACCCGATACCGCTTCAGT (VIC) CGACGGGCCATCTTGGATCAGCTAC
	VNTR 3171	54		GGTGCGCACCTGCTCCAGATAA (NED) GGCTCTCATTGCTGGAGGGTTGTAC
Mix F	VNTR 0424	51	1.5	CTTGGCCGGCATCAAGCGCATTATT GGCAGCAGAGCCCGGGATTCTTC (FAM)
	VNTR 0577	58		CGAGAGTGGCAGTGGCGGTTATCT (VIC) AATGACTTGAACGCGCAAATTGTGA
	VNTR 1895	57		GTGAGCAGGCCCAGCAGACT (NED) CCACGAAATGTTCAAACACCTCAAT
Mix G	VNTR 2401	58	3.0	CTTGAAGCCCCGGTCTCATCTGT (FAM) ACTTGAACCCCCACGCCCATTAGTA
	VNTR 3690	58		CGGTGGAGGCGATGAACGTCTTC (VIC) TAGAGCGGCACGGGGGAAAGCTTAG
	VNTR 4156	59		TGACCACGGATTGCTCTAGT GCCGGCGTCCATGTT (NED)

[a]VNTR 0577, 2461, 4156, 1895 are described in (28), and correspond to ETRC, ETRB, QUB 4156 and 1895, respectively (14,25).

Global dissemination

To determine the geographical spread of the low-IS*6110*-copy-number strains, *M. tuberculosis* isolates collected from Southern (Western Cape) (n = 47), Central (Free State) (n = 4) and Northern (Gauteng and Mpumalanga) (n = 4) regions in South Africa and from Harare and Gweru (n = 11) in Zimbabwe were subjected to spoligotyping, IS*6110*-RFLP and IS*6110* insertion mapping (6). This genotypic data was compared to previously published genotypic data (IS*6110* banding patterns (visual comparison),

IS6*110* insertion points and spoligotype patterns) on low-IS*6110*-copy-number strains from Europe (United Kingdom (n = 14) (7), Denmark (n = 2) (2)) and America (Michigan (n = 70) (6) and CDC1551 (12)).

Genetic relationship analysis

The evolutionary state(s) for the RFLP data were assigned according to the presence (indicated by "1") or the absence (indicated by "0") of a hybridizing band. Spoligotypes were assigned according to the presence or absence of spacer sequences, while the VNTR alleles were assigned according to the number of repeats present in the different loci. The complete set of evolutionary states for the different markers were subjected to phylogenetic analysis using the neighbor joining algorithm (PAUP 4.0*; Phylogenetic Analysis Using Parsimony (*Other Methods) Version 4b10. Sinauer Associates, Sunderland, Massachusetts). Bootstrapping was performed to establish a degree of statistical support for nodes within each phylogenetic reconstruction (10). A consensus tree was generated using the program contree (PAUP 4.0*) in combination with the majority rule formula. The resulting trees were rooted to the principal genetic group 1 isolate (SA CT(67)). Only branches which occurred in > 50 % of the bootstrap trees were included in the final tree and all branches with a zero branch length were collapsed.

Results

Selection and molecular characterization of strains

Between January 1992 and December 1998, *Mycobacterium tuberculosis* isolates were obtained from 1030 patients resident in the adjoining suburbs in Cape Town, South Africa. IS*6110*-RFLP analysis established that 186 (18.1 %) of these patients were infected with a strain containing ≤ 6 IS*6110* hybridizing bands. No isolate lacking the IS*6110* element was identified in this study setting. Isolates were available from 170 (91 %) of these patients for further genotypic analysis.

Analysis of the *katG* and *gyrA* gene sequences classified 169 of these isolates in principal genetic group 2 , while only one isolate was classified as principal genetic group 1 (31). All isolates were then subjected to further analysis using Southern hybridization in combination with probes complementary to the 3' domain of the IS*6110* element (IS-3') (35) (Figure 1), 5' domain of the IS*6110* element (IS-5') (42), spoligotyping (19) (Table 2) and VNTR allele typing based on 21 independent loci (14,21,25,32,33) and PGRS-RFLP typing (41) (Table 3). This set included (alias designation in brackets) MIRU-VNTR loci 2, 4 (ETR-D), 10, 16, 20, 23, 24, 26, 27, 31 (ETR-E), 39 and 40, and VNTR loci 424, 577 (ETR-C), 1895 (QUB-1895), 2347, 2401, 2461(ETR-B), 3171, 3690, 4156 (QUB-4156).

Table 2: Spoligotype patterns of Cape Town *M. tuberculosis* isolates with ≤6 IS*6110* elements

Spoligotype type[a]	Octal code[b]	Spoligotype pattern[b]
1	777777777760771	■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□□□■■■■■■■
2	777776777760601	■■■■■■■■■■■■■■■■■□■■■■■■■■■■■■■■□□□□■■□□□□■
3	777736777760601	■■■■■■■■■■■■□■■■■□■■■■■■■■■■■■■□□□□■■□□□□■
4	677776777760601	■■□■■■■■■■■■■■■■□■■■■■■■■■■■■■□□□□■■□□□□■
5	777776777760771	■■■■■■■■■■■■■■■■■□■■■■■■■■■■■■■□□□□■■■■■■■
6	777776777720601	■■■■■■■■■■■■■■■■■□■■■■■■■■■■■□■□□□□■■□□□□■
7	737776777760601	■■■□■■■■■■■■■■■■■□■■■■■■■■■■■■■□□□□■■□□□□■
8	767776777760601	■■■■■□■■■■■■■■■■■□■■■■■■■■■■■■■□□□□■■□□□□■
9	777776777760711	■■■■■■■■■■■■■■■■■□■■■■■■■■■■■■■□□□□■■■□□■■
10	743776777760601	■■■■□□□■■■■■■■■■■□■■■■■■■■■■■■■□□□□■■□□□□■
11	777776777760731	■■■■■■■■■■■■■■■■■□■■■■■■■■■■■■■□□□□■■■□■■■
12	700076777760771	■■■□□□□□□□□□■■■■■□■■■■■■■■■■■■■□□□□■■■■■■■
13	617776777760771	■■□□□■■■■■■■■■■■□■■■■■■■■■■■■■□□□□■■■■■■■
14	777776777760740	■■■■■■■■■■■■■■■■■□■■■■■■■■■■■■■□□□□■■■■□□□
15	700076774360771	■■■□□□□□□□□□■■■■■□■■■■■■■□□□■■■■□□□□■■■■■■■
16	700076777740371	■■■□□□□□□□□□■■■■■□■■■■■■■■■■■■■□□□□□□■■■■■■
17	700076777760671	■■■□□□□□□□□□■■■■■□■■■■■■■■■■■■■□□□□■■□■■■■
18	777776777560771	■■■■■■■■■■■■■■■■■□■■■■■■■■■□■■■□□□□■■■■■■■
19	777776617560771	■■■■■■■■■■■■■■■■■□■■□□□■■■■■□■■■□□□□■■■■■■■
20	757777777413731	■■■■□■■■■■■■■■■■■■■■■■■■■□□□□■□■■■■■□■■■

[a]Spoligotype type, arbitrary designation (see Table 3).
[b]Octal Code according to (8).
[c]Closed boxes infer the presence of DVR sequences, while open boxes infer the absence of DVR sequence.

In addition, IS*6110* insertion into the genes Rv0403c, Rv1758, Rv3018c and Rv2787c was determined by PCR amplification (Figure 1A). Table 3 summarizes the genotypic data of the different isolates as defined by the combined markers. For this data set, the number of principal genetic group 2 genotypes obtained with the different methods was ranked as follows: IS-3' (14 genotypes) < IS-5' (19 genotypes) < spoligotypes (20 genotypes) < IS-3' & IS-5' (22 genotypes) < VNTR loci (38 genotypes) < PGRS (45 genotypes). In combination, these different genotyping methods identified a total of 66 distinct principal genetic group 2 genotypes (Table 3).

Table 3: Genotypic classification of Cape Town *M. tuberculosis* isolates with ≤ 6 IS*6110* elements

Name	Principal genetic group	IS*6110* Copy	IS-3' type	IS-5' type	PGRS type	12 MIRU-VNTR type (allele combinations[a])	9 VNTR (allele combinations[b])	Spoligotype type[c] (Octal format[c])
SA CT(1)	2	1	1	1	1	1 (225125113322)	(144442353)	1 (777777777760771)
SA CT(2)	2	2	2	2	2	2 (223325153323)	(142442383)	2 (777776777760601)
SA CT(3)	2	2	2	2	3	3 (224325123422)	(242442343)	3 (777736777760601)
SA CT(4)	2	2	2	2	3	3 (224325123422)	(242442343)	2 (777776777760601)
SA CT(5)	2	2	2	2	4	4 (224325143223)	Nd	3 (777736777760601)
SA CT(6)	2	2	2	2	4	4 (224325143223)	(442442333)	2 (777776777760601)
SA CT(7)	2	2	2	2	5	5 (224325143324)	(242442343)	4 (677776777760601)
SA CT(8)	2	2	2	2	5	5 (224325143324)	(242442343)	3 (777736777760601)
SA CT(9)	2	2	2	2	5	5 (224325143324)	(242442343)	2 (777776777760601)
SA CT(10)	2	2	2	2	6	6 (224325153223)	(442442431)	2 (777776777760601)
SA CT(11 - 12)	2	2	2	2	2, 7	7 (224325153323)	(142442383)	2 (777776777760601)
SA CT(13)	2	2	2	2	2	7 (224325153323)	(142442383)	5 (777776777760771)
SA CT(14)	2	3	3	3[d]	8	8 (223325143323)	(2424422?3)	6 (777776777720601)
SA CT(15)	2	3	3	4	9	9 (224325133324)	(442442333)	7 (737776777760601)
SA CT(16)	2	3	3	4	9	10 (224325153323)	(442442333)	7 (737776777760601)
SA CT(17)	2	3	4	5	10	11 (223325163322)	(242442333)	8 (767776777760601)
SA CT(18)	2	3	4	5	10	12 (224325163322)	(242442333)	8 (767776777760601)
SA CT(19)	2	3	5	6	11	13 (224325153222)	(522442332)	9 (777776777760711)
SA CT(20)	2	3	5	7[d]	12	14 (224325153324)	(252441343)	10 (743776777760601)
SA CT(21)	2	3	6[d]	5	13	15 (223325153223)	(442442333)	2 (777776777760601)
SA CT(22)	2	3	6[d]	5	14	10 (224325153323)	(442442333)	2 (777776777760601)
SA CT(23)	2	3	6[d]	5	14	16 (224325153323)	(452442333)	2 (777776777760601)
SA CT(24)	2	3	7	8	15	17 (223325163433)	Nd	2 (777776777760601)
SA CT(25)	2	4	8	9	16	18 (223325143322)	(432442333)	5 (777776777760771)
SA CT(26 - 30)	2	4	8	9	16, 17, 18, 19, 20	19 (223325143324)	(432442333)	5 (777776777760771)
SA CT(31)	2	4	8	9	19	20 (223325143325)	(432442333)	5 (777776777760771)
SA CT(32)	2	4	8	9	17	21 (223325153324)	(432442333)	11 (777776777760731)
SA CT(33)	2	4	8	9	21	19 (223325143324)	Nd	5 (777776777760771)
SA CT(34)	2	4	8	9	22	21 (223325153324)	(432442333)	5 (777776777760771)
SA CT(35 - 36)	2	4	8	10	23, 24	22 (223325153224)	(332442343)	12 (700076777760771)
SA CT(37)	2	4	8	10	23	23 (224325153322)	(432442343)	12 (700076777760771)
SA CT(38 - 44)	2	4	8	10	23, 25, 26, 27, 28, 29, 30	24 (224325153324)	(432442343)	12 (700076777760771)
SA CT(45)	2	4	8	10	31	24 (224325153324)	Nd	5 (777776777760771)
SA CT(46)	2	4	8	11	32	25 (224325153324)	Nd	12 (700076777760771)
SA CT(47)	2	4	8	11	33	25 (224325153324)	(332442331)	13 (617776777760771)
SA CT(48)	2	4	9	12[d]	34	26 (224325153324)	(432442333)	14 (777776777760740)
SA CT(49 - 50)	2	4	9	9	34, 35	21 (223325153324)	(432442333)	14 (777776777760740)
SA CT(51 - 52)	2	4	9	9	34, 35	27 (224325133324)	(432442333)	14 (777776777760740)
SA CT(53)	2	4	9	9	34	28 (226325153324)	(432442333)	15 (777776777760740)
SA CT(54)	2	4	9	10	36	24 (224325153324)	(432442343)	15 (700076774360771)
SA CT(55)	2	4	9	10	37	29 (224325153325)	(432442343)	16 (700076777740371)
SA CT(56)	2	4	10	13	38	30 (224325153434)	Nd	2 (777776777760601)
SA CT(57)	2	4	11	14	39	31 (224225164434)	Nd	17 (700076777760671)
SA CT(58)	2	5	12	15[d]	40	32 (224325143324)	(432442334)	5 (777776777760771)
SA CT(59)	2	5	12	16	41	33 (224325153322)	(432442333)	18 (777776777560771)
SA CT(60)	2	5	12	16	42	33 (224325153322)	(432442333)	16 (777776777760771)
SA CT(61)	2	5	12	16	41	34 (234325133323)	(432442333)	18 (777776777560771)
SA CT(62)	2	5	12	16	41	35 (234325153323)	Nd	5 (777776777760771)
SA CT(63)	2	5	12	17	41	33 (224325153322)	(432442333)	19 (777776617560771)
SA CT(64)	2	5	13	18	43	36 (224325153222)	(532442433)	12 (700076777760771)
SA CT(65)	2	5	13	18	44	37 (224325153222)	(532442333)	12 (700076777760771)
SA CT(66)	2	6	14	19	45	38 (224325164335)	Nd	12 (700076777760771)
SA CT(67)	1	5	15	20	46	39 (254316?34613)	Nd	20 (757777777413731)

[a]MIRU-VNTR loci according to (33).
[b]VNTR loci 424, 577, 1895, 2347, 2401, 2461, 3171, 3690, 4156 (see Table 1 and (14,24), Supply *et al*. in preparation).
[c]Spoligotype nomenclature according to (8).
[d]Mutations in the IS*6110*-flanking domains other than in the DR region (43).
[e]Spoligotype type: arbitrary designation to demonstrate genetic diversity.
Nd not determined. ? missing VNTR allele.

Figure 1

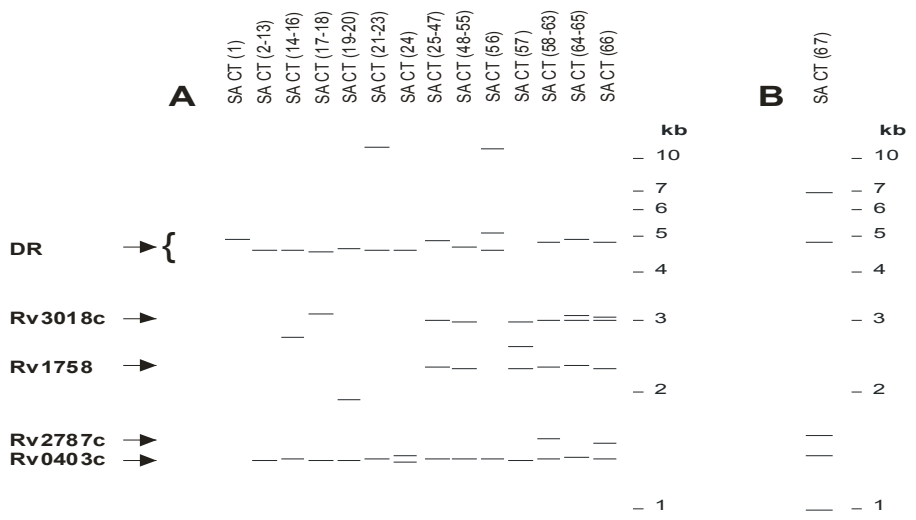


Figure 1. Southern blot analysis of Cape Town *M. tuberculosis* isolates with ≤ 6 IS*6110* insertions.
Genomic DNA was restricted with *Pvu*II, electrophoretically, fractionated on agarose gels and Southern hybridization was done with ECL-labeled IS-3' probe.
A) IS*6110* banding pattern of distinct principal genetic group 2 isolates from Cape Town, South Africa. Lanes are labeled according to the isolate names given in Table 3. The band representing the IS*6110* insertion in the DR repeat region and bands representing IS*6110* insertions in genes Rv0403c, Rv1758, Rv2787c, and Rv3018c are indicated.
B) IS*6110* banding pattern of the principal genetic group 1 isolate from Cape Town, South Africa. Lane labeled according to the isolate name given in Table 3. Co-hybridization between the IS*6110* and DR probes could not be demonstrated due to the presence of a *Pvu*II site situated between the 3' domain of the IS*6110* element and the flanking DR region

Clonal expansion of strains

Combined analysis of the above different molecular markers strongly supported the close genetic relatedness of the principal genetic group 2 isolates and their clear distinctiveness from the single isolate from principal genetic group 1. Six out of 21 VNTR loci (MIRU-

VNTR loci: 2, 20, 23, 24, and VNTR 2347 and 2401) were fully identical within principal genetic group 2 isolates analysed (Table 3), while 7 other loci (MIRU-VNTR locus 4, 16, 27, 39, VNTR 1895, 2461, 3171) displayed at most 3 variations compared to the predominant alleles among these isolates. PCR amplification of the Rv0403c region (Figure 1A), showed that only the principal genetic group 2 variants with ≥ 2 and ≤ 6 IS*6110* elements shared an IS*6110* in the same position, suggesting that these isolates were derived from a common ancestor. All principal genetic group 2 spoligotypes showed a deletion of direct variable repeat (DVR) sequences 33 to 36 (29), while isolates with ≥ 2 and ≤ 6 IS*6110* insertions also all showed an additional DVR 18 deletion (Table 2). DVR 34 was deleted in the principal genetic group 1 isolate (Table 2).

Phylogenetic analysis of the principle genetic group 2 isolates was done based on the whole set of markers, using the principle genetic group 1 member as an outgroup (SA CT(67)) (Figure 2). The overall branching order of the tree suggests that the principal genetic group 2 isolates evolved from a common progenitor by sequential replicative transposition of IS*6110* followed, in certain cases, by mutation in the regions flanking the IS*6110* elements (Table 3). According to this tree, these clones first evolved by replicative transposition of the IS*6110* element into Rv0403c (Figure 1A) and by the deletion of DVR 18 to generate a strain with two IS*6110* insertions (Figure 2). These two genotypic characteristics, along with the six conserved VNTR loci, were subsequently inherited in all the progeny. In different branches of this progeny, subsequent IS*6110* insertions were identified in Rv1758, Rv3018c and Rv2787c, to generate clonal variants with between three and six IS*6110* insertions (Figure 2). Along with these events

occurred the deletion of DVR's 39 to 42 and DVR's 4 to 12 (Table 2 and Figure 2). Broadly similar pictures of stepwise acquisition of IS*6110* were obtained by phylogenetic analysis using either IS*6110*-RFLP, spoligotyping and PGRS or VNTR genotypes alone (data not shown).

