MOLECULAR EPIDEMIOLOGY OF TUBERCULOSIS IN THE MIDLANDS

BY

JASON THOMAS EVANS

A thesis submitted to
The University of Birmingham
for the degree of
DOCTOR OF PHILOSOPHY

School of Immunity and Infection,
College of Medical and Dental Sciences,
The University of Birmingham,
Edgbaston, B15 2TT, United Kingdom.

July 2011

Abstract

Data from this thesis has extended our understanding of the molecular epidemiology and transmission of *Mycobacterium tuberculosis* in the Midlands. A novel DNA fingerprinting method called Mycobacterial Interspersed Repetitive Units containing Variable Number Tandem Repeats (MIRU-VNTR) typing provided equivalent results when compared to the current gold standard for DNA fingerprinting (IS*6110* RFLP). To improve our understanding of TB in the Midlands, MIRU-VNTR typing was then developed to be assayed by non-dHPLC for the first time. Using this high-throughput rapid method a prospective and universal typing study was undertaken. This work identified the predominance of the Euro-American and East African Indian global clades in the Midlands and linked them to particular human population groups using novel software based on names. DNA fingerprinting also discovered the most prevalent single strain in the Midlands. This strain is geographically restricted to the West Midlands within the UK and globally. From this geographical association, we have called this strain the "Mercian" strain. The Mercian strain was not associated with patients who originated from the Indian Sub-Continent but was significantly associated with UK-born, Black Caribbean patients in Wolverhampton. These findings show that strains have been imported into the Midlands from around the world and there has also been continued transmission of these and other strains which may have been present in the Midlands for years. Molecular tools developed in this thesis will have regional, national, and international impact on TB control.

Dedication

To Susan and Skye

Acknowledgements

I would like to personally thank my PhD supervisor Professor Peter Hawkey and Dr. Grace Smith for supervising my work during the course of this thesis. I would also like to thank Sarah Gardiner, Kerstin Boese, Shaina Khanom and all of the staff in the HPA Midlands Regional Centre for Mycobacteriology for carrying out DNA fingerprinting on a prospective routine basis.

Epidemiological data was collected by staff in HPA Health Protection Units across the Midlands including: Philip Monk (HPA East Midlands South Health Protection Unit, Leicester, UK); Gerry Bryan (Leicester Health Authority, Leicester, UK); Huda Mohamed (HPA West Midlands East Health Protection Unit, Birmingham, UK); Madhu Bardhan (HPA West Midlands East Health Protection Unit, Birmingham, UK), Nicholas Pugh (NHS Walsall, West Midlands, UK); and Beck Taylor (Department of Public Health, Epidemiology and Biostatistics, University of Birmingham, Birmingham, UK). Implementation of MIRU-VNTR typing using non-dHPLC was greatly aided by Dr George Hong (Transgenomic, Omaha, Nebraska, USA). Origins software was originally provided by Professor Richard Webber (King's College London, London, UK). Geographical mapping of the Mercian strain was provided by Helen Bagnall and Jianxia Sun (HPA West Midlands Regional Surveillance Unit, Birmingham, UK). Epidemiological investigation of the Mercian strain in Wolverhampton was carried out by Dr. Robert Serafino Wani (Department of Infection, Guy's and St Thomas' NHS Foundation Trust, London, United Kingdom), Dr. Helen Jones (Department of Microbiology, New Cross Hospital, Wolverhampton, UK), and Dr. Jonathan Mann (Department of Respiratory Medicine, New Cross Hospital, Wolverhampton, UK). Dr. Laura Anderson (Respiratory Diseases Department - Tuberculosis Section, HPA Health Protection Services - Colindale, London, UK) provided the West Midlands regional data and carried out statistical analysis of the Wolverhampton and West Midlands epidemiological datasets. Dr. Annette Wood (Department of Public Health, Epidemiology and Biostatistics, School of Health and Population Sciences College of Medical and Dental Sciences, University of Birmingham, Edgbaston, Birmingham, UK), Dr. Babatunde Olowokure (HPA West Midlands Regional Epidemiology Unit, Birmingham, UK), Dr. Ibrahim Abubakar (Respiratory Diseases Department - Tuberculosis Section, HPA Health Protection Services - Colindale, London, UK) and Dr. Pam Sonnenberg (Department of Infection & Population Health, Centre for Sexual Health & HIV Research, University College London, London, UK) provided epidemiological advice and leadership. Spoligotyping was developed for a prior MRes project in collaboration with Helen Smith (School of Immunity and Infection, University of Birmingham, UK). Protocols and MS Excel analysis files were provided by Dr. Lauren Cowan (CDC, Atlanta, USA). Spoligotyping and 24 loci typing of Mercian strains were carried out in collaboration with Dr. Andrea Gibson (HPA MRCM, Birmingham). DNA microarrays were carried out under the supervision of Carmen Menéndez, Roger Buxton, and Joseph Colston when I visited the MRC National Institute for Medical Research, London, UK.