Figure 2

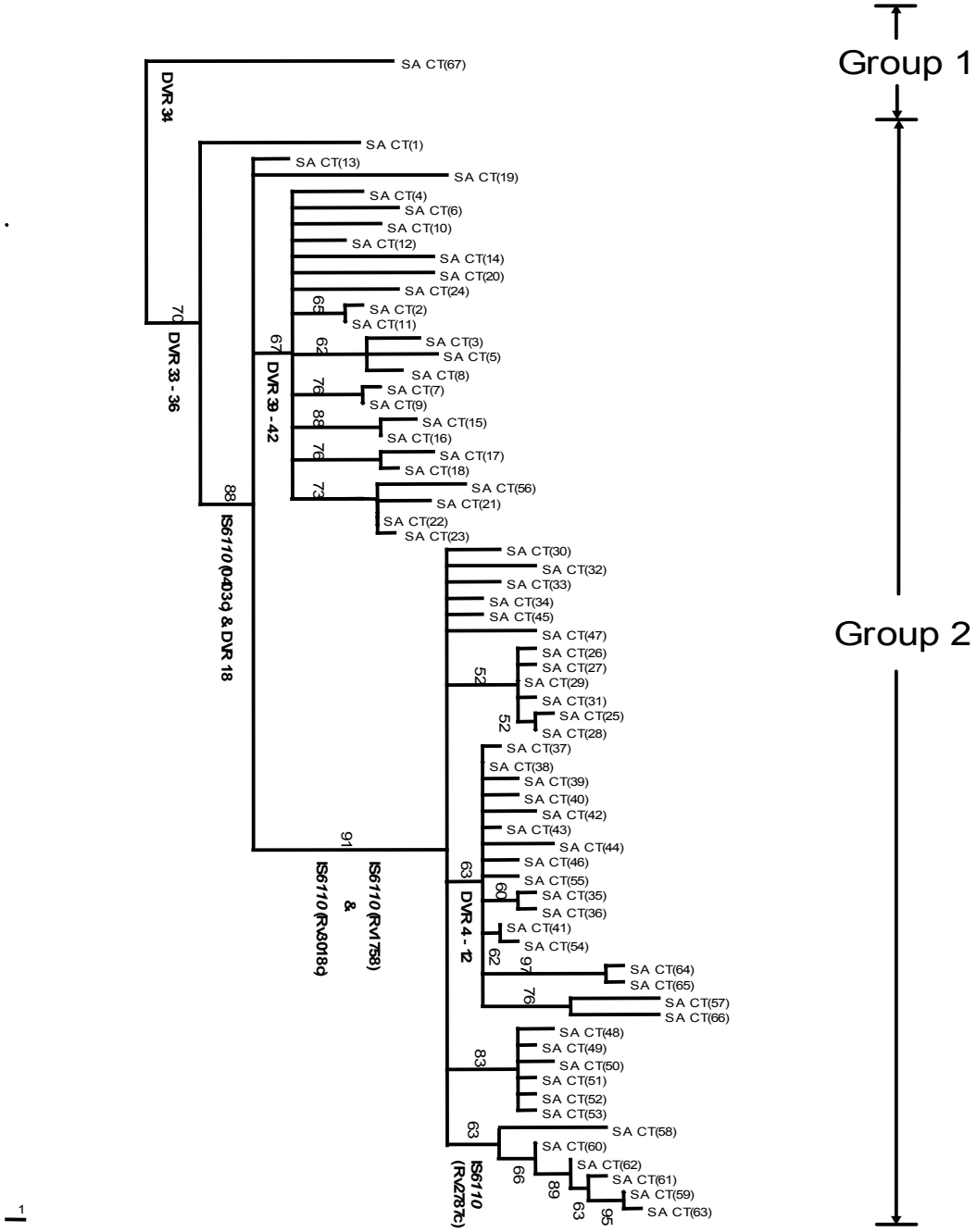


Figure 2. Phylogenetic tree of Cape Town *M. tuberculosis* isolates with ≤ 6 IS*6110* insertions. Genetic data from five different genotyping methods were subjected to phylogenetic analysis using the bootstrapping and neighbor joining algorithm In methods. The tree was rooted to the principal genetic group 1 clone (isolate SA CT(67)). Boostrap values are given at internal nodes. IS*6110* insertion in defined genes as well as DVR deletions are indicated at the nodes where they occurred. All branches with a zero length were collapsed. Isolate names are the same as in Table 3. Principle genetic groups 1 and 2 were assigned according to polymorphisms in the *katG* and *gyrA* genes (31). The scale indicates the number of steps per unit length

Geographical distribution of strains

To determine the evolutionary relationships between the low-IS*6110*-copy-number strains in different geographical regions, the genotypic data from a representative set of the Cape Town isolates (n = 16) was compared with those of a representative set of low-IS*6110*-copy-number isolates from other regions of Southern Africa (n= 17) (this study), Europe (London UK (n = 8) (7) and Denmark (n = 2) (2)), and America (Michigan (n = 47) (6) and the CDC1551 reference strain (6)). Phylogenetic analysis using these data indicated a close evolutionary relationship between the principal genetic group 2 isolates from these different geographical regions (Figure 3). Many of the isolates between the different settings shared identical IS-3' banding patterns, IS*6110* insertion points and spoligotypes, and only in rare instances were IS*6110* transposition and DVR deletion events found to be unique to a specific geographic region (Figure 3, IS*6110* insertion in Rv2787c of SA isolates). These observations suggest that these genotype properties were acquired in a distant past and prior to the global dissemination of the lineage. In contrast, comparison of MIRU-VNTR genotypes based on 12 loci, in common with those previously reported (6), failed to identify clones with identical IS-3', spoligotype and MIRU-VNTR genotypes between Cape Town and Michigan, indicating that the MIRU-VNTR loci are evolving more rapidly. A similar study could not be done for the PGRS genotypes as this methodology has not been internationally standardized.

Figure 3

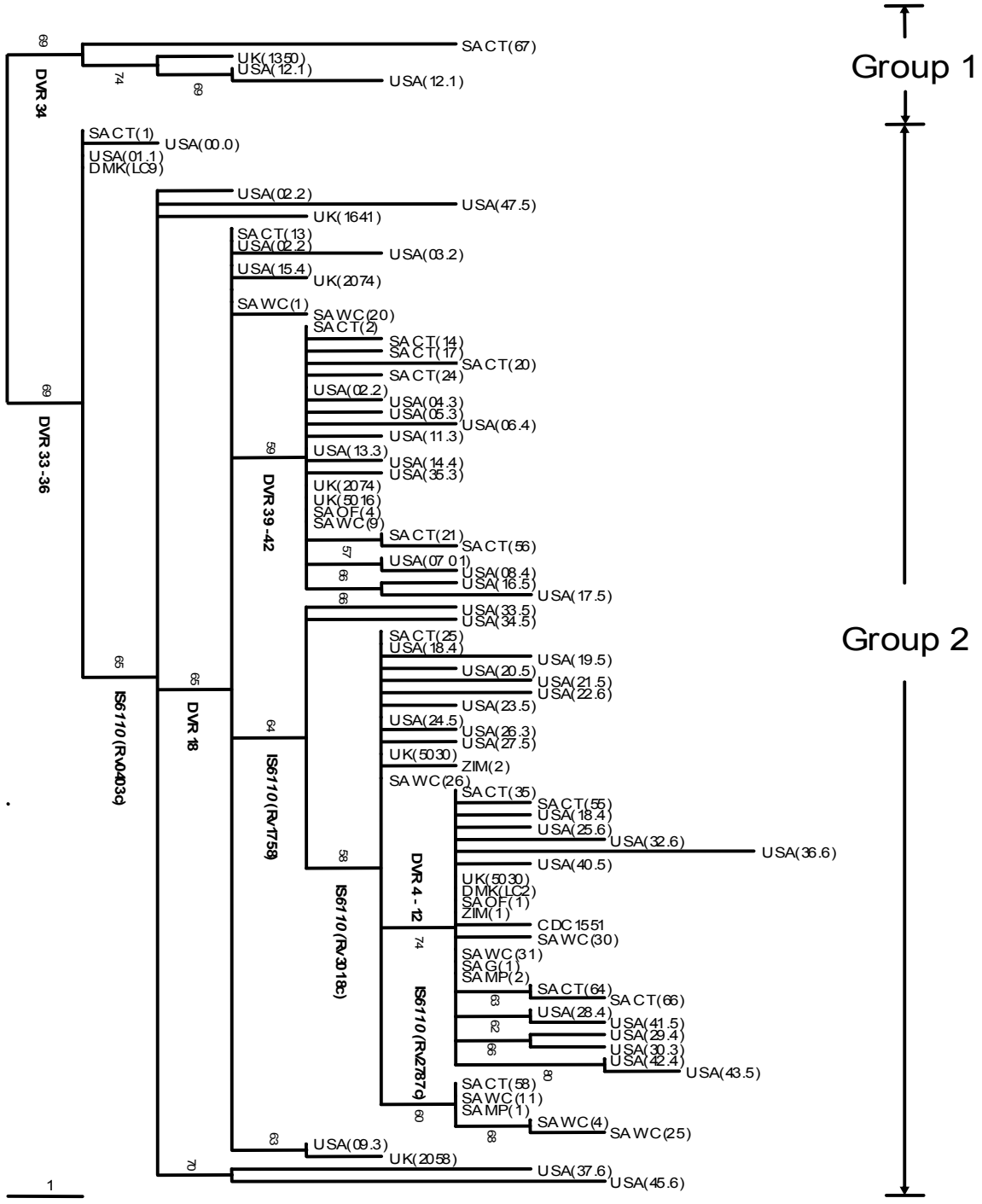


Figure 3. Phylogenetic tree of *M. tuberculosis* isolates with ≤ 6 IS*6110* insertions obtained from different geographical regions. Representative isolates (n = 91) cultured in Cape Town, Southern Africa, Europe (2,7), and the United States of America (6,12) were included. Genetic data from IS*6110* banding patterns, IS*6110* insertion points and spoligotyping were subjected to phylogenetic analysis using the bootstrapping and neighbor joining algorithm The tree was rooted to the principal genetic group 1 isolate (SACT(67)). Bootstrap values are given at internal nodes. IS*6110* insertion in defined genes as well as DVR deletions are indicated at the nodes where they occurred. All branches with a zero length were collapsed. The Cape Town isolate names are the same as in Table 3, while the isolates from other regions of Southern Africa are labeled as follows: SAWC (South Africa Western Cape), SAOF (South Africa Free Sate), SAG (South Africa Gauteng), SAMP (South Africa Mpumalanga) and ZIM (Zimbabwe). The isolates from Europe are labeled according to (7) and (2), while the isolates from the America are labeled according to (6) and (12) Principal genetic groups 1 and 2 were assigned according to polymorphisms in the *katG* and *gyrA* genes (31). The scale indicates the number of steps per unit length.

Discussion

This study provides evidence that nearly all *M. tuberculosis* isolates with ≤ 6 IS*6110* elements, collected in Cape Town, South Africa, are members of a lineage of the principal genetic group 2. Evidence for this is based on the inheritance of defined polymorphisms, which include; 1) principal genetic group 2 classification according to mutations in the *katG* and *gyrA* genes (31), and concordant deletion of DVR's 33 to 36, known to be specific to principal genetic group 2 and 3 strains (29); 2) the identification of a conserved IS*6110* insertion in Rv0403c (13) and the deletion of DVR 18 from the direct repeat region in principal genetic group 2 strains with ≥ 2 and ≤ 6 IS*6110* insertions; 3) the presence of six fully conserved VNTR loci. Only one isolate from this setting was identified as being a member of a distinct low-IS*6110*-copy-number lineage of the principal genetic group 1, which corresponds to Cluster I (15) or Group I (7) and has been primarily associated with patients from East-Africa and Asia (6,30).

The principal genetic group 2 lineage studied here encompasses the groups referred to as Group II and III (7) or clade X (27), and clusters IV and V defined by sSNP analysis (15). Our phylogenetic analysis based on fully independent markers in isolates from different geographical areas, supports the notion that strains in this principal genetic group 2 lineage evolved from a common progenitor containing a single IS*6110* element by sequential acquisition of up to five additional IS*6110* copies, as well as by expansion and contraction of VNTR sequences and the deletion of specific DVRs. Sequential acquisition of additional IS*6110* copies is consistent with the direct evolutionary relationship between the sSNP clusters IV and V, which include strains with 1-3 and 4-6

IS*6110* copies, respectively (15). Such congruence between phylogenies inferred from independent sets of markers (within our study or between our study and that of Guttacker *et al*., 2003) provides strong evidence for the robustness of the inferred phylogeny. Moreover, the deletion of DVRs suggested by our phylogenetical analysis (Figure 2 and 3) is consistent with previous findings supporting that evolution of the DR region is driven by loss of DVR sequences rather than by their duplication (9,36,44).

Interestingly, each IS*6110* transposition event appeared to occur only once within the phylogenetic tree, suggesting divergent evolution. This is in sharp contrast to previous suggestion that the IS*6110* banding patterns of low-IS*6110*-copy-number strains could have evolved convergently due to the presence of preferential IS*6110* integration sites (13). The limited number of IS*6110* variants identified may suggest that IS*6110* transposition is regulated in this lineage, raising the hypothesis of lineage-specific effects. Regulation in the number of transposable elements, referred to as taming, has been described in eukaryotic genomes, and might be a specific mechanism against mutagenic effects induced by these elements (18).

The preservation of certain IS-3' banding patterns and spoligotypes in isolates from Cape Town, from other Southern African regions, from Europe and from America suggests that these markers have remained stable over a long period of time. Therefore, we hypothesize that these genotypes represent distantly evolved clones that have become globally disseminated. The examination of the SpolD database (11) indicates that principal genetic group 2 isolates with the characteristic DVR 33 to 36 and DVR 18

deletions have been isolated in 27 different countries. By comparison, principal genetic group 1 clones with the characteristic DVR 34 deletion have been isolated in 26 countries, with a high prevalence in South Asia. Taken together, these findings suggest that in addition to other well-identified lineages like W-Beijing (4), the principal genetic group 1 and 2 low-IS*6110*-copy-number lineages also play an important role in the global tuberculosis epidemic.

The inferred stability of the IS-3' banding patterns and of certain spoligotypes for extended periods of time is likely too high to be informative for tracking ongoing transmission between patients in settings where this lineage is predominant. Conversely, comparison of our genotype data with those of Cowan *et al.* (6) failed to identify strains between Cape Town and Michigan in which the IS-3' banding pattern, spoligotype and MIRU-VNTR types are identical. Given the stability of MIRU-VNTR genotypes in epidemiologically-linked isolates (16,20,23,26) the absence of MIRU-VNTR matching between the two studies is in accordance with the above argument of distant relationships between shared IS-3' types and spoligotypes. Moreover, it supports VNTR typing as a useful tool for epidemiologic tracking across various epidemiological settings and bacterial populations. This is consistent with the contention that, as a multi-locus based method, VNTR typing is much less exposed to biases inherent to single loci or to copy numbers of a single genetic element, such as spoligotyping and IS*6110*-based typing, respectively.

Our study represents a step towards a better understanding of the evolutionary mechanisms modelling the genome in different *M. tuberculosis* lineages and of the different rates at which these events occur. This will provide new insights for the interpretation of molecular epidemiological data and enhance our understanding of how different strains contribute to the tuberculosis epidemic in specific regions and at a global scale.

Acknowledgments

This study was made possible by grants from the GlaxoSmithKline Action TB Initiative, IAEA (projects SAF6/003 and CRP 9925), Harry Crossely Foundation and the National Research Foundation (project 2054201), and from the Ministère de de la Recherche and Ministère des Affaires Etrangères, and the South African Medical Research Council (MRC). The work was also supported by Institut National de la Santé et de la Recherche Médicale (INSERM), Institut Pasteur de Lille, Région Nord-Pas-de-Calais. Ms VN Chihota was supported by Wellcome Trust Research Leave Fellowship Grant (01680/2/00/2 to V N. Chihota). Mrs E. Engelke, Mrs S. Carlini Mrs M. De Kock, and Frederique De Matos are thanked for their technical assistance P. S. is a Chercheur du Centre National de Recherche Scientifique. We thank Dr S. Charalambous for the provision of clinical isolates collected in the Free State, South Africa.

Reference list

1. **Alland, D., T. S. Whittam, M. B. Murray, M. D. Cave, M. H. Hazbon, K. Dix, M. Kokoris, A. Duesterhoeft, J. A. Eisen, C. M. Fraser, and R. D. Fleischmann**. 2003. Modeling bacterial evolution with comparative-genome-based marker systems: application to *Mycobacterium tuberculosis* evolution and pathogenesis. J.Bacteriol. **185**:3392-3399.

2. **Bauer, J., A. B. Andersen, K. Kremer, and H. Miorner**. 1999. Usefulness of spoligotyping To discriminate IS*6110* low-copy-number *Mycobacterium tuberculosis* complex strains cultured in Denmark. J.Clin.Microbiol. **37**:2602-2606.

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

5. **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.

6. **Cowan, L. S., L. Mosher, L. Diem, J. P. Massey, and J. T. Crawford**. 2002. Variable-number tandem repeat typing of *Mycobacterium tuberculosis* isolates with low copy numbers of IS*6110* by using mycobacterial interspersed repetitive units. J.Clin.Microbiol. **40**:1592-1602.

7. **Dale, J. W., H. Al Ghusein, S. Al Hashmi, P. Butcher, A. L. Dickens, F. Drobniewski, K. J. Forbes, S. H. Gillespie, D. Lamprecht, T. D. McHugh, R. Pitman, N. Rastogi, A. T. Smith, C. Sola, and H. Yesilkaya**. 2003. Evolutionary relationships among strains of *Mycobacterium tuberculosis* with few copies of IS*6110*. J.Bacteriol. **185**:2555-2562.

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

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. **Felsenstein, J.** 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution **39**:783-793.

11. **Filliol, I., J. R. Driscoll, D. van Soolingen, B. N. Kreiswirth, K. Kremer, G. Valetudie, D. A. Dang, R. Barlow, D. Banerjee, P. J. Bifani, K. Brudey, A. Cataldi, R. C. Cooksey, D. V. Cousins, J. W. Dale, O. A. Dellagostin, F. Drobniewski, G. Engelmann, S. Ferdinand, D. Gascoyne-Binzi, M. Gordon, M. C. Gutierrez, W. H. Haas, H. Heersma, E. Kassa-Kelembho, M. L. Ho, A. Makristathis, C. Mammina, G. Martin, P. Mostrom, I. Mokrousov, V. Narbonne, O. Narvskaya, A. Nastasi, S. N. Niobe-Eyangoh, J. W. Pape, V. Rasolofo-Razanamparany, M. Ridell, M. L. Rossetti, F. Stauffer, P. N. Suffys, H. Takiff, J. Texier-Maugein, V. Vincent, J. H. de Waard, C. Sola, and N. Rastogi**. 2003. Snapshot of moving and expanding clones of *Mycobacterium tuberculosis* and their global distribution assessed by spoligotyping in an international study. J.Clin.Microbiol. **41**:1963-1970.

12. **Fleischmann, R. D., D. Alland, J. A. Eisen, L. Carpenter, O. White, J. Peterson, R. DeBoy, R. Dodson, M. Gwinn, D. Haft, E. Hickey, J. F. Kolonay, W. C. Nelson, L. A. Umayam, M. Ermolaeva, S. L. Salzberg, A. Delcher, T. Utterback, J. Weidman, H. Khouri, J. Gill, A. Mikula, W. Bishai, J. W. Jacobs, Jr., J. C. Venter, and C. M. Fraser**. 2002. Whole-genome comparison of *Mycobacterium tuberculosis* clinical and laboratory strains. J.Bacteriol. **184**:5479-5490.

13. **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.

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

15. **Gutacker, M. M., J. C. Smoot, C. A. Migliaccio, S. M. Ricklefs, S. Hua, D. V. Cousins, E. A. Graviss, E. Shashkina, B. N. Kreiswirth, and J. M. Musser**. 2002. Genome-wide analysis of synonymous single nucleotide polymorphisms in *Mycobacterium tuberculosis* complex organisms. Resolution of genetic relationships among closely related microbial strains. Genetics **162**:1533-1543.

16. **Hawkey, P. M., E. G. Smith, J. T. Evans, P. Monk, G. Bryan, H. H. Mohamed, M. Bardhan, and R. N. Pugh**. 2003. Mycobacterial interspersed repetitive unit typing of *Mycobacterium tuberculosis* compared to IS*6110*-based restriction fragment length polymorphism analysis for investigation of apparently clustered cases of tuberculosis. J.Clin.Microbiol. **41**:3514-3520.

17. **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.

18. **Jensen, S., M. P. Gassama, and T. Heidmann**. 1999. Taming of transposable elements by homology-dependent gene silencing. Nat.Genet. **21**:209-212.

19. **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.

20. **Kwara, A., R. Schiro, L. S. Cowan, N. E. Hyslop, M. F. Wiser, H. S. Roahen, P. Kissinger, L. Diem, and J. T. Crawford**. 2003. Evaluation of the epidemiologic utility of secondary typing methods for differentiation of *Mycobacterium tuberculosis* isolates. J.Clin.Microbiol. **41**:2683-2685.

21. **Le Fleche, P., M. Fabre, F. Denoeud, J. L. Koeck, and G. Vergnaud**. 2002. High resolution, on-line identification of strains from the *Mycobacterium tuberculosis* complex based on tandem repeat typing. BMC.Microbiol. **2**:37.

22. **Mathema, B., P. J. Bifani, J. Driscoll, L. Steinlein, N. Kurepina, S. L. Moghazeh, E. Shashkina, S. A. Marras, S. Campbell, B. Mangura, K. Shilkret, J. T. Crawford, R. Frothingham, and B. N. Kreiswirth**. 2002. Identification and evolution of an IS*6110* low-copy-number *Mycobacterium tuberculosis* cluster. J.Infect.Dis. **185**:641-649.

23. **Mazars, E., S. Lesjean, A. L. Banuls, M. Gilbert, V. Vincent, B. Gicquel, M. Tibayrenc, C. Locht, and P. Supply**. 2001. High-resolution minisatellite-based typing as a portable approach to global analysis of *Mycobacterium tuberculosis* molecular epidemiology. Proc.Natl.Acad.Sci.U.S.A **98**:1901-1906.

24. **Roring, S., D. Brittain, A. E. Bunschoten, M. S. Hughes, R. A. Skuce, J. D. van Embden, and S. D. Neill**. 1998. Spacer oligotyping of *Mycobacterium bovis* isolates compared to typing by restriction fragment length polymorphism using PGRS, DR and IS*6110* probes. Vet.Microbiol. **61**:111-120.

25. **Roring, S., A. Scott, D. Brittain, I. Walker, G. Hewinson, S. Neill, and R. Skuce**. 2002. Development of variable-number tandem repeat typing of *Mycobacterium bovis*: comparison of results with those obtained by using existing exact tandem repeats and spoligotyping. J.Clin.Microbiol. **40**:2126-2133.

26. **Savine, E., R. M. Warren, G. D. van der Spuy, N. Beyers, P. D. van Helden, C. Locht, and P. Supply**. 2002. Stability of variable-number tandem repeats of Mycobacterial Interspersed Repetitive Units from 12 Loci in Serial Isolates of *Mycobacterium tuberculosis*. J.Clin.Microbiol. **40**:4561-4566.

27. **Sebban, M., I. Mokrousov, N. Rastogi, and C. Sola**. 2002. A data-mining approach to spacer oligonucleotide typing of *Mycobacterium tuberculosis*. Bioinformatics. **18**:235-243.

28. **Smittipat, N. and P. Palittapongarnpim**. 2000. Identification of possible loci of variable number of tandem repeats in *Mycobacterium tuberculosis*. Tuber.Lung Dis. **80**:69-74.

29. **Soini, H., X. Pan, A. Amin, E. A. Graviss, A. Siddiqui, and J. M. Musser**. 2000. Characterization of *Mycobacterium tuberculosis* isolates from patients in Houston, Texas, by spoligotyping. J.Clin.Microbiol. **38**:669-676.

30. **Soini, H., X. Pan, L. Teeter, J. M. Musser, and E. A. Graviss**. 2001. Transmission dynamics and molecular characterization of *Mycobacterium tuberculosis* isolates with low copy numbers of IS*6110*. J.Clin.Microbiol. **39**:217-221.

31. **Sreevatsan, S., X. Pan, K. E. Stockbauer, N. D. Connell, B. N. Kreiswirth, T. S. Whittam, and J. M. Musser**. 1997. Restricted structural gene polymorphism in the *Mycobacterium tuberculosis* complex indicates evolutionarily recent global dissemination. Proc.Natl.Acad.Sci.U.S.A **94**:9869-9874.

32. **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.

33. **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.

34. **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.

35. **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.

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

37. **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.

38. **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.

39. **Victor, T. C., A. M. Jordaan, A. van Rie, G. D. van der Spuy, M. Richardson, P. D. van Helden, and R. Warren**. 1999. Detection of mutations in drug resistance genes of *Mycobacterium tuberculosis* by a dot-blot hybridization strategy. Tuber.Lung Dis. **79**:343-348.

40. **Victor, T. C., A. van Rie, A. M. Jordaan, M. Richardson, G. D. Der Spuy, N. Beyers, P. D. van Helden, and R. Warren**. 2001. Sequence polymorphism in the rrs gene of *Mycobacterium tuberculosis* is deeply rooted within an evolutionary clade and is not associated with streptomycin resistance. J.Clin.Microbiol. **39**:4184-4186.

41. **Warren, R., M. Richardson, S. Sampson, J. H. Hauman, N. Beyers, P. R. Donald, and P. D. van Helden**. 1996. Genotyping of *Mycobacterium tuberculosis* with additional markers enhances accuracy in epidemiological studies. J.Clin.Microbiol. **34**:2219-2224.

42. **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. **81**:291-302.

43. **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.

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

Chapter 5

Distribution of *Mycobacterium tuberculosis* genotypes varies across regions of Southern Africa with high TB incidence[6]

Violet N. Chihota, Elizabeth L. Corbett, Beauty Makamure, Gunila Kanellius, Sayoki G. Mfinanga, Marnomorney Pillay, Elizabeth M. Streicher EM, Ludwig Apers L, Marisa Tait, Webster Kasongo, Grace Mbulo, Gian D. van der Spuy , Sophia O. Viegas, Zaina Cuna, Ramona Groenheit, Stanley Mungofa S, Sturm AW, Tommie VC, Cox H, Behr M, Nico Gey van Pittius, van Helden PD, Warren R M

[6] Manuscript in preparation and to be submitted to a peer reviewed journal

Abstract

Introduction

Different *Mycobacterium tuberculosis* genotypes contributing to tuberculosis have been described in countries of Southern Africa. However a comprehensive assessment of the population structure of *M. tuberculosis* in this region has not been done.

Aim

To compare the population structure of *M. tuberculosis* in Southern Africa regions and to describe the regional population, as an entity.

Methods

Spoligotyping and/or IS*6110*-restriction fragment length polymorphism (RFLP) data were received from collaborating institutes in Mozambique, Tanzania, South Africa, Zambia and Zimbabwe. Additional spoligotyping data was extracted from the SpolDB4 database and this was obtained for Madagascar, Malawi, Namibia and Zimbabwe, Further data from Tanzania was also obtained from published literature. Statistical analyses using X^2 test were performed to compare differences in proportions of the *M. tuberculosis* genotypes across these countries.

Results

A total of 5586 isolates from 8 countries in Southern Africa were included in this analysis, 53.2% of which were from South Africa. *M. tuberculosis* genotypes identified included Latin-American-Mediterranean (LAM) (34.2%), Beijing (16.1%), T (9.8%), Central Asian (CAS) (7.8%), X (6.9%) and East African Indian (EAI) (6.3%). LAM genotypes were predominantly isolated from Namibia (80% of isolates), 65% of isolates in both Zambia and Zimbabwe and 50% of isolates from Malawi. 88.3% of Beijing

genotypes identified were from South Africa. Other genotypes including T genotype were more prevalent in Madagascar, the CAS genotype in Tanzania, the EAI more prevalent in Mozambique (31.3%) and the S and X genotypes in South Africa (7.3%) and 11.2% respectively.

Conclusion

A large diversity of *M. tuberculosis* genotypes, with some genotypes localized within specific geographic regions was identified. These observations suggest that particular *M. tuberculosis* strains may be adapted to particular host populations in the region. This may have implications for development of diagnostic tools, vaccines and drugs**.**

Introduction

Mycobacterium tuberculosis, an important human pathogen causing tuberculosis (TB), kills close to 2 million people each year and infects 2 billion people globally (1). The TB incidence in Southern Africa has increased due to the HIV epidemic, worsening the public health problem posed by TB. It is estimated that there are ≥500 new TB cases per 100 000 population in each of the Southern African countries. The HIV prevalence in new TB cases is ≥50% and therefore TB remains the leading cause of death from an infectious disease (1,2).

In recent years genetic loci in *M. tuberculosis* have been identified and molecular tools developed to better understand the epidemiology of TB (3). The Direct Repeat locus (DR) consisting of DRs and variable spacers has proved to be invaluable in assessing the genetic variability of tuberculosis (4). Spoligotyping is an easy, robust, cheap and highly reproducible molecular tool to assay this DR locus (4). Combining Spoligotyping with other molecular techniques such as variable number tandem repeat (VNTR) typing (5), Mycobacterial-interspersed-repetitive-units (MIRU) (6) or IS*6110*-restriction fragment length polymorphism (RFLP) (7) increases the precision in identifying the different *M. tuberculosis* clones.

Different genotypes contributing to disease have been described globally with specific genotypes causing disproportionate disease in some populations (8). The availability of genotypic data, including SpolDB4 (9), not only allows comparability analysis of the global phylogeographical population structure but regional structures as well, together

with TB transmission and underlying spatial and temporal evolution of strains of the different genotypes (9). These are important as they contribute to the development of new tuberculosis control tools. However a comprehensive assessment of the genetic variability of *M. tuberculosis* in Southern African regions has not been done despite the availability of some genetic diversity data from countries in this region. Thus our current knowledge of the genetic diversity is limited to a study by Gagneaux *et al* (10). Using comparative genomic and molecular epidemiological tools on a global sample of 875 strains from 80 countries, the study defined six main lineages of *M. tuberculosis* that are highly geographically structured.

Predominant clades have been reported from different countries in the Southern Africa region varying from Central Asia (CAS)_Kili strains making up 30% of the population in Northern Tanzania and Dar es Salaam (11, 12), Latin American- Mediterranean (LAM)strains in 47.2% and 65% of isolates from Zimbabwe and Zambia respectively (13) and five major clades Beijing, LAM, Low copy clade (LCC), Haarlem (H) and Quebec have been reported as making up 84% of the population structure in a community in Western Cape (14). Invariably some of these strains have been associated with drug resistance. Beijing genotype has been associated with multidrug-resistant (MDR) and extensively drug-resistant (XDR) TB in provinces of South Africa (15, 16, 17) and is overrepresented in children with drug resistant TB (18), LAM being associated with XDR in KwaZulu Natal (19) and LCC strains in the Boland/Overberg and Southern Cape/Karoo (BOKS) regions in Western Cape (20).

Early indications from these studies show that the *M. tuberculosis* population structure may be less homogenous within countries and across different countries than previously thought (9, 10). The distribution and heterogeneity could possibly reflect the different host population, migration patterns and colonization history across the region. These variations may have implications for TB prevention and control strategies. To better understand the population structure across this region the current study will collect data from different regions of Southern Africa with the aim of describing the clades of *M. tuberculosis* contributing to the TB problem and the spread these strains across the region. It is envisaged that by describing the distribution of strains at a regional level, the disease dynamics will be better understood. These data on *M. tuberculosis* population structure will help contribute towards understanding clinical presentation and in the development and design of vaccines, drugs and diagnostics.

We hypothesize that specific genotypes of *M. tuberculosis* predominate in defined regions of Southern Africa and vary within and across different countries in this region. Such genotypes predominate and have a propensity to spread due specific bacterial and host population, migration and environmental factors. The aims of the study were to compare the population structure of *M. tuberculosis* in Southern Africa and to describe the regional population as an entity.

Methods

Study design

The study was a descriptive study using genotyping data available from countries that are part of Southern Africa to describe the population structure of *M. tuberculosis* complex.

Study population

The *M. tuberculosis* spoligotypes or IS*6110* RFLP data included in the study were received from co-investigators and collaborating institutes from Southern Africa. IS*6110* DNA fingerprint patterns were assigned to genotypes (families) accordingly to previously described criteria (21, 22). Where spoligotyping data was not available from investigators or institutes, data was searched on the SpolDB4 database or on Pub Med for published literature reporting *M. tuberculosis* strain diversity and spoligotypes in regions of Southern Africa.

All the spoligotyping data was entered in separate Microsoft Excel sheets and converted to STATA files using STATA 10 (Statacorp, College Station, Texas USA).

Briefly, isolates contributing to this analysis were collected as follows:

Mozambique

Sputum samples were collected from 2007-2008 from pulmonary TB patients with no prior history of tuberculosis as part of a nationwide drug resistance surveillance (DRS) program. Patients included in the surveillance were 40 randomly selected districts in seven provinces of Mozambique including, Maputo City, Maputo, Gaza, Inhambane, Nampula, Cabo Delgado and Niass Provinces. Only tuberculosis patients aged 15 years or older were included in the study. Sputum specimens were sent to the district

laboratories for smear microscopy and to National Reference Laboratory for culture and drug susceptibility testing. Heat killed cultures were sent to the Centre of Biotechnology of Eduardo Mondlane University, in Maputo city and to the Swedish Institute for Infectious Disease Control, in Solna, for genotyping.

South Africa

M. tuberculosis isolates included in this study were from three sites in Western Cape, as well as sites in Eastern Cape and KwaZulu Natal.

In the first site in Western Cape, as part of a long term molecular epidemiological project started in 1992, *M. tuberculosis* isolates were collected from patients residing in and attending primary health care facilities in an epidemiological field site. In this community approximately 350 new bacteriologically confirmed adult tuberculosis cases are reported per 100 000 population annually. Each isolate was classified by IS*6110* RFLP using internationally standardized protocol (7, 21).

As part of a prospective study done in a second site in Cape Town from April 2002 to 31 March 2004, patients presenting with symptoms of pulmonary TB were enrolled to the study. In addition, individuals identified as TB patients, from whom sputum samples were taken to determine whether or not they were infectious, were also enrolled in the study. Two sputum specimens were collected from each TB suspect and sent to the National Health Laboratory Services (NHLS) reference laboratory in Cape Town for

smear microscopy and culture. Positive Löwenstein-Jensen (LJ) slants were sent to the Stellenbosch University for IS*6110* RFLP.

Similarly, in the third suburb in Cape Town, South Africa, TB suspects having sputum taken for TB diagnosis and individuals identified as TB patients having sputum taken to determine whether or not they were infectious were enrolled to the study. Sputum samples were sent to the NHLS reference laboratory in Cape Town for smear microscopy and culture. Positive LJ slants were sent to the Stellenbosch University for spoligotyping.

Isolates from KwaZulu Natal were from sputum specimens collected from pulmonary TB patients hospitalized at two provincial hospitals in KwaZulu Natal from 2005-2006. Prior to 2006 specimens from patients in KwaZulu Natal were performed in 2 laboratories, one at Nelson R Mandela School of Medicine and at King George V Hospital in Durban. After February 2006, the samples were sent to NHLS for smear microscopy and culture. Positive LJ slants were sent to the University of KwaZulu Natal for genotyping using IS*6110* -RFLP and spoligotyping. Only isolates sensitive to anti-TB drugs were included in this analysis.

From July 2008-June 2009, tuberculosis cases that were suspected MDR-TB cases, as defined by clinical response to treatment, identified at primary health clinics in Eastern Cape Province were referred to the NHLS in Port Elizabeth for routine drug susceptibility testing. Only isolates sensitive to anti-TB drugs were included in this analysis.

Tanzania

As part of a routine drug resistance surveillance programme from 2001-2007, sputum samples were collected from both new and retreatment TB cases identified at TB clinics in 26 regions of Tanzania. Smear microscopy was done at local facilities and specimens were sent for culture to one of the three tuberculosis reference laboratories in Tanzania namely; the Central Tuberculosis Reference Laboratory (CTRL), Dar es Salaam, the Bugando Medical Centre (BMC), Mwanza and the Kilimanjaro Christian Medical Centre (KCMC), Moshi. Positive cultures were heat killed and shipped to Norway for spoligotyping.

Zambia

The isolates from Zambia (n=114) were part of a national drug resistance surveillance study done between December 1999 and January 20001. Specimens were obtained from new smear positive patients who had not started tuberculosis treatment, presenting at various medical centres in different provinces. The samples were cultured on LJ slants at the Chest Diseases Laboratory (CDL) in Zambia. All culture positive isolates were sent to the Tropical Disease Research Centre (TDRC) for DNA extraction and the DNA shipped to Stellenbosch University, for spoligotyping.

Zimbabwe

The *M. tuberculosis* isolates included in this study were from studies done in Harare and Gweru in Zimbabwe. For the first study; sputum specimens were collected from patients presenting with pulmonary TB at Beatrice Road Infectious Diseases Hospital (BRIDH),

the main referral centre for infectious diseases in Harare, from October to December 2001 (n=120). The second study included sputum specimens collected from pulmonary TB patients presenting at the Gweru Provincial Hospital, recruited from September 2000 to September 2001 (n=300) as part of a study examining the characteristics of different laboratory methods and the accuracy of diagnosis in routine circumstances and these results are described elsewhere (13, 23). In the third study sputum specimens were collected from patients presenting with chronic cough at primary health care clinics (n=134) (24) and from work place clinics in Harare (n=108). All sputum specimens collected in these studies were sent to the Biomedical Research and Training Institute (BRTI) in Harare for microscopy and culture. Positive cultures were sent to Stellenbosch University for spoligotyping and IS*6110*-RFLP.

Statistical analysis

To assess the population structure of *M. tuberculosis* in Southern Africa, individual databases for datasets from each data source were analyzed using STATA (Statacorp, College Station, Texas USA). After cleaning each dataset, the STATA files were merged by appending all the files. The Pearson's chi-squared test was used for testing the difference in proportion of *M. tuberculosis* genotypes within provinces and also among the different provinces. A level of $p<0.05$ was considered statistically significant.

Ethical considerations

The current study was part of long-term molecular epidemiology study which has been approved by the ethics committee of the Faculty of Health Sciences, Stellenbosch

University Ethics Committee. Approvals were also sought from Ethics committees of represented countries and from individual investigators to collaborate in the study. Participants in the study were identified by a unique study number. No personal identifiers were included in the data sent to Stellenbosch University

Results

Spoligotyping and/or IS*6110* RFLP data was received from co-authors in five countries including ten provinces of Mozambique (n=543), three provinces of South Africa including Western Cape Province (n=2648), KwaZulu Natal (n=119) and Eastern Cape (n=245); 20 regions of Tanzania (725), Zambia (n=114) and two provinces in Zimbabwe (Harare n=260, Midlands n=174). Additional spoligotyping data from Harare, Zimbabwe (n=245) was also extracted from SpolDB4 database together with data from Madagascar (n=323), Namibia (n=78) and Malawi (n=112). From the published literature additional data was obtained for Tanzania (Dar es Salaam n=146 and Northern Tanzania n=130) (11, 12). No data were available from Angola, Botswana, Democratic Republic of Congo (DRC), Lesotho and Swaziland.

***M. tuberculosis* genotype identified in Southern African countries**

The regional distribution of *M. tuberculosis* genotypes was assessed within 8 of the 14 (57.1%) countries in Southern Africa that are members of the Southern Africa Development Community (SADC). *M. tuberculosis* isolates from Western Cape Province were disproportionately represented and those from Namibia, Zambia and Malawi

underrepresented (Table 1). Of the 3012 isolates from South Africa, 2258 were genotyped using IS*6110* RFLP and lineages inferred from this DNA fingerprinting data.

Table 1: Number of *M. tuberculosis* isolates from the countries in Southern Africa

Country	*M. tuberculosis* Isolates (n)	Percentage (%)
Madagascar	323	5.8
Malawi	112	2.0
Mozambique	543	9.7
Namibia	78	1.4
South Africa	3012	53.9
Tanzania	725	13.0
Zambia	114	2.0
Zimbabwe	679	12.2
Total	5586	100.0

Major *M. tuberculosis* genotypes identified in regions of Southern Africa

The six most frequent genotypes identified among 5586 *M. tuberculosis* isolates from the eight Southern African countries were LAM (34.2%), Beijing (16.1%), T (9.8%), CAS (7.8%), X (6.9%) and East Asia India (EAI;6.3%),(Table 2).

In six countries *M. tuberculosis* isolates belonging to the LAM genotype were identified in high proportions and these included Namibia (81%), Zambia and Zimbabwe in similar proportion (65%), Malawi (50%), Mozambique (34%) and South Africa (31%) (Figure 1). A total of 68 shared types (STs) could be identified among these genotypes. Of the 78 LAM strains identified in Namibia 62 (79%) were LAM 1 genotype, LAM 11_ZWE genotypes were common in isolates from Mozambique (32.6%), Zambia (79.7%) and

Zimbabwe (64.3). Of 125 LAM strains from South Africa that had sub lineage data 121 were LAM 3 strains.

M. tuberculosis genotypes of the Beijing family were isolated from 899 patients and a majority of these, 794 (88.3%) were from South Africa. These Beijing strains represented 26.4% of the isolates from this country and a majority of these were typical Beijing strains (data not shown).

The ill defined T genotypes were isolated in high proportion among isolates from Madagascar (31.6%). These were also identified among isolates from the other seven countries (Figure 1). A majority of the *M. tuberculosis* isolates identified as T genotypes were T1 strains. The T1 strains were prevalent among T genotypes identified in each of the eight countries; 76 (74.5%) in Madagascar, all 18 isolates in Malawi, 22 (37.9%) in Mozambique, all 9 isolates from Namibia, 67 (43.5%) in South Africa, 50 (57.5%) in Tanzania, 24 (85.7%) in Zambia and 78 (86.7%) in Zimbabwe.

The CAS genotype which included the CAS1-Delhi and CAS1-Kilimanjaro (CAS1-KILI) was predominantly isolated from *M. tuberculosis* isolates from Tanzania (41.2%) and Madagascar (12.1%) and in less than 10% of the isolates from each of the remaining six countries (Figure 1).

The proportion of *M. tuberculosis* genotypes identified in the Southern African region (Table 2) differed significantly across the countries ($p=<0.001$).