Procurement of laboratory equipment for prospective DNA fingerprinting was funded by the Department of Health. Spoligotyping and 24 loci typing of the Mercian strain were funded by a HPA Strategic R&D Grant.

Finally, I would like to thank my parents and family for supporting me throughout this PhD.

Table of Contents

List of Figures

List of Tables

Abbreviations

AFB	Acid Fast Bacilli
ATCC	American Type Culture Collection
BCG	Bacillus Calmette-Guèrin
bp	Base Pair
CAS	Central Asian Strain
CCDC	Consultant in Communicable Disease Control
CDC	Centers For Disease Control
cDNA	Complementary Deoxyribonucleic Acid
CEL	Culture, Ethnic, and Linguistic
CI	Confidence Interval
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
CTAB	Cetyl Trimethylammonium Bromide
dCTP	Deoxycytidine Triphosphate
dHPLC	denaturing High-Performance Liquid Chromatography
DNA	Deoxyribonucleic Acid
dNTP	Deoxyribonucleotide Triphosphate
DR	Direct Repeat
EAI	East African Indian
EDC	1-Ethyl-3-[3-dimethylaminopropyl]carbodiimide hydrochloride
EDTA	Ethylenediaminetetraacetic Acid
ELISA	Enzyme-Linked Immunosorbent Assay
ESAT	Early Secretory Antigenic Target
ETR	Exact Tandem Repeat
fAFLP	Fluorescent Amplified Fragment Length Polymorphism
GIS	Geographic Information System
HD	Highly Discriminatory
HGDI	Hunter-Gaston Discrimination Index
HIV	Human Immunodeficiency Virus
HPA	Health Protection Agency
HPU	Health Protection Unit
IGRA	Interferon-Gamma Release Assay
IS	Insertion Sequence
ISC	Indian Subcontinent
LAM	Latin American Mediterranean
LJ	Lowenstein Jensen
LA	Local Authority
LSP	Large Sequence Polymorphism
MDR-TB	Multidrug-Resistant Tuberculosis
MES	Morpholineethanesulfonic Acid
MGIT	Mycobacteria Growth Indicator Tube

MIRU	Mycobacterial Interspersed Repetitive Unit
MIRU-VNTR	Mycobacterial Interspersed Repetitive Units containing Variable Number Tandem Repeat
MLST	Multilocus Sequence Typing
MPTR	Major Polymorphic Tandem Repeat
MRC	Medical Research Council
MRCM	Midlands Regional Centre for Mycobacteriology
MTBC	*Mycobacterium tuberculosis* Complex
NHS	National Health Service
NTM	Nontuberculous Mycobacteria
OADC	Oleic Albumin Dextrose Catalase
ORF	Open Reading Frame
PANTA	Polymixin B, Amphotericin B, Nalidixic acid, Trimethoprim, Azlocillin
PCR	Polymerase Chain Reaction
PCT	Primary Care Trust
PE	Proline (P) Glutamic acid (E)
PGRS	Polymorphic GC-rich Repetitive Sequence
PGG	Principal Genetic Group
PPE	Proline (P) Proline (P) Glutamic acid (E)
PS-DVB	Polystyrene-Divinylbenzene
QUB	Queen's University Belfast
RD	Region of Difference
RFLP	Restriction Fragment Length Polymorphism
RNA	Ribonucleic acid
RNase	Ribonuclease
rRNA	Ribosomal Ribonucleic acid
rDNA	Ribosomal Deoxyribonucleic Acid
SA	South Asian
SDS	Sodium Dodecyl Sulphate
SET	Sucrose Ethylenediaminetetraacetic Acid Tris
SI	South Indian
SNP	Single Nucleotide Polymorphism
SSC	Saline Sodium Citrate
TB	Tuberculosis
TBE	Tris Borate Ethylenediaminetetraacetic Acid
TE	Tris Ethylenediaminetetraacetic Acid
TEAA	Triethylammonium Acetate
TIFF	Tagged Image File Format
TMAC	Tetramethylammonium Chloride
TST	Tuberculin Skin Test
UK	United Kingdom
UPGMA	Unweighted Pair Group Method using Arithmetic Averages

UV	Ultraviolet
VNTR	Variable Number Tandem Repeats
WGS	Whole genome sequencing
WT	Wildtype
XDR-TB	Extensively Drug-Resistant Tuberculosis

Publications Arising From This Thesis

Chapter 3

Hawkey, P.M., Smith, E.G., Evans, J.T., Monk, P., Bryan, G., Mohamed, H.H., Bardhan, M., & Pugh, R.N. (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.

Chapter 4

Evans, J.T., Hawkey,P.M., Smith,E.G., Boese,K.A., Warren,R.E., & Hong,G. (2004) Automated high-throughput mycobacterial interspersed repetitive unit typing of *Mycobacterium tuberculosis* strains by a combination of PCR and nondenaturing high-performance liquid chromatography. *J Clin Microbiol* **42**: 4175-4180.

Chapter 5

Evans, J.T., Gardiner,S., Smith,E.G., Webber,R., & Hawkey,P.M. (2010) Global Origin of *Mycobacterium tuberculosis* in the Midlands, UK. *Emerg Infect Dis* **16**: 542-545.

Chapter 6

Evans, J.T., Serafino Wani, R.L., Anderson, L., Gibson, A.L., Smith, E.G., Wood, A., Olowokure, B., Abubakar, I., Mann, J.S., Gardiner, S., Jones, H., Sonnenberg, P., & Hawkey, P.M. (2011) A geographically-restricted but prevalent *Mycobacterium tuberculosis* strain identified in the West Midlands Region of the UK between 1995 and 2008. *PLoS One* **6**: e17930.

Chapter 7

Menendez, M.C., Buxton, R.S., Evans, J.T., Gascoyne-Binzi, D., Barlow, R.E., Hinds, J., Hawkey, P.M., & Colston, M.J. (2007) Genome analysis shows a common evolutionary origin for the dominant strains of *Mycobacterium tuberculosis* in a UK South Asian community. *Tuberculosis (Edinb)* **87**: 426-436.

Other Publications

Brudey, K., Driscoll, J.R., Rigouts, L., Prodinger, W.M., Gori, A., Al-Hajoj, S.A., Allix, C., Aristimuno, L., Arora, J., Baumanis, V., Binder, L., Cafrune, P., Cataldi, A., Cheong, S., Diel, R., Ellermeier, C., Evans, J.T., Fauville-Dufaux, M., Ferdinand, S., Garcia de Viedma, D., V, Garzelli, C., Gazzola, L., Gomes, H.M., Guttierez, M.C., Hawkey, P.M., van Helden, P.D., Kadival, G.V., Kreiswirth, B.N., Kremer, K., Kubin, M., Kulkarni, S.P., Liens, B., Lillebaek, T., Ho, M.L., Martin, C., Martin, C., Mokrousov, I., Narvskaia, O., Ngeow, Y.F., Naumann, L., Niemann, S., Parwati, I., Rahim, Z., Rasolofo-Razanamparany, V., Rasolonavalona, T., Rossetti, M.L., Rusch-Gerdes, S., Sajduda, A., Samper, S., Shemyakin, I.G., Singh, U.B., Somoskovi, A., Skuce, R.A., van Soolingen, D., Streicher, E.M., Suffys, P.N., Tortoli, E., Tracevska, T., Vincent, V., Victor, T.C., Warren, R.M., Yap, S.F., Zaman, K., Portaels, F.,

Rastogi, N., & Sola, C. (2006) *Mycobacterium tuberculosis* complex genetic diversity: mining the fourth international spoligotyping database (SpolDB4) for classification, population genetics and epidemiology. *BMC Microbiol* **6**:23:

Evans,J.T., Parveen,A., Smith,G.E., Xu,L., Chan,E.W., Chan,R.C., & Hawkey,P.M. (2009) Application of denaturing HPLC to rapidly identify rifampicin-resistant *Mycobacterium tuberculosis* in low- and high-prevalence areas. *J Antimicrob Chemother* **63**: 295-301.