Table 2: *M. tuberculosis* genotypes identified among isolates from eight countries in Southern Africa

Genotype	Madagascar	Malawi	Mozambique	Namibia	South Africa	Tanzania	Zambia	Zimbabwe	Total
Beijing	2 (0.6)	0 (0.0)	33 (6.1)	0 (0.0)	794 (26.4)	47 (6.5)	1 (0.9)	22 (3.2)	899 (16.1)
*CAS	39 (12.1)	8 (7.1)	12 (2.2)	0 (0.0)	39 (1.3)	299 (41.2)	4 (3.5)	33 (4.9)	434 (7.8)
**EAI	49 (15.2)	18 (16.1)	170 (31.3)	1 (1.3)	3 (0.1)	90(12.4)	4 (3.5)	17 (2.5)	352 (6.3)
***H/H like	28 (8.7)	0 (0.0)	15 (2.8)	0 (0.0)	220 (7.3)	10 (1.4)	1 (0.9)	16(2.4)	290 (5.2)
****LAM	30 (9.3)	56 (50.0)	187 (34.4)	63 (80.8)	909 (30.2)	148 (20.4)	74 (64.9)	442 (65.1)	1909 (34.2)
MANU	10 (3.1)	3 (2.7)	6 (1.1)	0 (0.0)	3 (0.1)	3 (0.4)	1 (0.9)	0 (0.0)	26 (0.5)
S	4 (1.2)	1 (0.9)	10 (1.8)	0 (0.0)	204 (6.8)	6 (0.8)	0 (0.0)	4 (0.6)	229 (4.1)
T	102 (31.6)	18 (16.1)	58 (10.7)	9 (11.5)	154 (5.1)	87 (12.0)	28 (24.6)	90 (13.3)	546 (9.8)
U	48 (14.9)	0 (0.0)	0 (0.0)	2 (2.6)	2 (0.1)	8 (1.1)	0 (0.0)	1 (0.2)	61 (1.1)
X	3 (0.9)	2 (1.8)	15 (2.8)	0 (0.0)	337 (11.2)	1 (0.1)	1 (0.9)	28 (4.1)	387 (6.9)
orphan	0 (0.0)	0 (0.0)	37 (6.8)	0 (0.0)	75 (2.4)	17 (2.3)	0 (0.0)	18 (2.7)	147 (2.6)
other	8 (2.5)	6 (5.4)	0 (0.0)	3 (3.9)	272 (9.0))	9 (1.2)	0 (0.0)	8 (1.2)	306 (5.5)
Total	323 (100.0)	112 (100.0)	543 (100.0)	78 (100.0)	3012(100.0)	725 (100.0)	114 (100.0)	679 (100.0)	5586 (100.0)

*CAS: Central Asian; LAM:
**Latin-American-Mediterranean;
***EAI: East African Indian;
****H/H-like: Haarlem/Haarlem-like

The well characterized IS*6110* low-banding family, the X genotype family was identified among 387 of the 5586 isolates included in the study. This family was most prevalent in South Africa where 337 isolates were classified as X genotype. No isolates belonging to the X genotype were identified in Namibia but these were isolated in low proportions among isolates from the other six countries (Figure 1).

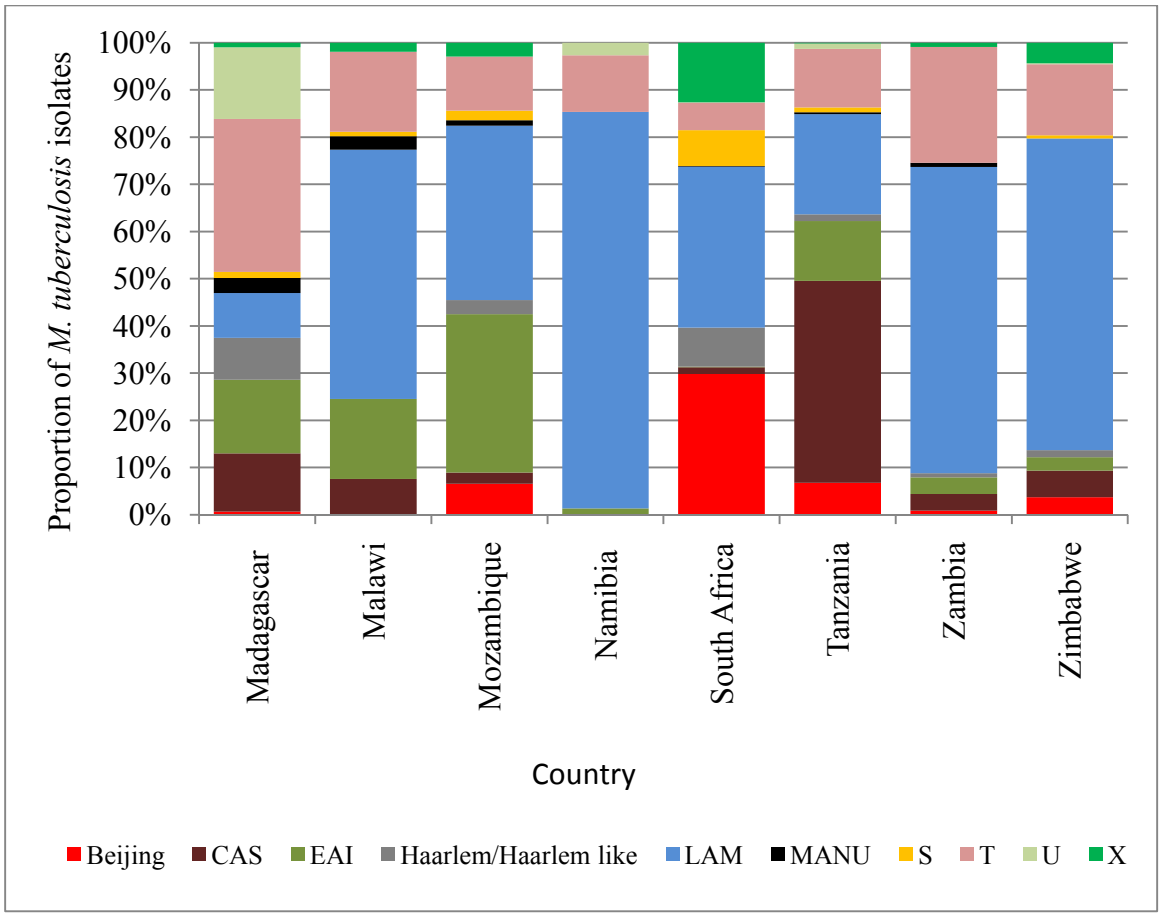


Figure 1: Percentage of ten main spoligotyping-defined *M. tuberculosis* genotype families identified in eight Southern African countries.

The EAI genotypes were identified among 352 isolates across all 8 countries but were isolated in high proportions from *M. tuberculosis* strains from Mozambique (31.3%), Madagascar (15.2%) and Tanzania (12.4%). The sub lineages were undefined in some

isolates characterized as EAI genotypes (n=16), but the rest included EAI5 (n=121), EAI1-Somalia (n=82), EAI8-Madagascar (n=46), EAI6-Bangladesh (n=38), EAI5/EAI3 (n=34) and others (n=15). Among the three most common EAI sub lineages, EAI5 was isolated in high proportions from Mozambique (n=54) and Tanzania (n=34), whilst EAI_Somalia was prevalent in Mozambique (n=75) and EAI_Madagascar in Madagascar (n=46).

The S genotype was more common in South Africa than in any of the other countries. The genotype was identified in 7.3% of the isolates from South Africa and ranging from very low proportions in five countries (0.6%-1.2%) to none in Zambia and Namibia.

Distribution of genotypes in regions of Southern Africa

Figure 2 is a pictorial map of Southern African countries showing the distribution of the main lineages identified in each of the eight countries. The LAM genotype was widely distributed across this region, most prevalent in the central parts of Southern Africa. The Beijing, X, Haarlem (H) and S genotypes were more likely to be isolated from South Africa than the other countries in Southern Africa. In addition the H genotype was also more likely to be identified in isolates from the island of Madagascar. The CAS genotype found mainly in India,the Indian subcontinent and Tanzania (9), was reconfirmed as being the most prevalent genotype in Tanzania. The EAI genotypes were isolated at considerable frequency in Mozambique, Malawi Madagascar and Tanzania.

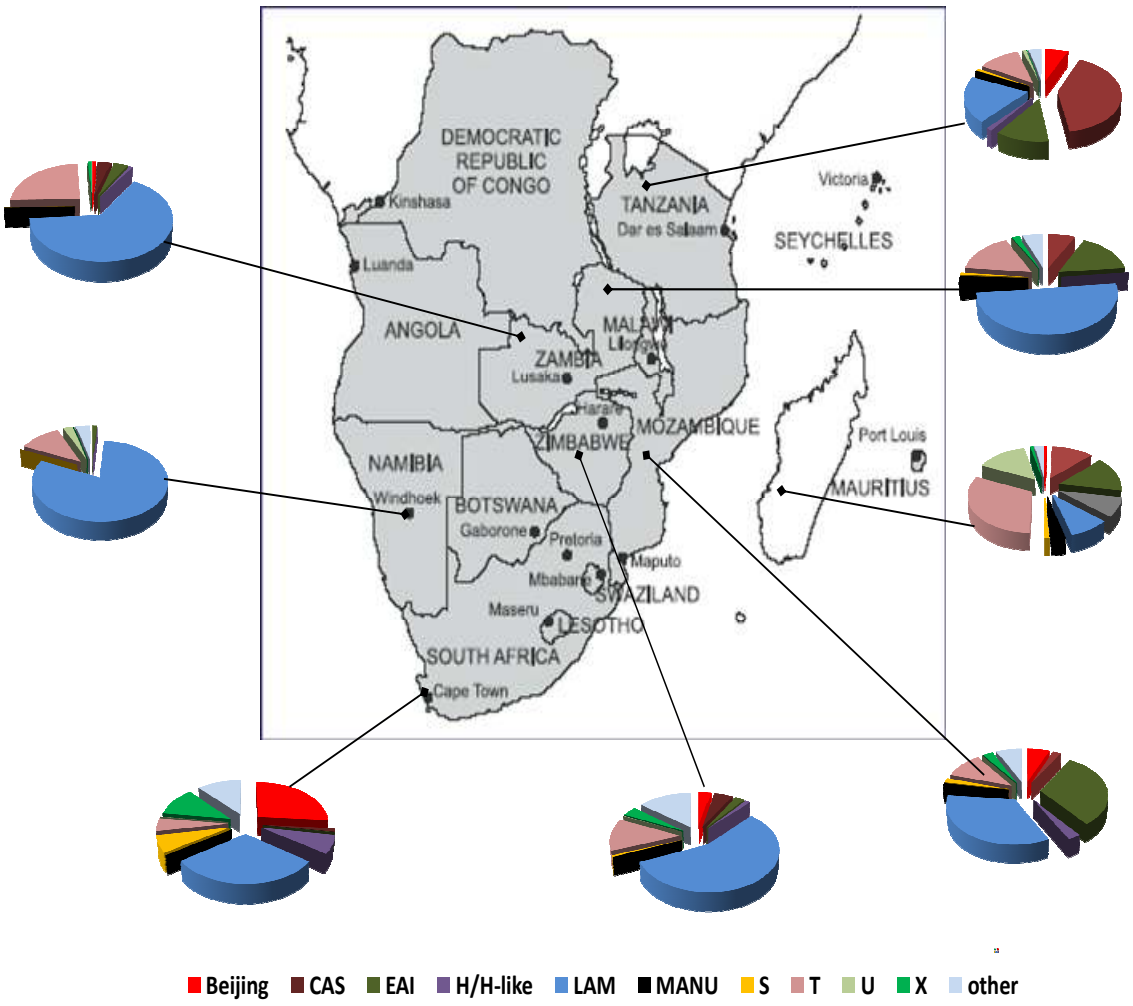


Figure 2: Distribution of *Mycobacterium tuberculosis* genotypes in Southern African countries

Discussion

In this study we analyzed genotypes of *M. tuberculosis* isolated from patients presenting with pulmonary tuberculosis in eight Southern African countries that are part of the SADC region. To our knowledge this is this is the first collaborative effort to describe genotypes of *M. tuberculosis* as an entity, driving the tuberculosis epidemic in this region of high TB incidence and HIV prevalence. Using spoligotyping data submitted from different countries and the SpolDB4 to compare and extract additional data enabled uniform and accurate comparisons to be made across different countries. These data could also be compared to genotypes inferred from IS*6110* RFLP data (22).

The study used a combination of data sources; including unpublished data collected as part of studies from some countries and published data that is now available on the public domain. This was done to ensure that as much of the genotyping data that is available in the region from studies is included in the analysis in order to get a representative sample to infer the population structure in the region.

The diversity of genotypes within this region seemed to be well structured along specific regional lines, but a few genotypes including LAM, Beijing, T, CAS, X and EAI were identified across most of the countries. Interestingly this comparison of the population structure in the different countries of Southern Africa demonstrates finer national differences in population structure, contrary to findings from earlier studies (10), where less evidence of phylogenetic structure was found in individual nations but more so in broader geographic regions. The observations that the LAM genotypes with a very rich

diversity were predominantly isolated from the Central parts of Southern Africa (Malawi, Namibia, Zimbabwe and Zambia) supports this notion. Similarly, EAI genotypes which have been linked to Asian countries (9) were predominantly found in countries on the Indian Ocean cost, including Madagascar, Mozambique and Tanzania. The prevalence of the EAI in Malawi may be explained by the fact that Malawi is a neighbouring country of both Mozambique and Tanzania and cross boarder migration may play a role in the spread of these strains. In part, this observation together with an observation of specific genotypes that are prevalent may reflect migratory and trade routes over the past 3 decades. A study by Brudey *et al* (9), also supported our observations by demonstrating that the genetic diversity of *M. tuberculosis* is strongly linked to finer geographical scale.

It was also interesting to note, however, that the Beijing genotype, which is known to be prevalent in Asia (Far East and Middle East Asia), emerging in other regions of the world and associated with drug resistance (25),was not prevalent in other countries on the Indian Ocean coast but more prevalent in South Africa. The Beijing genotypes which were primarily identified among isolates from Western Cape Province in South Africa were previously found to increase exponentially over time with a doubling time 4.86 years, unlike other predominant genotypes in this province (14). This together with the relative fitness of Beijing strains may suggests that they may have a selective advantage associated with transmission in this community. This may explain in part the differences in prevalence of Beijing genotypes in South Africa and other countries on the Indian Ocean coast. We may also be drawn to speculate that the identification of Beijing genotypes in low proportions in other countries of Southern Africa including

Mozambique, Tanzania and Zimbabwe may represent an emerging genotype in the rest of Southern Africa.

Even though the same genotypes were identified across the eight countries, certain genotypes caused disproportionate numbers of tuberculosis cases, specifically the LAM genotype in Malawi, Zambia and Zimbabwe. One sub lineage LAM11_ZWE was the most prevalent in all three countries. These were designated SAF1/RD^{Rio} strains in earlier studies and have been shown in an earlier chapter in this thesis (Chapter 6) and in other studies to be associated with more severe forms of tuberculosis (26). Though in part we can speculate that it could be explained by cross boarder migration and environmental factors, the finer details remain unknown. One possible explanation is that the country specific population structures observed could have evolved by drift after a single entry and spread of the *M. tuberculosis* clone between countries or as the population of *M. tuberculosis* expanded in each country. Other factors like prevalence of HIV and host specific factors may also contribute to selection of these specific clones which may be more transmissible in these specific conditions. Our findings confirm those of early studies showing that *M. tuberculosis* strains are associated with specific host population (8, 10, 27). This in part could also explain the association seen between other prevalent genotypes identified in regions of Southern Africa including LAM, CAS, EAI, S, T and X. No demographic and clinical data was available in this study to better understand why in specific countries certain genotypes cause a disproportionate number of tuberculosis cases. However an earlier study looking at potential risk factors and clinical outcomes of

disease due to a sublineage of the LAM family suggested that disease due to these strains was more likely to cause cavitary disease (Chapter 2).

Despite our attempt to get data from the region, it was still underrepresented in some of the countries (Namibia, Malawi) and completely absent in 6 of the 14 SADC countries. Consequently a clear picture of what is happening in the rest of the region is unknown and may be over represented for some countries such as South Africa which contributed 53.2% of the data analyzed. Therefore our findings may not give a representative snapshot picture of the regional population structure of *M. tuberculosis*. Despite these limitations, stratification of data by country allowed us to define the main genotypes that constitute the population structure of *M. tuberculosis* in 57% of the SADC region.

In conclusion, *M. tuberculosis* genotype data is increasingly becoming available from high burden countries including regions of Southern Africa. Our results suggest that the distinct genotypes are associated with specific countries and regions of Southern Africa with a geographical localization of some genotypes probably implying a complex interplay between geography, histories, host population migration and the TB epidemic. The observation that *M. tuberculosis* exhibits distinct phylogeographic population structure in these countries confirms that the epidemic in the region is a result of several outbreaks of specific genotypes requiring different approaches to diagnosis, treatment and prevention. As the different genotypes are likely to have different phenotypes this may affect the development of new diagnostics, vaccines and drugs. As more and more data becomes available it will be worthwhile repeating a similar study to increase the

resolution of the phylogeographical structure of *M. tuberculosis* in Southern Africa. Such information may help in influencing disease control efforts such as development of new diagnostic, drugs and vaccines. Within these regions this may influence the evolutionary order of the various genotypes, thereby underscoring the need for region specific approaches to controlling TB.

Acknowledgements

The data used in the study was contributed by the authors on the paper. Such a population based study would not have been possible without accurate diagnosis of TB, isolation of *M. tuberculosis* isolates and genotyping of the identified strains. This study was also made possible by the Wellcome Trust Research Leave Fellowship Grant (01680/2/00/2 to V N. Chihota).

References

1. **World Health Organization (WHO).** Global tuberculosis control: Short update to the 2009 report. WHO/HTM/TB/2009.426: WHO 2009 Geneva. Available from http://www.who.int/tb/publications/global_report/2009/update/en/index.html accessed 19 January 2010

2. **Corbett EL, Watt CJ, Walker N, Maher D, Williams BG, Raviglione MC and Dye C.** The growing burden of tuberculosis: Global trends and interactions with the HIV epidemic. *Arch Intern Med* 2003;163: 1009-1021

3. **Mostowy S, Behr MA**. The origin and evolution of *Mycobacterium tuberculosis*. *Clin Chest Med* 2005; 26:207-216

4. **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 Apr; 35(4):907-14.

5. **Frothingham R, Meeker-O'Connell WA**. Genetic diversity in the *Mycobacterium tuberculosis* complex based on variable numbers of tandem DNA repeats. *Microbiol* 1998; 144:1189-1196.

6. **Supply P, Mazara 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. **van Embdden JD, Cave MD, Crawford JT, dale JW, Eisenach B, Hermans P, Martin C, McAdam R and Shinnick T**. Strain identification of *Mycobacterium tuberculosis* by DNA fingerprinting: recommendations for a standardized methodology. *J Clin Microbiol* 31:406-409.

8. **Gagneaux S and Small P**. Global phylogeography of *Mycobacterium tuberculosis* and the implications for tuberculosis product development. *Lancet Infect Dis* 2007;7:328-37

9. **Brudey K, Driscoll JR, Rigouts L, Prodinger WM, Gori A, Al-Hajoj SA, Allix C, Aristimuño L, Arora J, Baumanis V, Binder L, Cafrune P, Cataldi A, Cheong S, Diel R, Ellermeier C, Evans JT, Fauville-Dufaux M, Ferdinand S, Garcia de Viedma D, Garzelli C, Gazzola L, Gomes HM, Guttierez MC, Hawkey PM, van Helden PD, Kadival GV, Kreiswirth BN, Kremer K, Kubin M, Kulkarni SP, Liens B, Lillebaek T, Ho ML, Martin C, Martin C, Mokrousov I, Narvskaïa O, Ngeow YF, Naumann L, Niemann S, Parwati I, Rahim Z, Rasolofo-Razanamparany V, Rasolonavalona T, Rossetti ML, Rüsch-Gerdes S, Sajduda A, Samper S, Shemyakin IG, Singh UB, Somoskovi A, Skuce RA, van Soolingen D, Streicher EM, Suffys PN, Tortoli E, Tracevska T, Vincent V, Victor TC, Warren RM, Yap SF, Zaman K, Portaels F, Rastogi N, Sola C.** *Mycobacterium, tuberculosis* complex genetic diversity: mining the fourth international spoligotyping database (SpolDB4) for classification, population genetics and epidemiology. *BMC Microbiology* 2006;6:23

10. **Gagneux S, DeRiemer K, Van T, Kato-Maeda M, de Jong BC, Narayanan S, Nicol M, Niemann S, Kremer K, Gutierrez MC, Hitty M, Hopewell PC and Small P** Variable host-pathogen compatibility in *Mycobacterium tuberculosis*. *Proc Natl Acad Sci U S A* 2006, 103(8):2869-2873

11. **Kibiki GS, Mulder B, Dolmans WMV, de Beer JL, Boeree M, Sam N, van Soolingen D, Sola C and van der Zanden AGM**. *Mycobacterium tuberculosis* genotypic diversity and drug susceptibility pattern in HIV-infected and non HIV infected patients in Northern Tanzania. *BMC Microbiology* 2007; 7:51

12. **Eldholm V, Matee M, Mfinanga GMS, Heun M and Dahle UR**. A first insight into the genetic diversity of *Mycobacterium tuberculosis* in Dar es Salaam, Tanzania, assessed by spoligotyping. *BMC Microbiology* 2006; 6:76

13. **Chihota V, Apers L, Mungofa S, Kasongo W, Nyoni IW, Tembwe R, Mbulo G, Tembo M, Streicher EM, van der Spuy GD, Victor T, van Helden P and Warren RM.** Predominance of a single genotype of *Mycobacterium tuberculosis* in regions of Southern Africa. *Int J Tuberc Lung* Dis; 11:311-318

14. **Van de Spuy GD, Kremer K, Ndabambi SL, Beyers N, Dunbar R, Marais BJ, van Helden P.D and Warren RM.** Changing *Mycobacterium tuberculosis* population highlights clade specific pathogenic characteristics. *Tuberculosis* 2009; 89: 120-125

15. **van Rie A, Warren RM, Beyers N, Gie RP, Classen CN, Richardson M, Sampson SL, Victor TC, van Helden PD**. Transmission of a multidrug-resistant *Mycobacterium tuberculosis* strain resembling "strain W" among non-institutionalized, human immunodeficiency virus-seronegative patients. *J Infect Dis* 1999; 180: 1608-15

16. **Johnson R, Warren RM, van der Spuy GD, Gey van Pittius NC, Theron D, Streicher EM, Bosman M, Coetzee GJ, van Helden PD and Victor TC**. Drug-resistant tuberculosis epidemic in the Western Cape driven by a virulent Beijing genotype strain. *Int J Tuberc Lung Dis* 2010; 14:119-121

17. **Mlambo CK, Warren RM, Poswa X, Victor TC Duse AG, Marais E**. Genotypic diversity of extensively drug resistant tuberculosis (XDR-TB) in South Africa. *Int J Tuberc Lung Dis* 12:99-104

18. **Marais BJ, Victor TC, Hesseling AC, Barnard M, Jordaan A, Brittle W, Reuter H, Beyers N, van Helden PD, Warren RM, Schaaf HS**. Beijing and

Haarlem genotypes are overrepresented among children with drug-resistant tuberculosis in the Western Cape Province of South Africa. *J Clin Microbiol*. 2006 Oct; 44(10):3539-43. Epub 2006 Aug 23.