Evans, J.T., Smith, E.G., Banerjee, A., Smith, R.M., Dale, J., Innes, J.A., Hunt, D., Tweddell, A., Wood, A., Anderson, C., Hewinson, R.G., Smith, N.H., Hawkey, P.M., & Sonnenberg, P. (2007) Cluster of human tuberculosis caused by *Mycobacterium bovis*: evidence for person-to-person transmission in the UK. *Lancet* **369**: 1270-1276.

Mandal, S., Bradshaw, L., Anderson, L.F., Brown, T., Evans, J.T., Drobniewski, F., Smith, G., Magee, J.G., Barrett, A., Blatchford, O., Laurenson, I.F., Seagar, A.L., Ruddy, M., White, P.L., Myers, R., Hawkey, P., & Abubakar, I. (2011) Investigating Transmission of *Mycobacterium bovis* in the United Kingdom in 2005 to 2008. *J Clin Microbiol* **49**: 1943-1950.

Thorne, N., Borrell, S., Evans, J., Magee, J., Garcia de Viedma, D., V, Bishop, C., Gonzalez-Martin, J., Gharbia, S., & Arnold, C. (2011) IS*6110*-based global phylogeny of *Mycobacterium tuberculosis*. *Infect Genet Evol* **11**: 132-138.

Thorne, N., Evans,J.T., Smith,E.G., Hawkey,P.M., Gharbia,S., & Arnold,C. (2007) An IS*6110*-targeting fluorescent amplified fragment length polymorphism alternative to IS*6110* restriction fragment length polymorphism analysis for *Mycobacterium tuberculosis* DNA fingerprinting. *Clin Microbiol Infect* **13**: 964-970.

1. INTRODUCTION

1.1. GLOBAL EPIDEMIOLOGY

Mycobacterium tuberculosis is the single leading global cause of infectious disease. There were an estimated 9.4 million incident cases of TB world-wide in 2008 of which 80% were in 22 countries (Figure 1.1). South-East Asia has the highest total number of cases (35% of global total) whereas rates in Sub-Saharan Africa are nearly twice that in South-East Asia at over 350 cases per 100,000 population. Of the 15 countries with the highest estimated TB incidence rates, 13 are in Africa with half of all new cases in six Asian countries - Bangladesh, China, India, Indonesia, Pakistan, and the Philippines. There were 1.3 million deaths caused by TB in 2008 with more than two billion people infected. The overall global TB incidence rate is very slowly decreasing at a rate of less than 1% each year. However, the slow decline in incidence rates is offset by global population growth. (World Health Organisation, 2009).

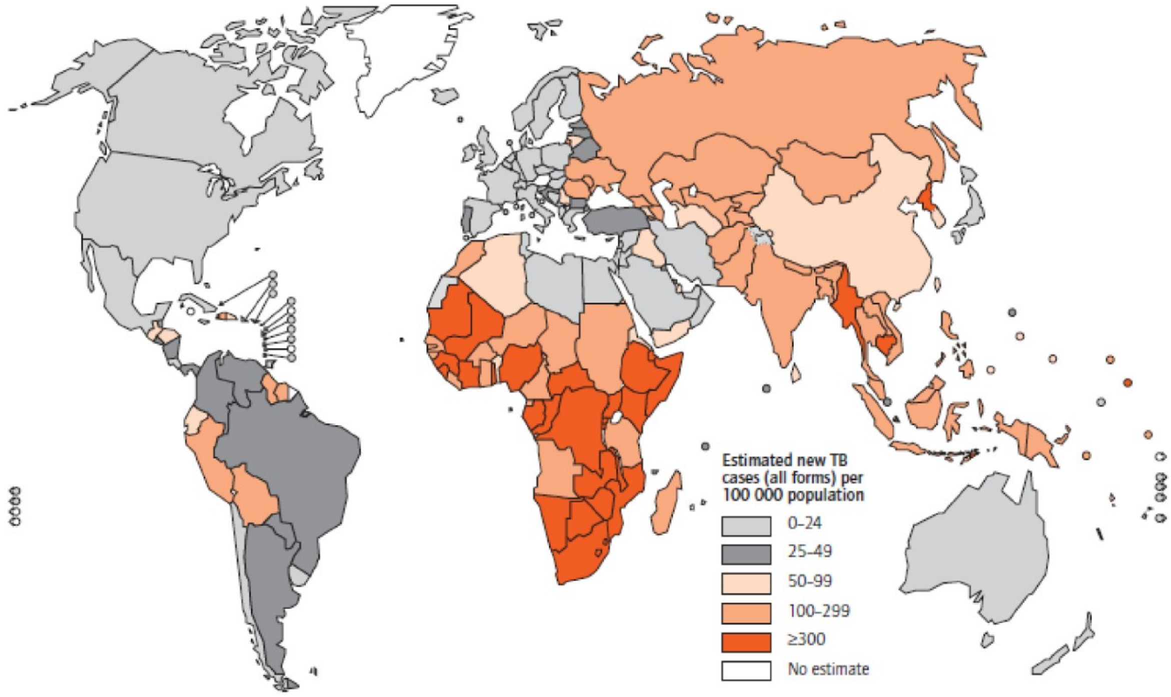


Figure 1.1. Estimated global TB incidence rates, 2008.

Figure from (World Health Organisation, 2009)

1.1.1. UK epidemiology

In the UK, the number of clinical cases reached an all-time low in 1987 with 5,085 cases. However, since then, the numbers of cases have slowly and consistently increased by approximately 2-3% each year with 9,040 cases in 2009 representing a national rate of 14.6 cases per 100,000 population (Figure 1.2). Just over half of all patients in 2009 were male (4,980/8,987, 55%) and patients aged 15-44 years old accounted for 60% of cases (5,425/9,040) (Health Protection Agency, 2010).

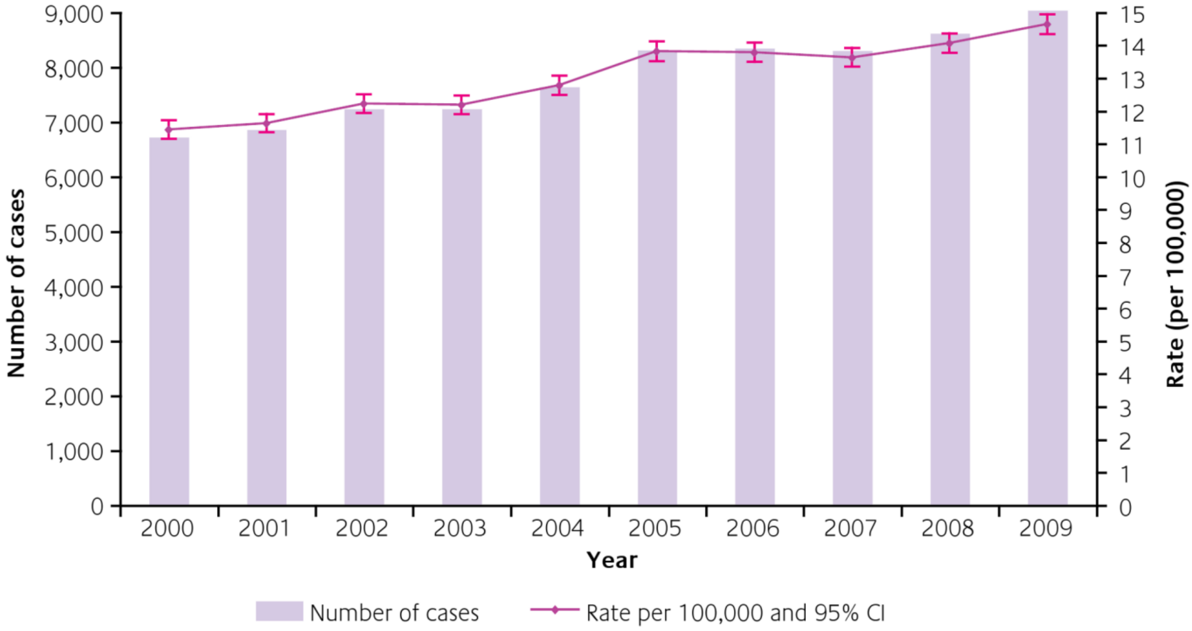


Figure 1.2 Tuberculosis case reports and rates in the UK, 2000-09.

Figure from (Health Protection Agency, 2010)

1.1.2. Midlands epidemiology

The Health Protection Agency (HPA) Midlands Regional Centre for Mycobacteriology (MRCM) laboratory receives specimens and positive cultures from >35 National Health Service (NHS) laboratories in the West and East Midlands for mycobacterial identification, speciation, drug sensitivity testing, and DNA fingerprinting. The Midlands region of the UK encompasses the West and East Midlands (Figure 1.3) which had a total population of 5.4 and 4.4 million people respectively in 2009 (Office for National Statistics, 2010).