19. **Brittle W, Reuter H, Beyers N, van Helden PD, Warren RM, Schaaf HS. Pillay M and Sturm W**. Evolution of the extensively drug resistant F15/LAM4/KZN strain of *Mycobacterium tuberculosis* in KwaZulu Natal, South Africa. *Clin Infect Dis* 2007;45:1409–1414

20. **Victor TC, Streicher EM, Kewley C, Jordaan AM, van der Spuy GD, Bosman M, Louw H, Murray M, Young D, van Helden PD, Warren RM**. Spread of an emerging *Mycobacterium tuberculosis* drug-resistant strain in the western Cape South Africa. *Int J Tuberc L*ung Dis. 2007 Feb; 11(2):195-201.

21. **Richardson M, van Lill SW, van der Spuy GD, Munch Z, Booysen CN, Beyers N, vn Helden PD, Warren RM**. Historic and recent events contribute to the disease dynamics of Beijing-like *Mycobacterium tuberculosis* isolates in a high incidence region. *Int J Tuberc Lung Dis* 2002; 6(11):1001-1011.

22. **Streicher EM, Victor TC, van der Spuy G, Sola C, Rastogi N, van Helden PD and Warren RM**. Spoligotype signatures in the *Mycobacterium tuberculosis* complex. J Clin *Micro2007; 45:237-240*

23. **Apers L, Mutsvangwa J, Magwenzi J, Chigara N, Butterworth A, Mason P, van Stuyft P.** A comparison of direct microscopy, the concentration method and the Mycobacteria Growth Indicator Tube for the examination of sputum for acid-fast bacilli. *Int J Tuberc Lung Dis 2003;* 7(4):376-381

24. **Munyati SS, Dhoba T, Makanza ED, Mungofa S, Wellington M, Mutsvangwa J, Gwanzura L, Hakim J, Nyakabau M, Mason PR, Robertson V, Rusakaniko S, Butterworth AE and Corbett EL**. Chronic cough in primary

health care attendees, Harare, Zimbabwe: Diagnosis and Impact of HIV infection. *CIin Infect Dis* 2008; 40:1818-1827

25. Glynn JR, Whiteley J, Bifani PJ, Kremer K and van Soolingen D. Worldwide occurrence of the Beijing/W strains *of Mycobacterium tuberculosis*: A Systematic Review. *Emerg Infect Dis* 2002; 8:843-849

26. **Lazzarini LCO, Spindola Sm, Bang H, Gibson AL, Weisenberg S, da Silva Carvalho, Augusto CJ, Huard RC, Kritski AL and Ho JL**. RD^{rio} *Mycobacterium tuberculosis* infe*ction is ass*ociated with higher frequency of cavitary pulmonary disease. *J Clin Microbiol* 2008; 46:2175-2183

27. **Hirsch AE, Tsolaki AG, DeRiemer K, Feldman MW and Small P**. Stable association between strains of *Mycobacterium tuberculosis* and their human host populations. *PNAS 2004; 101:4*871-4876.

Chapter 6

Population structure of multidrug and extensively drug-resistant tuberculosis (M/XDR-TB) in South Africa[7]

Violet N. Chihota, Borna Müller, Charmaine K. Mlambo, Manormoney Pillay, Marisa Tait, Elizabeth M. Streicher, , , Madeleine Hanekom, Nico Gey van Pittius, Gerrit Coetzee, Andre Trollip, Cindy Hayes, Sturm AW, Elsie Marais, Tommie C. Victor, Paul D. van Helden and Robin M. Warren

[7] Manuscript in preparation and to be submitted to a peer reviewed journal

Abstract

Background

Drug-resistant tuberculosis (TB) and particularly multidrug-resistant (MDR) and extensively drug-resistant (XDR) TB, threaten TB control efforts throughout the world

Aim

To describe the population structure of *M. tuberculosis* strains isolated from patients with drug resistant TB in four provinces of South Africa

Methods

Multidrug-resistant (MDR), pre-extensively drug resistant (pre-XDR) TB and extensively drug-resistant (XDR) *M. tuberculosis* isolates were genotyped using spoligotyping and/or IS*6110*-restriction fragment length polymorphism (RFLP) using internationally standardized protocols. Differences in proportions of the *M. tuberculosis* genotypes among MDR-TB, pre-XDR-TB and XDR-TB in the four provinces were compared using Pearson's chi-squared test.

Results

A total of 3795 drug resistant isolates from Western Cape (n=2489), Eastern Cape (n=307), KwaZulu Natal (n=234) and Gauteng (n=765) were included in the analysis. Beijing genotypes were the most prevalent genotypes identified from Eastern and Western Cape, 42.5% of MDR-TB isolates, 96.8% of pre-XDR-TB isolates and 92.5% of XDR-TB and 47.9% among MDR-TB, 68.0% among pre-XDR and 92.3% among XDR-TB isolates respectively. Among 765 MDR-TB isolates from Gauteng the most prevalent genotypes were Latin American Mediterranean (LAM; 26.7%), Beijing (16.7%) and T (17.7%) genotypes. The S genotype (33.5%) and LAM genotype (24.9%) were the most

common genotypes identified from KwaZulu-Natal MDR-TB isolates. The proportion of LAM strains increased to 41.7% and 80% among PreXDR-TB and XDR-TB isolates respectively and were predominantly LAM4 genotype previously named F15/LAM4/KZN. The proportion of genotypes varied significantly across the three resistance classes and among the provinces ($p<0.001$).

Conclusion

In conclusion, the population structure of drug-resistant TB in South Africa is geographically localized. The diversity of *M. tuberculosis* genotypes decreased in all settings from MDR-TB to XDR-TB suggesting an increased ability of a limited number of strains to develop XDR-TB. These results support the need for identifying genetic markers in drug resistant *M. tuberculosis* which can aid in developing diagnostic tests that allow the identification of cases at high risk of developing XDR-TB. and early initiation of adapted treatment regimens.

Introduction

An estimated 440 000 cases of multidrug- resistant (MDR) tuberculosis (TB) occurred worldwide in 2008 of which 360 000 were likely due to transmission of an MDR strain and 94 000 were likely acquired resistance (1). In South Africa MDR has been reported in 1.8% new TB cases and 6.7% among previously treated cases. High levels of extensively drug-resistant (XDR) TB (10.5%) were reported from 5451 MDR cases tested for second line drugs (1). Drug resistant TB, including MDR- and XDR-TB, is a major challenge to the TB control because they have the most profound effects on patient treatment outcomes. The most effective antibiotics with the least side effects, isoniazid (INH) and rifampicin (RIF) must be replaced with less effective, more expensive and more toxic drugs in the treatment of MDR- and XDR-TB.

As predicted by mathematical models the future of MDR-TB and XDR-TB will depend largely on the efficiency of their transmission (2, 3) and hence cannot be contained without specific efforts to limit transmission. Development of drug resistance in Mycobacteria is due to genomic events unlike in other infectious agents where exchange of genes plays a major role (4). Selection of resistant mutants therefore happens during treatment with anti-tuberculous drugs when patients are treated with inappropriate regimens or when patients do not adhere to treatment. Non adherence is worsened by the large number of drugs that should be taken, side effects and the long duration of treatment (six months in new TB cases, ≥9 months for retreatment cases and even longer for patients with drug resistant disease). Hence human behaviour is very important in the development of drug resistance in TB.

In recent years availability of culture, drug susceptibility testing of both first and second line drugs and genotyping has allowed strain diversity studies to be done to determine the disease dynamics of drug resistant strains. As such genotyping data is available to allow comparability analysis of regional phylogeographical population structure as well as transmission and underlying spatial and temporal evolution of drug resistant strains of the different genotypes (5, 6). These are important as they will help either confirm or refute predictions on the future of drug resistant tuberculosis made through mathematical modeling. These molecular epidemiology studies have also affirmed the role of human behavior in the transmission of drug resistant TB (5). More recently temporal analysis of drug resistant tuberculosis in communities in Western Cape Province showed that the genetic background of *M. tuberculosis* strains contributed to transmission and acquisition of drug resistant TB (6).

Worldwide, molecular genotyping of *M. tuberculosis* strains has shown diverse lineages contributing to drug resistant TB worldwide. Specific genotypes have been suggested as being more successful than others in causing drug resistant TB, including Beijing (7, 8, 9, 10) LAM (11) and strains belonging to the IS*6110* low copy number clade (4). The Beijing strains can be grouped as modern (typical) or old (atypical) according to the presence or absence of an IS*6110* insertion in the NTF locus of the *M. tuberculosis* genome (12). Characterization of strains using these molecular methods has enabled development of rapid tests to detect outbreak strains (13). Subsequently these strains have been hypothesized as having acquired unique properties that allow them to have a fitness advantage over other strains (14, 15, 16, 17).

Our current knowledge of the genetic diversity is limited to studies done in *M. tuberculosis* in general and not specific to drug resistant strains (18, 19). The studies show that the population structure of *M. tuberculosis* is highly structured along specific geographic lines and to a lesser extent in individual countries. To date paucity of studies looking at the occurrence of drug resistant strains within provinces and across provinces of South Africa has prevented comparative studies to be done. However, in recent years genotyping data has become available from other provinces including Gauteng and KwaZulu-Natal, but no comprehensive assessment of the genetic variability of drug resistant *M. tuberculosis* has been done. This information may be helpful in guiding TB control programmes (20).

We hypothesize that specific lineages and therefore a distinct phylogeographic structure exists among drug resistant *M. tuberculosis* strains isolated from TB patients across the different provinces of South Africa. These genotypic differences exist because programmatic constraints and fitness of resistant *M. tuberculosis* strains that have now adapted to specific human populations.

In this study we use recently collected data on drug resistant *M. tuberculosis* isolates from 4 provinces of South Africa, namely, the Eastern Cape, Gauteng, KwaZulu Natal and Western Cape. The aim of the study was to describe the population structure of *M. tuberculosis* strains isolated from patients with drug resistant TB in four provinces of South Africa

Methods

Study design

The study was a descriptive study which included *M. tuberculosis* isolates collected as part of studies done in four provinces in South Africa, including Western Cape, Eastern Cape, KwaZulu Natal and Gauteng.

Study population

Isolates from Western Cape were collected during the period August 2000 to January 2010. Sputum specimens were routinely collected from all patients whose sputum failed to convert to smear negative after two months of intensive phase therapy for TB or at diagnosis for patients who had previously been treated with anti-tuberculous drugs (retreatment cases), attending primary health clinics in Cape Metropole, Boland, Overberg, Karoo and the Southern Cape regions of Western Cape. Sputum specimens were sent to the National Health Laboratory Services (NHLS) in the Western Cape Province or Stellenbosch University Laboratory.

Following decontamination using sodium hydroxide-N-acetyl-L-cystein (NALC-NaOH), auramine stained slides were examined for acid fast bacilli and culture done on Mycobacterial Growth Indicator Tube (MGIT). Using the indirect proportion method on Middlebrook 7H11 medium, drug susceptibility testing (DST) was done for isoniazid (INH) 0.2µg/ml; rifampicin (RIF) 1.0 µg/ml; ethambutol (EMB) 7.5µg/ml; streptomycin (SM) 2.0 µg/ml; kanamycin (KAN) 6.0µg/ml; ethionamide (ETH) 10.0 µg/ml; and ofloxacin (OFX) 2.0 0.2µg/ml (21). Susceptibility testing for amikacin (AMK) was done

using BACTEC 960 using a concentration of 1.0 µg/ml as recommended for routine screening (22). Testing for second line drugs was only started from 2006 onwards. All drug resistant isolates were sent to Stellenbosch University for genotypic analyses. Spoligotyping was done on all isolates and IS*6110* on a subset of isolates.

From July 2008–June 2009, isolates of tuberculosis cases with a suspicion of MDR-TB as defined by lack of clinical response to treatment, identified at primary health clinics in Eastern Cape Province were referred to the NHLS in Eastern Cape Province for routine drug susceptibility testing (DST). The same laboratory procedures for testing were used as described above for Western Cape Province.

In KwaZulu Natal sputum samples were collected from patients suspected of having MDR- and/or XDR-TB referred to a tertiary hospital and a provincial hospital in KwaZulu Natal. The two hospitals are referral hospitals for the KwaZulu Natal Province and data was collected during the period 2005-2006. Sputum samples were sent to either of two laboratories, one at the Nelson R Mandela School of Medicine and the other at King George V Hospital in Durban. After February 2006, the samples were sent to NHLS for smear microscopy, culture and DST. At the School of Medicine laboratory the 1% proportion method on Middlebrook 7H10 agar was used at the following concentrations; INH 0.2 mg/L and/or 1 mg/L; RIF 1 mg/L; EMB 7.5 mg/L, SM 2 and/or 10 mg/L. For specimens sent to King George hospital, DST was done on LJ media at the following drug concentrations INH 0.2 mg/L and 1 mg/L; RIF 40mg/L; EMB 2 mg/L, SM 4 mg/L;

ETH, 20 mg/L; KAN 20 mg/L; OFX 2.5 mg/L. Positive LJ slants were sent to University of KwaZulu-Natal for genotyping using IS*6110* –RFLP and spoligotyping.

During the period 2004-2007 isolates from suspected drug resistant TB cases were collected from the seven regions of Greater Johannesburg and sent to NHLS laboratories for drug susceptibility testing for four first line drugs including INH, RIF, SM and EMB. Drug susceptibility testing was done using the indirect proportion method on Middlebrook medium containing critical concentrations of 0.2 µg/ml INH and 30 µg/ml RIF.

Definitions

In this study drug-resistant strains were defined according to WHO definitions: MDR-TB isolates defined as *M. tuberculosis* isolates resistant to at least INH and RIF and XDR-TB strains as isolates that were resistant to at least INH, RIF, a fluoroquinolone such as OFX and one of three injectable second line drugs, either capreomycin (CPM), KAN or AMK (23). Pre-XDR-TB strains which are essentially one mutation away from becoming XDR-TB were defined as MDR-TB isolates (i.e. at least resistant to both INH and RIF) and either a fluoroquinolone or a second-line injectable drug, but not both (24). In this analysis the MDR-TB population in the Eastern and Western Cape and KwaZulu Natal, excluded all identified pre-XDR-TB and XDR-TB isolates.

Genotypic classification

M. tuberculosis isolates from patients were subcultured on LJ medium and genotypes were determined using spoligotyping (25) and/or IS6110 DNA fingerprinting (26) according to internationally standardized methods.

Sub-classification of all the Beijing strains into typical and atypical strains, identified in Eastern Cape, a sample from Western Cape and another sample from Gauteng was done using a polymerase chain reaction (PCR) amplification method as previously described (27). None of the Beijing strains identified in KwaZulu Natal were further characterized.

Comparison of isolates

The spoligotype patterns were compared to spoligotype data deposited in the SpolDB4 database maintained at the Pasteur Institute in Guadeloupe (28). Strains were grouped according to different clades/lineages as defined in the SpolDB4 database. Genotypes from IS*6110*-RFLP were assigned to families as previously described (29).

Inclusion criteria

Only the first *M. tuberculosis* isolate from each case was included in this analysis. Isolates were included in the analysis if they were MDR, pre-XDR or XDR and if the province where the patient resided was known.

Statistical analysis

Statistical analyses were performed using STATA 10.0 (Statacorp, College Station, Texas USA). The Pearson's chi-squared test was used for testing the difference in proportion of *M. tuberculosis* genotypes in the different resistance classes within provinces and also among the different provinces. A level of $p<0.05$ was considered statistically significant.

Ethics considerations

The current study was part of a long-term molecular epidemiology study which has been approved by the Ethics Committee (Institutional Review Board) of the Faculty of Health Sciences Stellenbosch University Ethics Committee. Approval was also sought from individual investigators to collaborate in the study. Participants in the study were identified by a unique study number and no personal identifiers were included.

Results

***Mycobacterium tuberculosis* isolates**

The number of isolates with spoligotyping data received from the provinces is shown in Supplemental table 5. In this analysis spoligotyping data for MDR-TB, pre-XDR-TB and XDR-TB isolates (n=3795, received from 4 provinces in South Africa namely; Western Cape (n=2489), Eastern Cape (n=307), KwaZulu Natal (n=234) and Gauteng (n=765) were included (Supplemental table 5 column 3, 4 and 5).

***M. tuberculosis* genotypes identified among drug resistant isolates from all provinces**

Overall of the 3795 *M. tuberculosis* identified in this study as MDR Pre-XDR or XDR, 1664 (43.9%) belonged to the Beijing genotype, 544 (14.3%) to the LAM genotype, 478 (12.6) to the X genotype, 462 (12.2%) to the more modern T lineage, 189 (5.0%) to the S lineage, 57 (1.5%) to the U genotype and 28 (0.7%) to CAS. Other genotypes were identified among 260 isolates and these included mixed strains. A small proportion of isolates (n=113) had spoligotype patterns that were not matched to the SpolDB4.

***M. tuberculosis* genotypes identified among drug resistant isolates from Eastern Cape**

Of the 307 *M. tuberculosis* isolates from Eastern Cape, 134 (43.6%) MDR-TB isolates were identified, 93 (30.3%) pre-XDR-TB and 80 (26.1%) XDR-TB isolates. The Beijing genotype was predominant among all drug resistant strains: MDR-TB isolates (57/134), pre-XDR-TB isolates (90/93) and XDR-TB isolates (74/80) (Figure 1). Of the 57 Beijing isolates that were MDR, 39 (68.4%) were atypical Beijing genotypes and 18 (31.6%) were typical Beijing genotypes (Figure 1). All Beijing isolates identified among pre-XDR isolates (n=90) and XDR isolates (n=74) were atypical Beijing genotypes (Figure 1).

Close to two-thirds of isolates identified as LAM were MDR-TB isolates and the proportion of LAM strains decreased among pre-XDR-TB and XDR-TB isolates (Figure 1). In total 39 LAM strains were identified and these were predominantly LAM 4 genotypes (n=30; 76.9%). *M. tuberculosis* isolates of the T genotype were isolated among 18 MDR strains and four sub lineages were identified, a majority (38.9%) belonged to the T1 sub lineage. The number of *M. tuberculosis* isolates belonging to

other genotypes decreased as we moved from MDR-TB to XDR-TB while that of Beijing genotypes increased, in particular the atypical Beijing genotype. The proportion of *M. tuberculosis* genotypes differed across the three resistance classes (p<0.001).