In 2009, there were 1,564 clinical cases (16.0 per 100,000) in the Midlands region of the UK with 1,018 cases (18.7) in the West Midlands and 546 (12.3) in the East Midlands. Only London accounted for a higher proportion of all cases in the UK in 2009 with 3,440/9,040 (38%) and a rate of 44.4 per 100,000. There has been a 45% increase in case numbers in the West Midlands with a 29% increase in the East Midlands since 2000 (Figure 1.4). In London, there has been a 30% increase in case numbers since 2000. Birmingham is the largest city in the Midlands with a total population of one million and 429 cases (42.4 cases per 100,000) in 2009 (Personal Communication, Helen Bagnall, HPA West Midlands Regional Surveillance Unit) (Office for National Statistics, 2003).

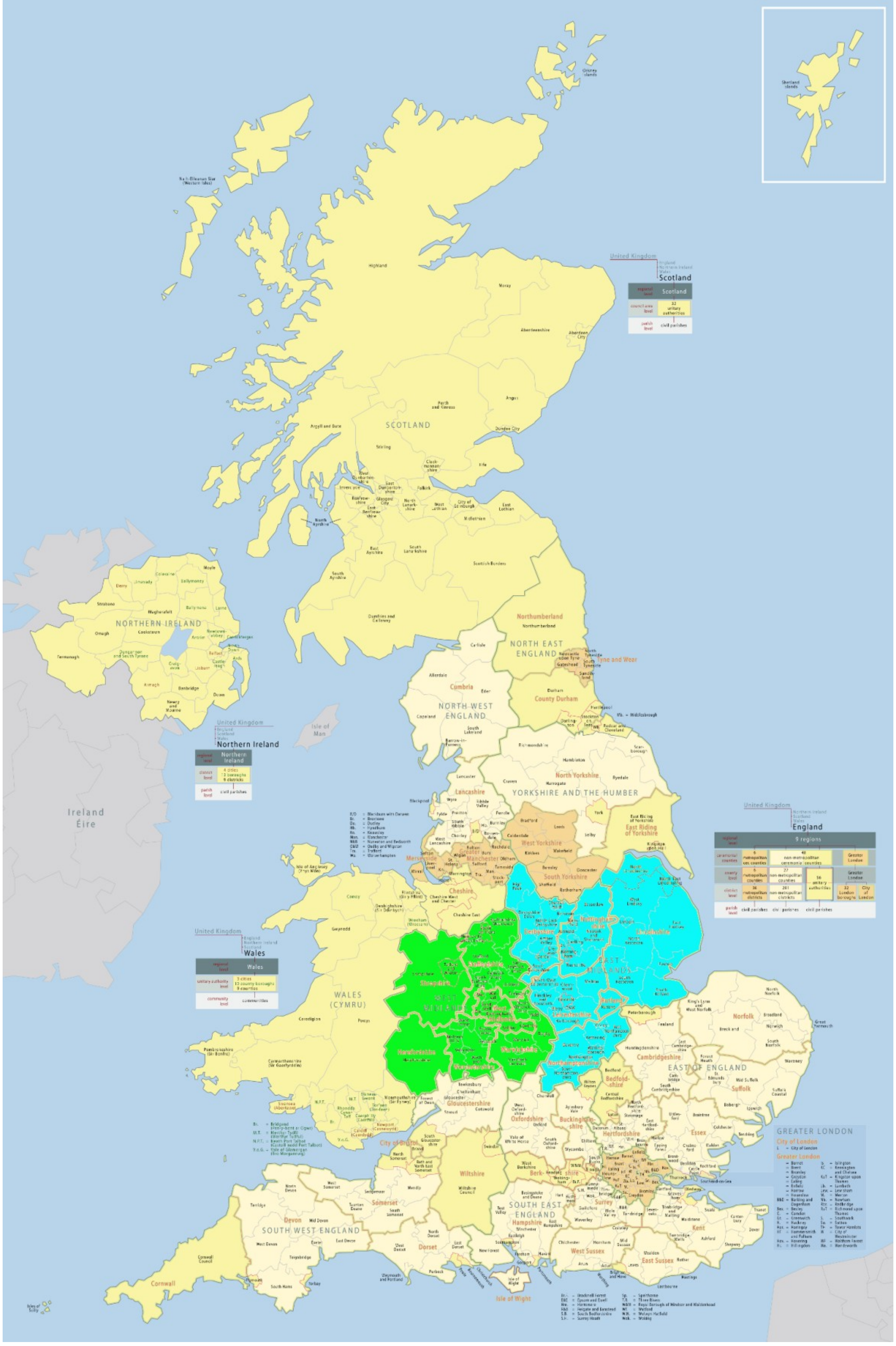


Figure 1.3. Map of the administrative geography of the United Kingdom.