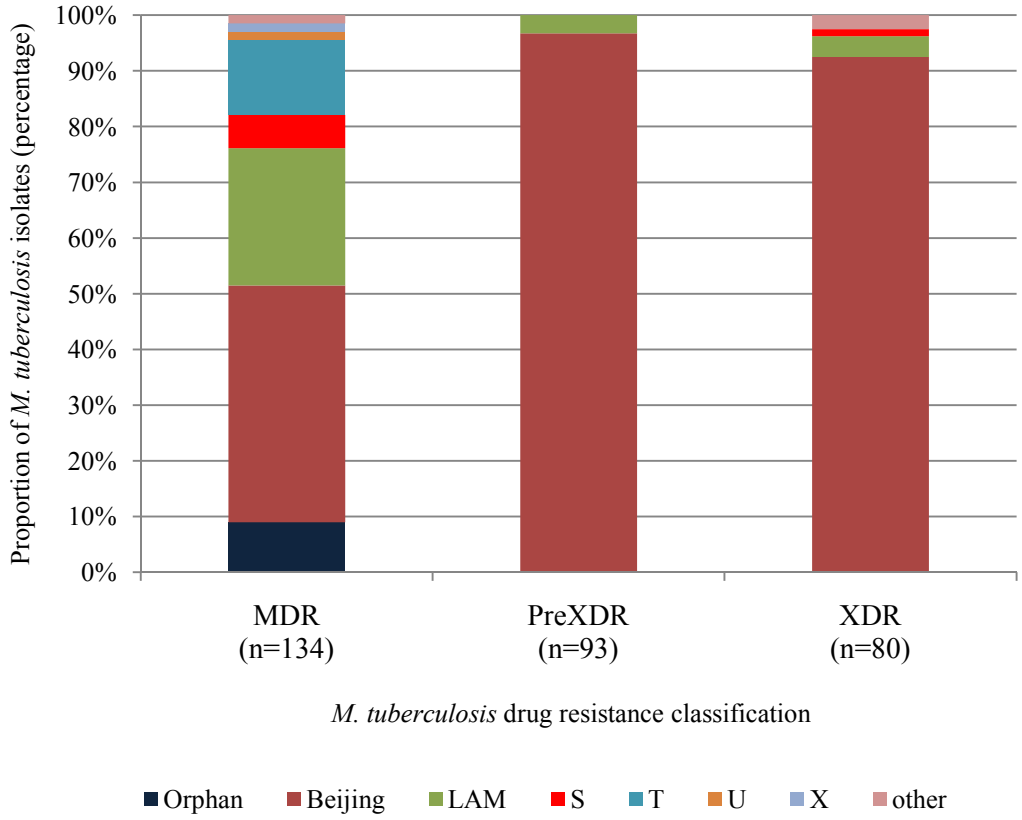


Figure 1: *M. tuberculosis* genotypes associated with MDR-TB, Pre-XDR-TB and XDR-TB in Eastern Cape Province, South Africa

***M. tuberculosis* genotypes identified among drug resistant isolates from Gauteng Province**

All the *M. tuberculosis* isolates from Gauteng province (n=765) were classified as MDR-TB isolates (Table 1). There was no data on second line testing, therefore inference to either pre-XDR or XDR could not be done. It is clear from Table 1 that the proportion of *M. tuberculosis* genotypes differed in this province. The most prevalent genotypes associated with MDR were LAM (26,7%), followed by T (17.7%) and Beijing (16.7%). Among the 204 MDR isolates identified as LAM, 129 (63.2%) belonged to the LAM4

lineage. One hundred and thirty-five isolates were identified as T genotypes and 87% of these belonged to the T1 sub lineage.

Of the 128 Beijing strains, a sample (n=81) were sub-classified further as either typical or atypical Beijing strains. A majority 67 (82.7%) were typical Beijing and 14 (17.3%) were atypical Beijing strains (Figure 4).

Table 1: Genotypes identified among *M. tuberculosis* MDR isolates

Genotype	Resistance class MDR	Percentage
LAM	204	26.7
T	135	17.6
Beijing	128	16.7
H	106	13.9
EAI	63	8.2
S	54	7.1
X	46	6.0
Orphan	15	2.0
CAS	12	1.6
MANU	2	0.3
Total	765	100.0

***M. tuberculosis* genotypes identified among drug resistant isolates from KwaZulu Natal Province**

Of the 234 *M. tuberculosis* isolates from KwaZulu-Natal, 196 (74.8%) were MDR-TB isolates, 13 (5.0%) were preXDR-TB isolates and 25 (9.5%) were XDR-TB isolates. Among MDR-TB isolates, 66 (33.5%) were S genotype, 49 (24.9%) were LAM genotypes and 33 (16.8%) T genotypes (Figure 2). Of the 196 MDR-TB isolates only 15 (7.6%) were Beijing strains and were not characterised further into typical or atypical Beijing lineages.

Four LAM sub lineages were identified and LAM 4 was the most prevalent being identified in 39 (79.6%) isolates. The proportion of LAM strains increased to 41.7% and 80% among preXDR-TB and XDR-TB isolates respectively (Figure 2). LAM 4 represented the majority of LAM strains within the pre-XDR and XDR groups, at 80% and 90% respectively. The proportion of other genotypes decreased and the Beijing genotype was not identified among pre-XDR-TB or XDR-TB isolates. The proportion of *M. tuberculosis* genotypes differed significantly among MDR-TB, preXDR-TB and MDR-TB isolates ($p<0.001$).

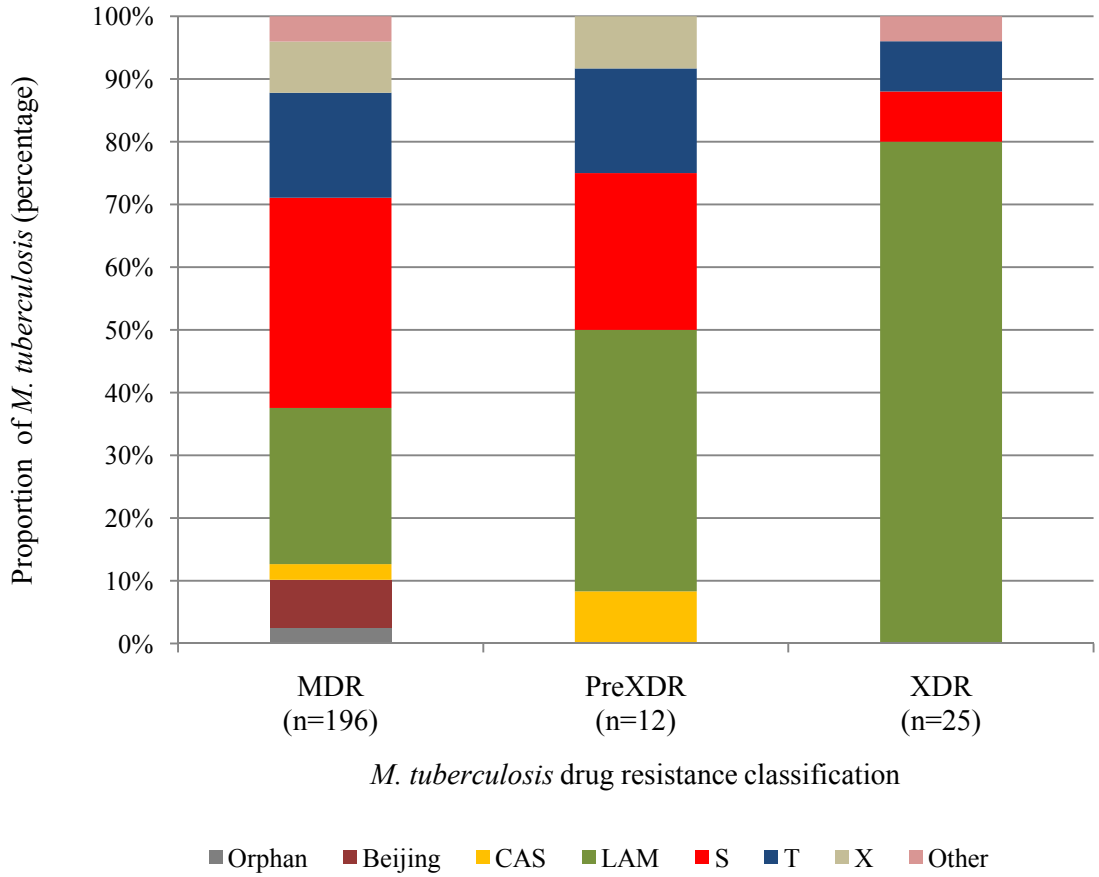


Figure 2: *M. tuberculosis* genotypes associated with MDR-TB, pre-XDR-TB and XDR-TB in KwaZulu Natal, South Africa

M. tuberculosis genotypes identified among drug resistant isolates from Western Cape Province

Of the 2489 isolates identified from TB patients in Western Cape, 2091 (84.0%) were MDR-TB isolates, 281 (11.3%) preXDR-TB isolates and 117 (4.7%) were XDR-TB isolates. Strains belonging to the Beijing genotype were predominant among all the three drug resistant isolates (MDR-TB, pre-XDR-TB and XDR-TB isolates) (Figure 3).

Beijing strains were identified from 1001 (47.9%) MDR-TB isolates, 191 (67.9%) pre-XDR-TB isolates and 108 (92.3%) XDR-TB isolates (Figure 3). A subset of the isolates identified as Beijing genotype (312/1300) were sub-classified as either typical or atypical strains. Typical Beijing strains were isolated in higher proportions from MDR isolates (159/189; 84.1%) and pre-XDR isolates (42/66; 63.6%), compared to XDR isolates (25/57; 43.8%) (Figure 4-6). On the contrary, atypical Beijing strains were isolated in lower proportions from MDR (30/189; 15.9%) and pre-XDR isolates (24/66; 36.4%) compared to XDR isolates (32/57; 56.1%).

The X genotype (X1, X2 and X3) was identified in 413 isolates (16.6%) and the majority of these (93.5%) were identified from MDR-TB isolates.

Across the three drug resistance classes, the proportion of *M. tuberculosis* genotypes was significantly different ($p>0.001$)

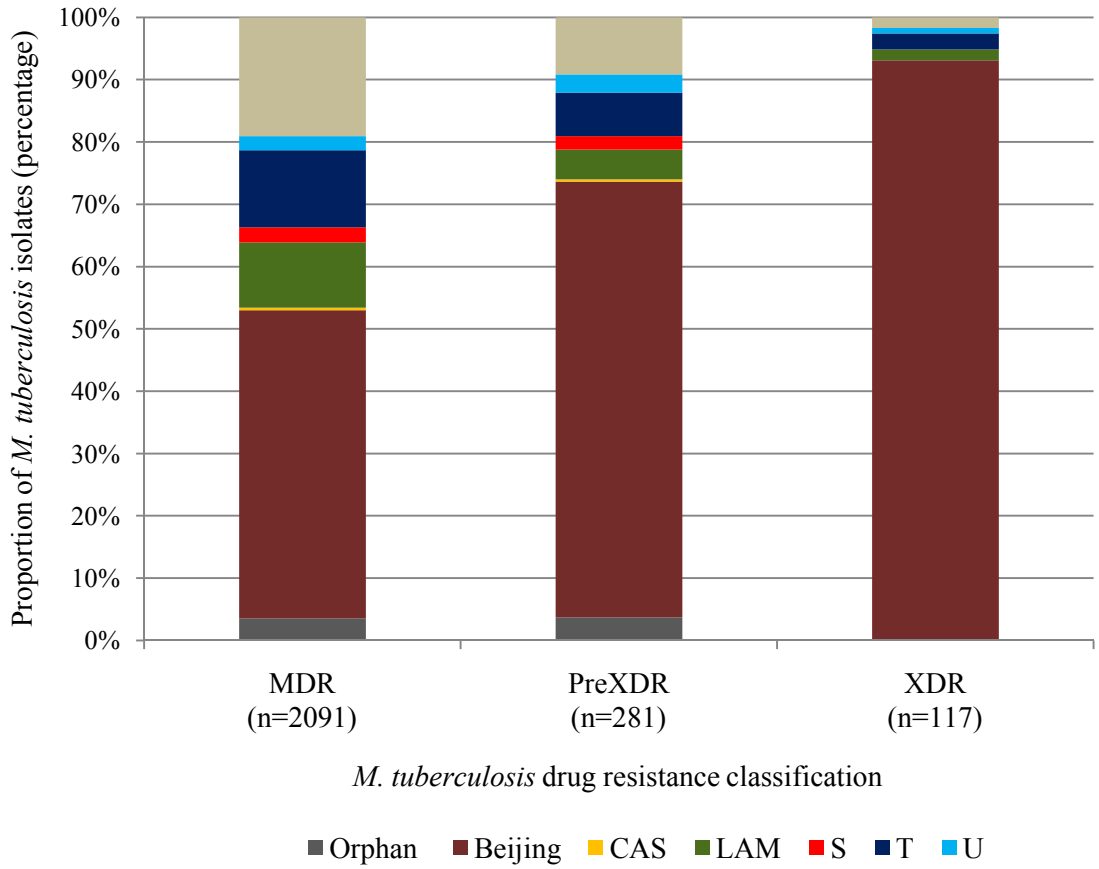


Figure 3: *M. tuberculosis* genotypes associated with MDR-TB pre-XDR-TB and XDR-TB in Western Cape Province, South Africa

Mycobacterium tuberculosis genotypes and drug resistance: comparison between the provinces

MDR-TB

In general *M. tuberculosis* isolates of the Beijing genotype were the most common cause of MDR-TB in the four provinces of South Africa. In the Eastern and Western Cape provinces MDR-TB isolates were predominantly Beijing genotypes, (n=57; 42.5%) and (n=1001; 47.9%), respectively (Table 2). Among the 57 Beijing MDR-TB isolates from Eastern Cape, 68.4% were atypical Beijings and 31.6% typical Beijings. In the Westen

Cape province, *M. tuberculosis* isolates belonging to the X genotype represented close to one-fifth of the MDR-TB isolates.

In Gauteng *M. tuberculosis* isolates of the LAM genotype (26.7%) were predominantly isolated from MDR-TB isolates. This changed in KwaZulu Natal with *M. tuberculosis* isolates of the S genotype reported in a majority of the MDR isolates (33.5%) followed by LAM genotypes (25.0%). The distribution of MDR isolates across the provinces of South Africa is shown in Figure 4 below.

Comparing across the provinces the genotypes associated with MDR-TB varied significantly ($p<0.001$) (Table 2).

Table 2: Comparison of *M. tuberculosis* genotypes identified among multi drug-resistant isolates from four provinces in South Africa

M. tuberculosis genotype	Province Eastern Cape	Gauteng	KwaZulu Natal n (%)	WesternCape n (%)	Total
Beijing	57 (42.5)	128 (16.7)	15 (7.7)	1001 (47.9)	1201 (37.7)
LAM	33 (24.6)	204 (26.7)	49 (25.0)	212 (10.1)	498 (15.6)
T	18 (13.4)	135 (17.7)	33 (16.8)	250 (12.0)	436 (13.7)
X	2 (1.5)	46 (6.0)	16 (8.1)	386 (18.5)	450 (14.1)
S	8 (6.0)	54 (7.1)	66 (33.7)	49 (2.3)	177 (5.6)
Other	16 (11.9)	198 (25.9)	17 (8.7)	193 (9.2)	424 (13.3)
Total	134 (100.0)	765 (100.0)	196 (100.0)	2091 (100.0)	3186(100.0)

Pre-XDR-TB

M. tuberculosis isolates that were at least resistant to both INH and RIF and either a fluoroquinolone or an injectable second line anti-TB drug were identified in 386 isolates from Eastern Cape (n=93), KwaZulu Natal (n=12) and Western Cape (n=281). Of the 93 Pre-XDR-TB isolates from Eastern Cape, 90 (96.8%) were Beijing strains, all of which were of the atypical Beijing genotype and 3 (3.2%) were LAM strains (Figure 5). Five genotypes could be identified among 12 pre-XDR isolates from KwaZulu Natal. Five of the 12 isolates (41.7%) were LAM genotypes and 4 (80%) of these were LAM4 sub lineage. Although a number of *M. tuberculosis* genotypes were identified from 281 pre-XDR isolates from the Western Cape Province, the Beijing genotype was the most predominant genotype (n=191; 68.0%), and less frequently were the T (6.8%) and LAM genotypes (4.6%). Further characterization of a subset of the isolates (n=69), showed that

a majority of pre-XDR isolates that belonged to the Beijing genotype were typical Beijing strains (42/69). The distribution of the genotypes across the three provinces is shown in Figure 5. Genotypes identified among the pre-XDR isolates varied significantly across the three provinces ($p<0.001$).

XDR-TB

XDR-TB was identified in 222 isolates from the Eastern Cape (n=80), KwaZulu Natal (n=25) and the Western Cape (n=117) and none from Gauteng Province.

Of the 80 XDR isolates from the Eastern Cape, all 74 (92.5%) isolates identified as Beijing were atypical Beijing strains, 3 (3.8%) LAM, 1 (1.3%) S type and 2 (2.5%) were other genotypes (Figure 6). *M. tuberculosis* isolates from KwaZulu Natal that were XDR were predominantly LAM genotypes (n=20; 80.0%). A majority of the LAM genotypes were of the LAM4 sub lineage (90%). The S and T genotypes were isolated in similar proportions (8.0%). As with the pre-XDR isolates identified in the Western Cape, the XDR-TB isolates belonged to a variety of *M. tuberculosis* genotypes (Figure 6), however, the Beijing genotype still predominated being represented in 92.3% of the XDR-TB isolates (Figure 6). Of the subset of isolates that could be characterized as typical or atypical (n=60), 32 (53.3%) were atypical Beijing genotype and 25 (41.6%) were typical Beijing types and 3 were mixed. The genotypes identified among the XDR-TB isolates were significantly associated with the origin of the isolates ($p<0.001$).

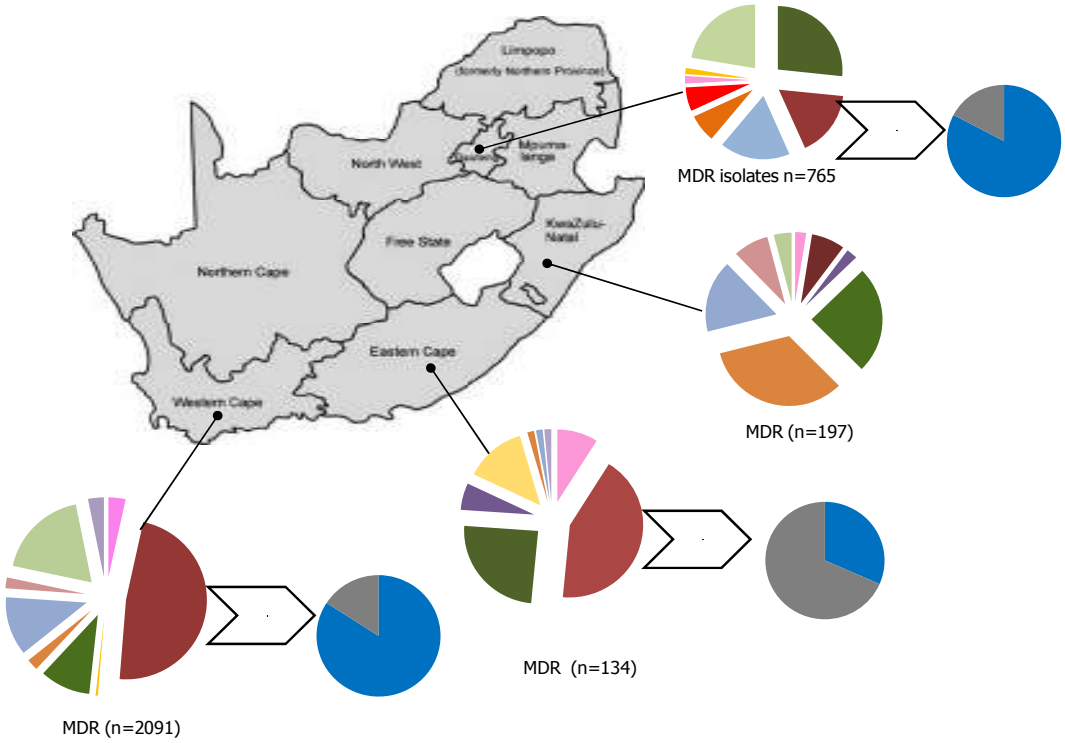


Figure 4: Distribution of *Mycobacterium tuberculosis* genotypes identified from MDR-TB isolates from Eastern Cape, Gauteng, KwaZulu-Natal and Western Cape provinces in South Africa. The blue and gray pie charts show the sub-classification of Beijing strains into typical Beijing strains and atypical Beijing strains respectively, for Eastern Cape (on all Beijing isolates n=57), Gauteng (on a subset of Beijing isolates n=81) and Western Cape (on a subset of Beijing isolates (n=189)

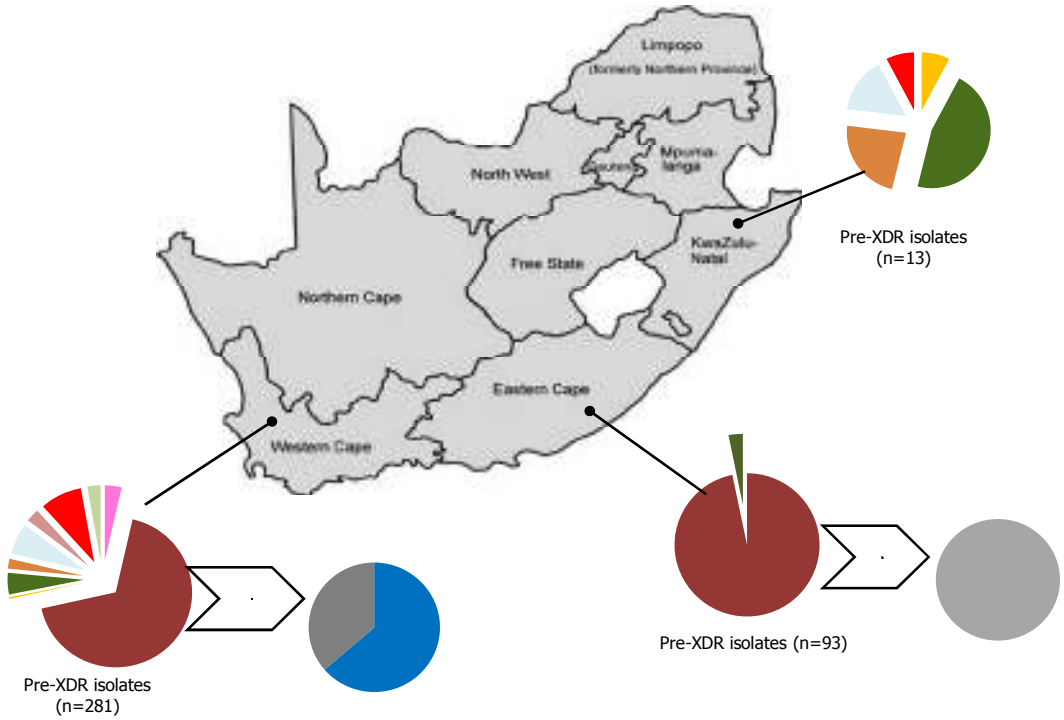


Figure 5: Distribution of *Mycobacterium tuberculosis* genotypes identified from Pre-XDR-TB isolates from Eastern Cape, KwaZulu-Natal and Western Cape provinces in South Africa. The blue and/or gray pie charts show the sub-classification of Beijing strains into typical Beijing strains and atypical Beijing strains respectively, from Western (on a subset of Beijing isolates n=66) and Eastern Cape (on all Beijing isolates n=90) into typical Beijing strains and atypical Beijing strains.