The West Midlands is highlighted in green and the East Midlands in blue (adapted from http://en.wikipedia.org/wiki/File:Map_of_the_administrative_geography_of_the_United_Kingdom.png).

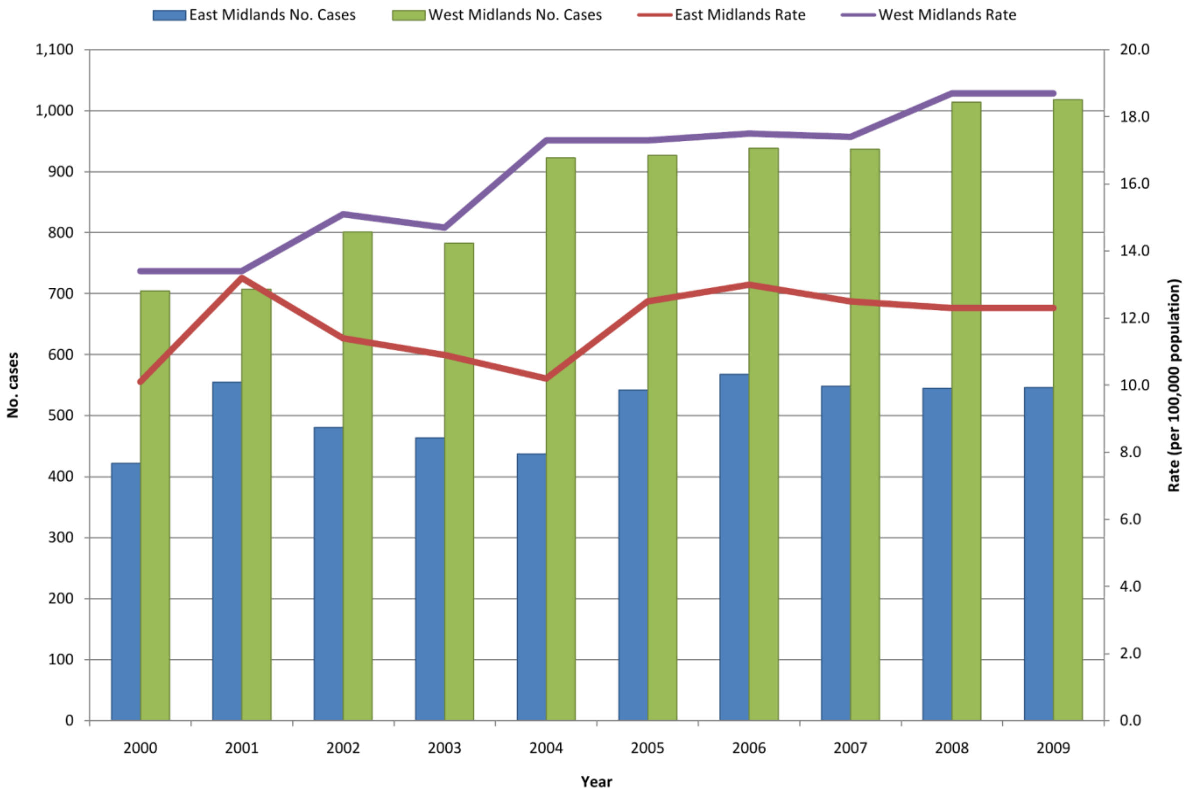


Figure 1.4. Tuberculosis case reports and rates in the East and West Midlands, England, 2000-2009.

1.2. THE AETIOLOGICAL AGENT OF TUBERCULOSIS

The clinical disease of tuberculosis is caused by various members of the *M. tuberculosis* complex (MTBC) and comprises *M. tuberculosis*, *Mycobacterium africanum*, *Mycobacterium bovis*, *Mycobacterium bovis* BCG, *Mycobacterium microti*, *Mycobacterium canettii*, *Mycobacterium pinnipedii*, *Mycobacterium caprae*, and the recently proposed *Mycobacterium mungi* (Table 1.1) (Brosch *et al.*, 2002;Alexander *et al.*, 2010). The members of the MTBC display different phenotypic traits including colony morphology, biochemical profiles, and host ranges (Levy-Frebault and Portaels, 1992) but the complex represents extreme genetic homogeneity when compared with other bacterial species with <0.03% synonymous nucleotide variation within the complex (Sreevatsan *et al.*, 1997).