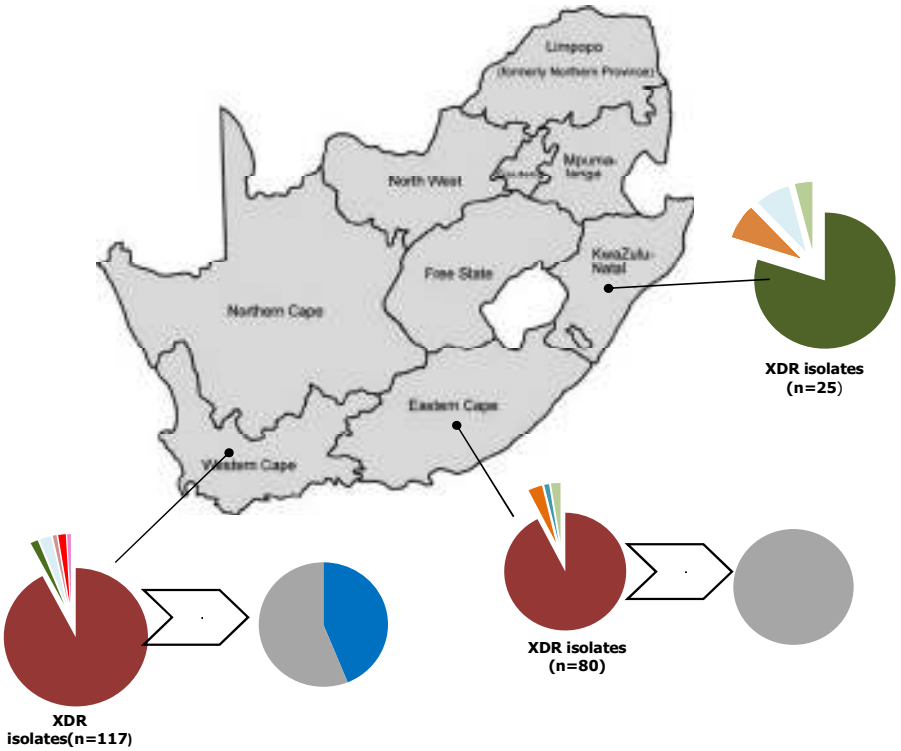


Figure 6: Distribution of *Mycobacterium tuberculosis* genotypes identified from XDR-TB isolates from Eastern Cape, KwaZulu-Natal and Western Cape provinces in South Africa. The blue and/or gray pie charts show the sub-classification of Beijing strains into typical Beijing strains and atypical Beijing strains respectively, from Eastern Cape (on all Beijing isolates n=74) and Western Cape (on a subset of Beijing isolates n=57) into typical Beijing and atypical Beijing strains.

Discussion

We have used genotyping data from four provinces in South Africa including, Gauteng, KwaZulu Natal, Western and Eastern Cape to study the phylogeography of drug resistant *M. tuberculosis*, to test the hypothesis that specific lineages and therefore distinct phylogeographic structure exists across the different provinces. Our results show a within country geographically localized population of *M. tuberculosis* strains that are associated with MDR-TB, pre-XDR-TB or XDR-TB. The population structure was clearly different in Eastern and Western Cape and KwaZulu-Natal provinces, differing in both the type and frequency of occurrence of each genotype. Distinct genotypes were associated with each of the three resistance classes and the province from which the patients originated.

The Beijing genotypes were predominantly associated with MDR-TB, pre-XDR-TB and XDR-TB in Eastern and Western Cape making up 72% and 52% respectively of all genotypes identified among the resistant isolates. The Beijing genotype could be classified further into the typical or atypical genotypes using a novel multiplex PCR method (13, 27). The results showed geographic localization at this level; with the atypical Beijing strains more likely to be isolated from MDR-TB, pre-XDR-TB and XDR-TB isolates from Eastern Cape than from the subset of isolates analyzed from Western Cape and Gauteng. We believe that the results from the subset from Western Cape and Gauteng were representative of the total isolates analyzed in this study. In an earlier study analyzing the genome structure *M. tuberculosis* Beijing isolates from rural and urban populations of Western Cape, an overabundance of the more recently evolving typical Beijing (sublineage 7) was observed (27). The typical Beijing sublineage 5, 6 and

7 were associated with MDR-TB, corroborating the findings among MDR-TB isolates from Western Cape in this study. However, in our study we found a trend towards increasing proportion of atypical Beijing strains that was also seen as isolates evolved from MDR-TB to XDR-TB in Western Cape. This could imply migration from Eastern Cape to Western Cape. A similar analysis could not be done for the isolates from KwaZulu Natal as these were not further characterized.

Previously typical Beijing strains have been defined as the more recently evolved strains that are well adapted to spread and cause disease given their frequency of occurrence when compared to the distantly evolved atypical Beijing strains (27, 30, 31). On the contrary, based on the frequency at which atypical Beijing strains occur in Eastern Cape Province, in this and other studies (14), the atypical Beijing genotypes could be well adapted to spread and cause disease in these communities. Further north in Gauteng province, though Beijing strains were also identified, the majority of strains identified among MDR-TB isolates belonged to the LAM genotype and these were predominantly LAM4 genotype. As with the Eastern and Western Cape, a variety of other genotypes were also identified from MDR-TB isolates in Gauteng but in lower proportions. Towards the North Eastern coast in KwaZulu-Natal a variety of genotypes were associated with MDR-TB but the S and LAM genotypes predominated and as in Gauteng the LAM4 was the predominant sublineage identified among the LAM strains. However, the proportions changed as the isolates evolved from MDR-TB to XDR-TB and the LAM4 genotypes predominated whereas the Beijing genotypes disappeared completely. From earlier studies the LAM4 genotype appeared to be unique to the KwaZulu Natal

Province and had a distinct spoligotype pattern and referred to as F15/LAM4/KZN (11). In this study however LAM4 strains were also identified from Gauteng among MDR-TB cases indicating that they may be more wide spread than initially thought.

The possible explanation for the disappearance of the Beijing strains among XDR-TB isolates in this population could be that they have a relatively lower ability to acquire resistance to second-line drugs. Alternatively they may have been introduced in this setting recently and are therefore less frequently transmitted as drug resistant strains. As a result they are less likely to become XDR-TB strains and there is a selection bias for the LAM genotypes. The LAM genotypes appear to have adapted well in the KwaZulu Natal and Gauteng population. Transmission may be an important driver for the occurrence of MDR-TB in all settings and it seems reasonable to assume that transmitted MDR strains much more often become XDR-TB strains than initially pan sensitive or non-MDR strains. The data suggests that within each region, sub-populations of *M. tuberculosis* genotypes have been selected which have now gone to select pre-XDR-TB and XDR-TB population suggestive of outbreaks. The reasons for this are not apparent but the populations seem to be able to escape second-line treatment regimens.

Our data shows that the genotypes and their frequency vary across the provinces. Within each province where we were able to describe genotypes among all three resistance classes, there was decreasing diversity of *M. tuberculosis* stains towards XDR-TB. The genotypes did not remain uniform but changed in proportion as isolated evolved to XDR-TB.

On the contrary, the observation of a wider variety of genotypes among MDR–TB isolates possibly reflects the underlying *M. tuberculosis* population structure in each province that we may expect to see even among drug susceptible strains. In the previous chapter (Chapter 5) we showed a similar population structure among drug sensitive strains. In that study avariety of genotypes were identified from South Africa with LAM (30.2%), Beijing (26.4%), X (11.5%), H (7.4%) and S (6.8%), being the most prevalent. MDR strains evolving from such drug susceptible strains during exposure to first line treatment would be expected to show similar diversity of strains with a few predominating. This diversity decreases as we move towards XDR-TB indicating that there must be a very successful group of MDR-TB strains that are able to transmit efficiently, despite additional drug resistance.

In this study we have shown that atypical Beijing and LAM 4 strains are more likely to be associated with XDR-TB in the Eastern/Western Cape and in KwaZulu-Nata, respectively. These strains may not necessarily transmit as XDR-TB strains but more likely as MDR. As second line drugs are added to the regimen, this results in acquisition of resistance to second line drugs. This notion is supported by the identification of different drug resistance mutations for second line drugs in the same clone (17).

A few other genotypes were also identified among XDR-TB isolates but were represented in very low proportions varying from 1.0-13.0%, probably showing that these are less efficient in transmitting XDR-TB in these provinces. In addition, if there are other MDR-TB strains that transmit in these settings they appear to be less efficient in acquiring

additional mutations and becoming XDR-TB strains than the atypical Beijing and LAM4 clones that are specific to either Eastern and Western Cape and KwaZulu Natal. However, it seems quite reasonable that a well standardized drug regimen that still offers opportunities for strains to transmit eventually creates an ecological niche that strains will adapt to.

Our study has some limitations. Data from the provinces was collected over different periods. The Western Cape Province dataset was collected over a period of 10 years (2000-2010) and DST to second-line drugs was only done from 2006 onwards. Therefore the population of XDR-TB isolates may be under-represented. In addition, the number of isolates contributed from the Western Cape was overrepresented when compared to Eastern Cape (2008-2009), Gauteng (2004-2007) and KwaZulu Natal (2005-2006). This together with the different levels of selection in the different provinces may have introduced a selection bias. The isolates from Gauteng were only tested for susceptibility to first-line drugs and MDR-TB isolates based on the definition used in this study may have been over-represented. A convenience sample of isolates was used to sub-classify Beijing strains from Western Cape and Gauteng; therefore we may have over or underestimated the proportions of atypical and typical Beijing strains. No data on this sub-classification was available from KwaZulu Natal and it was not possible to compare the Beijing strains from this province to those from other provinces.

Despite these observations our results provide evidence of few successful clones that are driving M/XDR TB. In this study, the observation that atypical Beijing strains were the

most prevalent genotypes among XDR strains in the Eastern and Western Cape and LAM4 genotype in KwaZulu Natal strongly suggests that these are very effective *M. tuberculosis* clones. These very effective clones including atypical Beijing and LAM4 genotypes have been shown to have an i*nh*A promoter mutation (6, 14, 17). A significant association between the *inh*A promoter mutations and XDR-TB has been shown in two South African provinces, Eastern and Western Cape which were part of this study (Muller B, personal communications). In this study the proportion of isolates with an *inh*A promoter mutation also increased significantly from MDR-TB to XDR-TB (data not shown). These observations together with the observations made in our study on the population structure of XDR-TB isolates suggest a selective advantage of *M. tuberculosis* strains harbouring an *inh*A promoter mutation to become XDR-TB strains

The predominance of certain genotypes could be due to geographical localization of genotypes before the implementation of standardized treatment regimens. The population that became effective in transmitting as drug-resistant strains, most likely, originated from the most prevalent drug sensitive group. However treatment practices in specific regions may also have contributed to the selection of specific lineages by suppressing wild type drug susceptible strains and selecting for these highly successful/adapted drug resistant clones. As soon as a group of strains has adapted to this new ecological niche, the frequencies were likely to change radically as the more adapted strains outcompeted less adapted strains and clustering of the adapted genotypes became more evident. Although other factors, such as HIV/AIDS may have contributed to the selection of specific lineages in the past two decades, the currently observed picture could possibly

also be explained without these other factors if we assume that geographical localization may have existed before the introduction of standardized regimens. The observation that certain drug-resistant genotypes can be present in high frequencies in specific populations despite having a heavy burden of drug resistant mutations may suggest that that they are more fit than initially presumed and may have specific but as yet unknown virulence factors.

In conclusion the population structure of drug-resistant TB is geographically localized to different South African regions and the diversity of *M. tuberculosis* decreases as isolates move from MDR-TB to XDR-TB. These observations put emphasis on the need to identify prevalent cases though active screening including close contacts of such cases. The identification of genetic markers in drug resistant *M. tuberculosis* strains is important in identifying cases at risk of developing M(X)DR-TB. The tools for diagnosing such cases already exist, for example the Genotype® MTBDR*plus* assay which detects the principal mutation in the *inh*A promoter conferring INH resistance that are closely associated with these very successful strains and mutations conferring RIF resistance. The assay has been implemented in several countries including some provinces in South Africa to assist culture based DST. The results from the assay should allow for rapid detection of high risk patients and adjustment of treatment regimens.

References

1. **World Health Organization (WHO).** Multidrug and extensively drug-resistant TB (M/XDR-TB). 2010 Global report on surveillance and response WHO/HTM/TB/2010.3 Geneva. Available from http://www.who.int/tb/publications/2010/978924599191/em/index.html accessed 21 September 2010

2. **Cohen T, Murray M**. Modelling epidemics of multidrug resistant *M. tuberculosis* of heterogeneous fitness. *Nat Med* 2004; 10:1117-1121.

3. **Borrell S, Gagneux S**. Infectiousness, reproductive fitness and evolution of drug resistant *Mycobacterium tuberculosis*. *Int J Tuberc Lung Dis*.2009; 13:1456-1466

4. **Petrini B, Hoffner S**. Drug-resistant and muti-drug resistant tubercle bacilli. *Int J Antimicrob Agents* 1999; 13: 93-7

5. **Victor TC, Streicher EM, Kewley C, Jordaan AM, van der Spuy GD, Bosman M, Louw H, Murray M, Young D, van Helden PD, Warren RM**. Spread of an emerging *Mycobacterium tuberculosis* drug-resistant strain in the western Cape of South Africa. *Int J Tuberc Lung Dis*. 2007 Feb;11(2):195-201.

6. **Johnson R, Warren RM, van der Spuy GD, Gey van Pittius NC, Theron D, Streicher EM, Bosman M, Coetzee GJ, van Helden PD, Victor T**C. Drug-resistant tuberculosis epidemic in the Western Cape driven by a virulent Beijing genotype strain. *Int J Tuberc Lung Dis* 2010; 14:119-121

7. **van Rie A, Warren RM, Beyers N, Gie RP, Classen CN, Richardson M, Sampson SL, Victor TC, van Helden PD**. Transmission of a Transmission of a multidrug-resistant *Mycobacterium tuberculosis* strain resembling "strain W" among noninstitutionalized, human immunodeficiency virus-seronegative patients. *J Infect Dis* 1999; 180: 1608-15

8. **Mlambo CK, Warren RM, Poswa X, Victor TC Duse AG, Marais E**. Genotypic diversity of extensively drug resistant tuberculosis (XDR-TB) in South Africa. *Int J Tuberc Lung Dis* 12:99-104

9. **Glynn JR, Whiteley J, Bifani PJ, Kremer K and van Soolingen D**. Worldwide occurrence of the Beijing/W strains of Mycobacterium tuberculosis: A Systematic Review. *Emerg Infect DIs* 2002; 8:843-849

10. **European concerted Action** on New Generation Genetic Markers and Techniques for the Epidemiology and control of Tuberculosis. Beijing /W Genotype *Mycobacterium tuberculosis* and drug resistance. *Emerg Infect Dis* 2006;12:736-742

11. **Pillay M and Sturm AW**. Evolution of the extensively drug resistant F15/LAM4/KZN strain of *Mycobacterium tuberculosis* in KwaZulu Natal, South Africa. *Clin Infect Dis* 2007;45:1409–1414

12. Mokrousov I, Jian WW, Sun GZ, Liu JW, Valcheva V, Li M, Narvaskaya O, Shen DA. Evolution of drug resistance in different sublineages of *Mycobacterium tuberculosis* Beijing genotype. *Antimicrob Agents and Chemotherapy* 2006; 50: 2820-2823

13. **Johnson R, Warren R,, Strauss OJ, Jordaan AM, Falmer AA, Beyers N, Schaaf HS, Murray M, Cloete K, van Helden PD, Victor TC**. An outbreak of drug-resistant tuberculosis caused by a Beijing strain in the Western Cape, South Africa. *Int J Tuberc Lung Dis* 2006; 10: 1412-1414

14. **Strauss OJ, Warren RM, Jordaan A, Streicher EM, Hanekom M, Falmer AA, Albert H, Trollip A, Hoosain E, van Helden PD, Victor TC**. Spread of a low-fitness drug resistant *Mycobacterium tuberculosis* strain in a setting of high Human Immunodeficieny virus prevalence. *J Clin Micro* 2008; 46:1514-1516

15. **Bifani PJ, Mathema B, Kurepina NE, Kreiswirth BN**. Global dissemination of the *Mycobacterium tuberculosis* W-Beijing family strains. *Trends Microbiol* 2002; 10:45-52.

16. **Rad ME, Bifani P, Martin Cm, Kremer K, Samper S, Rauzier J, Kreiswirth B, Blazquez J, Jouan M, van Soolingen D, Gicquel B**. Mutations in putative mutator genes of *Mycobacterium tuberculosis* strains of the W-Beijing family. *Emerg Infect Dis* 2003; 9: 838-845.

17. **Ioerger TR, Koo S, No E, Chen X, Larsen MH, Jacobs WR, Pillay M, Strum AW, Sacchettini JC**. Genome analysis of multi-extensively-drug-reseistant tuberculosis from KwaZulu-Natal, South Africa. *Plos One* 2009;4:e7778. Doi:10.1371/journal.pone.0007778

18. **Hirsh A, Tsolaki A, DeRiemer K, Feldman MW, Small P**. Stable association between stains of *Mycobacterium tuberculosis* and their human host populations. *PNAS* 2004;101: 4871-4876

19. **Gagneaux 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 P**. Variable host-pathogen compatibility in *Mycobacterium tuberculosis*. *PNAS* 2006; 103:2869-2873

20. **Gagneaux S and Small P**. Global phylogeography of *Mycobacterium tuberculosis* and the implications for tuberculosis product development. *Lancet Infect Dis* 2007;7:328-37

21. World Health Organization. Policy guidance on drug-susceptibility testing (DST) of second-line antituberculosis drugs. Avaialble at: http://www.stoptb.org/cb/meetings/20081028_Bagamoyo_Tanzania/assets/documents/2.08-11.3%20DST%20policy%20.pdf. Accessed February 2011

22. **Springer, B., K. Lucke, R. Calligaris-Maibach, C. Ritter, and E. C. Bottger**. 2009. Quantitative drug susceptibility testing of Mycobacterium tuberculosis by use of MGIT 960 and EpiCenter instrumentation. J.Clin.Microbiol. **47**:1773-1780.

23. **World Health Organization**. WHO media report. WHO Global Task Force outlines measures to combat XDR-TB worldwide. Geneva Switzerland: WHO 2006. http://www.who.int/mediacentre/news/notes/2006/np29/en/index.html. Accessed 4 August 2010

24. **Banerjee R, Allen J, Westenhouse J, Oh P, Elms W, Desmond E, Nitta A, Royce S, Flood J.** Extensively drug-resistant tuberculosis in California 1993-2006. *CID* 2008; 47:450-456

25. **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(4):907-14.

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

27. **Hanekom M, van der Spuy GD, Gey van Pittius NC, McEvoy CRE, Hoek KGP, Ndabambi SL, Jordaan AM, Victor TC, van Helden PD, Warren RM**. Discordance between Mycobacterial interspersed repetitive-unit-variable-number tandem repeat typing and IS*6110* restriction fragment length polymorphism genotyping for analysis of *Mycobacterium tuberculosis* Beijing strains in a setting of high incidence tuberculosis. *J. Clin Microbiol* 2008; 46: 3338-3345

28. **Brudey K, Driscoll JR, Rigouts L, Prodinger WM, Gori A, Al-Hajoj SA, Allix C, Aristimuño L, Arora J, Baumanis V, Binder L, Cafrune P, Cataldi A, Cheong S, Diel R, Ellermeier C, Evans JT, Fauville-Dufaux M, Ferdinand S, Garcia de Viedma D, Garzelli C, Gazzola L, Gomes HM, Guttierez MC, Hawkey PM, van Helden PD, Kadival GV, Kreiswirth BN, Kremer K, Kubin M, Kulkarni SP, Liens B, Lillebaek T, Ho ML, Martin C, Martin C, Mokrousov I, Narvskaïa O, Ngeow YF, Naumann L, Niemann S, Parwati I, Rahim Z, Rasolofo-Razanamparany V, Rasolonavalona T, Rossetti ML, Rüsch-Gerdes S, Sajduda A, Samper S, Shemyakin IG, Singh UB, Somoskovi A, Skuce RA, van Soolingen D, Streicher EM, Suffys PN, Tortoli E, Tracevska T, Vincent V, Victor TC, Warren RM, Yap SF, Zaman K, Portaels F, Rastogi N, Sola C**. *Mycobacterium, tuberculosis* complex genetic diversity: mining the fourth international spoligotyping database (SpolDB4) for classification, population genetics and epidemiology. *BMC Microbiology* 2006;6:23

29. **Richardson M, van Lill SW, van der Spuy GD, Munch Z, Booysen CN, Beyers N, van Helden PD, Warren RM**. Historic and recent events contribute to the disease dynamics of Beijing-like *Mycobacterium tuberculosis* isolates in a high incidence region. *Int J Tuberc Lung Dis* 2002; 6(11):1001-1011.

30. **Mokrousov L, Narvskaya O, Otten T, Vyazovaya E, Limeschenko E, Steklova L and Vyshnevskyi B**. Phylogenetic reconstruction within *Mycobacterium tuberculosis* Beijing genotype in north western Russia. *Res Microbiol.* 2002; 153:629-637.

31. **Toungoussova OS, Sandven P, Mariandyshev AO, Nizovtseva NI, Bjune G and Caugant DA.** Spread of drug resistant *Mycobacterium tuberculosis* strains of te Beijing genotype in Archangekl Oblast Russia. *J. Clin Microbiol*. 2002; 40: 1930-1937

Conclusions

Tuberculosis (TB) has affected mankind for thousands of years. Despite the noteworthy progress made in understanding the disease over the past 100 years, the global incidence of TB continues to increase especially in Southern Africa. This is largely due to poor TB control programmes and the concurrent pandemic of HIV. Together these have gradually enabled the emergence of multidrug-resistant TB with the concomitant evolution of extensively drug-resistant strains of *M. tuberculosis* strains.

In recent years the use of molecular biological techniques has confirmed that the genotypic and phenotypic characteristics of *M. tuberculosis* are associated with the epidemiology of TB. These findings have highlighted the need for the development of new tools for preventing, diagnosing and treating TB. Molecular epidemiological tools have recently been used in a number of different settings in Southern Africa.

As discussed in Chapter 1, several lessons have been learnt from molecular epidemiology studies in South Africa. Major lessons include information on the diversity and frequency of occurrence of *M. tuberculosis* strains in these communities. Through this, some strain genotypes, including Beijing, LAM and LCC, have been shown to play a major role in the epidemic, both locally and globally, suggesting that these strain genotypes have evolved unique properties. Not only have they allowed an understanding of strain diversity but also the spread of both drug sensitive and drug resistant strains.

In Chapter 2, using both spoligotyping and IS*6110*-RFLP, the diversity of *M. tuberculosis* strains in Zimbabwe has been described. To date only two manuscripts have reported molecular epidemiological data from this country. Our study identified a sub-group of LAM genotype strains which was over-represented in this region. This genotype was termed the SAF1 genotype. Subsequently, a study in Brazil confirmed the importance of this genotype in Rio de Janeiro. We have recently renamed the strain SAF1/RD^{Rio} due to its characteristic region of difference. We also confirmed the occurrence of this family of strains further North in Zambia at relatively similar proportions. This study provided evidence that the burden of TB in central parts of Southern Africa is largely due to a small subset of actively circulating strains. The restricted strain diversity observed may be interpreted to reflect ongoing transmission or that this strain genotype is endemic as it was one of the first genotypes to be introduced into Southern Africa. If the latter is correct this could significantly restrict the utility of IS*6110*-RFLP in these regions.

In chapter 2a we give a short description of the geographical distribution of TB cases in Gweru (one of the cities in Zimbabwe). A subset of *M. tuberculosis* isolates genotyped as part of the study in Chapter 2, were also studied to determine their geographic distribution. The predominant SAF1/RD^{Rio} strains were not clustered in one particular district in the city, further supporting their role in driving the TB epidemic in these settings.

Interestingly, comparison of the strain population structure with other settings in South Africa clearly demonstrates regional differences, however risk factors for disease due to SAF1/RD^{Rio} strains are not known. We postulated host pathogen compatibility. In Chapter 3 potential risk factors for disease due to SAF1/RD$^{\mathbf{Rio}}$ strains were explored. Our study also showed an

association with smoking. Smoking has long been associated with TB and interacts with this disease by increasing disease susceptibility and worsening outcomes. This study showed an association between disease due to SAF1/RD^{Rio} strains and cavitary pulmonary TB in all TB patients regardless of HIV status, but not among HIV co-infected patients. The association with cavitary disease suggested that this genotype induced a more severe form of disease in the Zimbabwean population. This confirms a previous finding which demonstrated a similar association in Brazil.A Further observation included a trend towards an association with smear positive disease, which suggests high levels of transmission and that early diagnosis and treatment are important in controlling TB in this community.

In Chapter 4 we describe the genetic relationships among Low Copy Clade (LCC) strains identified in Cape Town in comparison to those from Zimbabwe and other regions using IS*6110*-RFLP, principal genetic grouping, spoligotyping, IS*6110* insertion site mapping and variable-number tandem repeats (VNTR). At the time the LCC strains were a poorly defined genotype and it was not known whether LCC strains were genotypically related. Our analysis showed that LCC strains from Cape Town were evolutionary close to LCC strains from Zimbabwe and other regions. These findings showed that these strains were distributed on a wider scale, probably reflecting spread from a common source during colonization of the African continent.

Specific *M. tuberculosis* genotypes have been associated with particular human populations. Molecular epidemiological data has clearly demonstrated regional differences, suggesting host pathogen compatibility (Chapter 2). In the past paucity of data from Southern Africa region has prevented such comparisons to be made regionally. In Chapter 5 we report the first study to

analyse the population structure of *M. tuberculosis* in 8 of the 14 countries in Southern Africa, using spoligotyping and/or IS*6110*-RFLP data. This study showed that distinct *M. tuberculosis* genotypes are associated with specific host populations, further supporting the notion of host pathogen compatibility. As the different genotypes are likely to have different phenotypes this may affect the development of new diagnostics, vaccines and drugs

Realizing the importance of drug-resistant TB in Southern Africa, Chapter 6 builds on the findings in chapter 5 where we use a similar methodology to describe the distribution of drug-resistant *M. tuberculosis* genotypes in provinces in South Africa. This is the first study to describe the population structure MDR-, pre-XDR- and XDR-TB on this scale. We showed that MDR-, pre-XDR- and XDR-TB genotypes were geographically localized across the four provinces. The Beijing genotype is particularly prevalent in the Western and Eastern Cape, with the typical Beijing genotype being predominantly associated with MDR-, pre-XDR-TB and XDR-TB in Western Cape, and atypical Beijing genotype with pre-XDR- and XDR-TB in the Eastern Cape. Further north in Gauteng and to the North East in KwaZulu Natal the LAM4 genotypes were more prevalent and the proportion of these strains increased among XDR-TB strains in KwaZulu Natal.

Our findings clearly demonstrate that MDR-TB is associated with a high number of genotypes probably reflecting both acquisition and transmission. Of significant importance is our observation that the diversity of XDR-TB genotypes is restricted in three of the four provinces. This suggests that there is a strong selection for certain genotypes to evolve to XDR-TB. Our present hypothesis is that these genotypes are at an advantage when exposed to second-line

treatment regimens. These results support the need for identifying genetic markers in drug resistant *M. tuberculosis* which can aid in diagnostic tests. These tests should allow the simple identification of cases at high risk of developing XDR-TB.

Collectively our findings show the importance of using standardized molecular biological techniques to study the epidemiology of TB in the Southern Africa region. Our findings show that different genotypes predominate in different geographic settings implying that these strains could be well adapted in these populations. The predominance of closely related strains implies that the bacterial population is fairly homogenous suggesting recent transmission of disease. These findings have major implications for the development of vaccines, new diagnostics and TB drugs. In vaccine and anti-TB drug design, different bacterial populations that are predominant in different geographic settings should be targeted and strain diversity should be considered in the development of diagnostic test. Genetic markers should be identified particularly among M(X) DR-TB strains to aid in identifying individuals at risk of developing drug resistant disease.

Candidate's contributions

Chapter 1: Molecular Epidemiology of Tuberculosis in South Africa: Lessons learnt

- Primary author

Chapter 2: Predominance of a single genotype of *Mycobacterium tuberculosis* in regions of Southern Africa

- Planning of the genotyping of isolates
- Spoligotyping and IS*6110* DNA fingerprinting of isolates
- Construction of genotyping database
- Analysis of IS*6110* DNA fingerprints using Gelcompar
- Comparison of spoligotypes to the international spoligotyping database
- Data analysis
- Writing up of chapter

Chapter 2b: Geographical distribution of *Mycobacterium tuberculosis* strains and pulmonary tuberculosis cases notified during one year in Gweru, Zimbabwe

- Genotyping of isolates

Chapter 3: Tuberculosis due to *Mycobacterium tuberculosis* "SAF1/RD^{Rio}" strains is associated with smoking and cavitary lung disease:

- Planning of the genotyping of isolates
- Spoligotyping and IS*6110* DNA fingerprinting of isolates
- Construction of genotyping database
- Analysis of IS*6110* DNA fingerprints using Gelcompar
- Comparison of spoligotypes to the international spoligotyping database
- Data analysis

Writing up of chapter

Chapter 4: Clonal expansion of a globally disseminated lineage of *Mycobacterium tuberculosis* with low IS*6110* copy numbers.

- Genotyping of isolates
- Co-author

Chapter 5: Distribution of *Mycobacterium tuberculosis* genotypes varies across regions of Southern Africa with high TB incidence

- Conceptualizing the study
- Spoligotyping and IS*6110* DNA fingerprinting of some isolates South Africa and Zimbabwe
- Comparison of spoligotypes to the international spoligotyping database
- Data analysis
- Writing up of chapter

Chapter 6: Population structure of multidrug-resistant and extensively drug-resistant tuberculosis in South Africa

- Conceptualizing the study
- Data analysis
- Writing up of chapter

Supplemental table 1: Factors associated with disease due to *Mycobacterium tuberculosis* SAF1/RD^{Rio} strains among all culture positive TB patients: Univariable analysis

	Total	Outcome: *Mycobacterium tuberculosis* SAF1/RD^{Rio} strains		
	Total (All cases)	All culture positive TB cases (row %)	*Unadjusted OR (95% CI)	**P value
All patients	n=134	n=69		
Age (years)				
18-24	16	8(50.0)	1	0.24
25-29	41	18(42.9)	0.71(0.22-2.26)	
30-34	30	15(50.0)	1.0 (0.30-3.37)	
35-39	16	12(75.0)	3.0 (0.67-13.4)	
≥40	30	16(53.3)	1.14(0.34-3.84)	
Gender				
Male	75	39 (52.0)	1	0.89
Female	59	30 (50.9)	0. 95(0.48-1.88)	
TB treatment history				
No	112	58(51.8)	1	0.38
Yes	22	11(50.0)	0.93(0.37-2.32)	
HIV status				
Positive	119	58(48.7)	1	0.06
Negative	15	11(73.3)	2.89(0.87-9.59)	
Household TB contact				
No	40	17(42.5)	1	0.17
Yes	94	52(55.3)	1.67(0.79-3.54)	
Currently/ever smoked				
No	93	43(46.2)	1	0.06
Yes	41	26(63.4)	2.01(0.94-4.29)	
Smoking				
No	93	43(46.2)	1	0.11
Currently smoking	13	8 (61.5)	1.86(0.57-6.11)	
Ever smoked	20	14(70.0)	2.71(0.96-7.67)	
Exposure to dust or irritants				
No				
Yes	109	60(55.1)	1	0.12
	24	9(37.5)	0.49(0.20-1.21	

*Unadjusted Odds ratio from logistic regression representing the summary relative odds of having disease due to *Mycobacterium tuberculosis* Southern Africa 1 (SAF1/RD^{Rio}) strains compared with other *Mycobacterium tuberculosis* strains.
CI confidence interval; **P value for Univariable association

Supplemental table 2: Factors associated with disease due to *Mycobacterium tuberculosis* SAF1/RD^{Rio} strains, among HIV positive TB patients with culture positive disease-univariable analysis

	Outcome: *Mycobacterium tuberculosis* SAF1/RD^{Rio} strains			
	HIV positive TB cases	HIV positive n (row %)	*Unadjusted OR (95% CI)	**P value
All patients	n=119	n=58		
Age (years)				
18-24	13	6(46.2)	1	0.56
25-29	38	16(42.1)	0.85(0.24-3.01)	
30-34	28	14(50.0)	1.17(0.31-4.35)	
35-39	13	9(69.2)	2.63(0.53-13.07)	
≥40	27	13(48.2)	1.08(0.29-4.08)	
Gender				
Male	65	32(49.2)	1	0.90
Female	54	26(48.2)	0.96(0.47-1.97)	
TB treatment history				
No	97	47(48.5)	1	0.90
Yes	22	11(50.0)	1.06(0.42-2.68)	
Household TB contact				
No	36	13(36.1)	1	0.07
Yes	83	45(54.2)	2.09(0.94-4.69)	
Currently/ever smoked				
No	82	35(42.7)	1	0.04
Yes	37	23(62.2)	2.21(0.99-4.88)	
Smoking				
No	82	35(42.7)	1	0.10
Currently smoking	11	7 (63.6)	2.35(0.64-8.66)	
Ever smoked	18	12(66.7)	2.69(0.92-7.86)	
Exposure to dust or irritants				
No	96	51(53.1)	1	0.07
Yes	22	7(31.8)	0.41(0.15-1.10)	

*Unadjusted Odds ratio from logistic regression representing the summary relative odds of having disease due to *Mycobacterium tuberculosis* Southern Africa 1 (SAF1/RD^{Rio} strains compared with other *Mycobacterium tuberculosis* strains; CI confidence interval; **P value for Univariable association

Supplemental table 3: Factors associated with cavitary tuberculosis disease among culture positive TB patients- Univariable analysis

	Total	Outcome: Cavitary TB disease		
	Total (All cases)	All culture positive TB cases (row %)	Unadjusted OR (95% CI)	P value
All patients	n=134	n=42		
Age (years)				
18-24	16	10(62.5)	1	0.09
25-29	40	10(25.0)	0.20(0.58-0.69)	
30-34	30	9(30.0)	0.26(0.07-0.92)	
35-39	16	4(25.0)	0.20(0.04-0.91)	
≥40	30	8(26.7)	0.22(0.06-0.80)	
Gender				
Male	74	20(27.0)	1	0.20
Female	59	22(37.3)	1.61(0.77-3.35)	
Strain family				
Other	65	15(23.1)	1	0.04
SAF1/RD^{Rio}	68	27(39.7)	2.20(1.03-4.67)	
TB treatment history				
No	111	35(31.5)	1	0.98
Yes	22	7 (31.8)	1.01(0.38-2.71)	
HIV status				
Positive	118	34(28.8)	1	0.06
Negative	15	8 (53.3)	2.82(0.95-8.40)	
Household TB contact				
No				
Yes	40	13(32.5)	1	0.88
	93	29(31.2)	0.94(0.42-2.08)	
Currently/ever smoked				
No	93	31(33.3)	1	0.50
Yes	40	11(27.5)	0.76(0.33-1.71)	
Smoking				
No	93	31(33.3)	1	0.61
Currently smoking	13	3 (23.1)	0.60(0.15-2.33)	
Ever smoked	20	5 (25.0)	0.66(0.22-2.00)	
Exposure to dust or irritants				
No	108	36(33.3)	1	0.42
Yes	24	6 (25.0)	0.67(0.24-1.82)	

*Unadjusted odds ratio from logistic regression representing the summary relative odds of having cavitary tuberculosis; CI confidence interval; **P value for Univariable association

Supplemental table 4: Factors associated with cavitary tuberculosis disease, among HIV positive TB patients with culture positive disease –Univariable analysis

	Total	Outcome: Cavitary TB disease		
	Total (All cases)	HIV positive TB cases (row %)	Unadjusted OR (95% CI)	P value
All patients	n=134	n=134		
Age (years)				
18-24	13	8 (61.5)	1	0.12
25-29	37	9 (24.3)	0.20(0.05-0.77)	
30-34	28	8(28.6)		
35-39	13	3(23.1)		
≥40	27	6(22.2)		
Gender				
	64	16(25.0)	1	0.31
Male	54	18(33.3)	1.50(0.67-3.33)	
Female				
Strain family				
Other	61	15(24.6)	1	0.29
SAF1/RD^{Rio}	57	19(33.3)	1.53(0.69-3.41)	
TB treatment history				
No	96	27(28.1)	1	0.73
Yes	22	7 (31.8)	1.19(0.44-3.25)	
Household TB contact				
No				
Yes	36	9 (25.0)	1	0.54
	82	25(30.5)	1.31(0.54-3.20)	
Currently/ever smoked				
No				
Yes	82	25(30.5)	1	0.54
	36	9 (25.0)	0.76(0.31-1.85)	
Smoking				
No	82	25(30.5)	1	0.57
Currently smoking	11	2 (18.2)	0.51(0.10-2.51)	
Ever smoked	18	4 (22.2)		
Exposure to dust or irritants				
No				
Yes	95	30(31.6)	1	0.20
	22	4 (18.2)	0.48(0.14-1.55)	

*Unadjusted Odds ratio from logistic regression representing the summary relative odds of having cavitary tuberculosis among HIV positive TB patients; CI confidence interval; **P value for Univariable association

Supplemental table 5: Resistance profiles of *M. tuberculosis* isolates

Province	Resistance class								
	InhMono n (%)	RifMono n (%)	**MDR n (%)**	**PreXDR n (%)**	**XDR n (%)**	*Other n (%)	Sensitive n(%)	Unclassified n(%)	Total
Eastern Cape	0 (0.0)	0 (0.0)	**134 (24.3)**	**93 (16.9)**	**80 (14.5)**	0 (0.0)	245 (44.4)	0 (0.0)	552 (100.0)
Gauteng	0 (0.0)	0 (0.0)	**765 (100.0)**	**0 (0.0)**	**0 (0.0)**	0 (0.0)	0 (0.0)	0 (0.0)	765 (100.0)
KwaZulu Natal	11 (2.8)	6 (1.6)	**196 (50.5)**	**12 (3.1)**	**25 (6.4)**	13 (3.3)	122 (31.4)	3(0.8)	388 (100.0)
Western Cape	1433 (31.2)	424 (9.2)	**2091 (45.6)**	**281 (6.1)**	**117 (2.6)**	68 (1.5)	24 (0.5)	151 (3.3)	4589 (100.0)
**Other	28 (23.3)	8 (6.7)	**68 (57.1)**	**7 (5.8)**	**3 (3.5)**	1 (0.80	1 (0.8)	3 (2.5)	119 (100.0)
Total	1472(23.0)	438 (6.8)	**3255 (50.8)**	**393 (6.1)**	**225 (3.5)**	82 (1.3)	392 (6.1)	157 (2.5)	6413 (100.0)

*Other- isolates in this drug resistant category were either mono resistant to other first line drugs or were poly resistant (streptomycin mono resistant (n=3); Ofloxacin mono resistant (n=1); Amikacin mono resistant (n=1); Poly resistant (n=81)

**isolates from other provinces that were excluded from the analysis, Northern Cape (n=9); North West (n=17); Limpopo (n=1); Free State (n=1); Unknown (n=91)