M. tuberculosis is the predominant cause of tuberculosis in humans with 5,014/5,075 (99%) of culture positive human cases of tuberculosis in the UK caused by *M. tuberculosis* in 2009, 25/5,075 (0.5%) were identified with *M. bovis* and 36/5,075 (0.7%) with *M. africanum*. (Health Protection Agency, 2010).

Species	**Predominant host**
M. tuberculosis	Humans, domesticated animals
M. africanum	Humans and primates
M. bovis	Cattle and wild animals
M. bovis BCG	Vaccine strain
M. microti	Rodents/voles
M. canettii	Humans in Africa
M. pinnipedii	Seals
M. caprae	Goats
M. mungi	Banded mongoose

Table 1.1. Members of the *M. tuberculosis* complex and host specificity.

BCG: Bacillus Calmette-Guèrin. List created from (Liebana *et al.*, 1996;van Soolingen *et al.*, 1997;Aranaz *et al.*, 1999;Alexander *et al.*, 2010)

1.3. PATHOGENESIS

In all individuals infected with *M. tuberculosis*, the primary outcome of infection with *M. tuberculosis* is not disease but containment and latency as 90% will never develop any clinical, active disease (Kaufmann, 2001). Therefore, most individuals infected with *M. tuberculosis* mount an immune response that is sufficient to prevent disease. The current global transmission and epidemiology of *M. tuberculosis* in humans is a reflection of the continuous interaction between the host and strain over many bacterial and human generations but understanding the extent to which *M. tuberculosis* has adapted to the human host still remains to be fully elucidated.

Infection with *M. tuberculosis* in humans occurs by airborne transmission between close contacts. It is estimated that one untreated smear-positive patient results in about 10 secondary infections annually (Styblo, 1980). About 10 percent of immunocompetent people infected with TB develop the disease with persons that do not develop active disease within two years of acquisition having a 10% lifetime risk of developing TB (Sutherland, 1976). Disease presentation varies between patients which suggests that host factors play a significant role in an individual's susceptibility to infection and disease. Several studies have proposed a role for variation in host genetics which contribute to susceptibility to TB but verification of these proposed genetic mechanisms in different global human populations has not always been achieved (Stein, 2011).

The other 90% of persons infected with TB that do not progress to active disease will never develop active or infectious disease but will remain latently infected without complete eradication of *M. tuberculosis*. The major predisposing factor in progression to active disease is the interaction between *M. tuberculosis* and the host immune response as host co-factors that reduce immunosuppression such as co-infection with HIV increase the 10% lifetime risk to a 10% risk for each year of life (Kaufmann, 2001).

There are two forms of tuberculosis: TB affecting the lungs (respiratory disease) and TB causing infection elsewhere in the body (non-respiratory disease). Most patients with TB in the UK present with respiratory disease (4,851/8,968, 54%) (Health Protection Agency, 2010). *M. tuberculosis* is an intracellular pathogen that in respiratory disease is found primarily in the alveolar macrophage. Thus, the focus of the pathogenic properties of *M. tuberculosis* is based on the interaction with the hostile environment of the host alveolar macrophage and the exposed cell surface of the mycobacterial cell (McDonough *et al.*, 1993).

Airborne droplets containing *M. tuberculosis* are inhaled and can reach the pulmonary alveoli (Figure 1.5). *M. tuberculosis* then infects and can replicate within the alveolar macrophage. From this, a granulomatous lesion can develop which if the host is immunocompetent will contain and limit the mycobacteria. Within a granuloma, there is a highly heterogeneous immune cell population containing a variety of different T cell populations, B cells, macrophages at different stages of maturation, fibroblasts, and dendritic cells. If the cellular immune response is altered, then the granuloma may no longer be able to contain and limit spread of the mycobacteria. The granuloma liquefies and mycobacteria are released not only within the lung but can disseminate to other extra-pulmonary sites via the lymphatic system and blood, resulting in miliary (disseminated) disease, extra-pulmonary tuberculosis, or meningitis (Kaufmann, 2002;Donoghue, 2009).

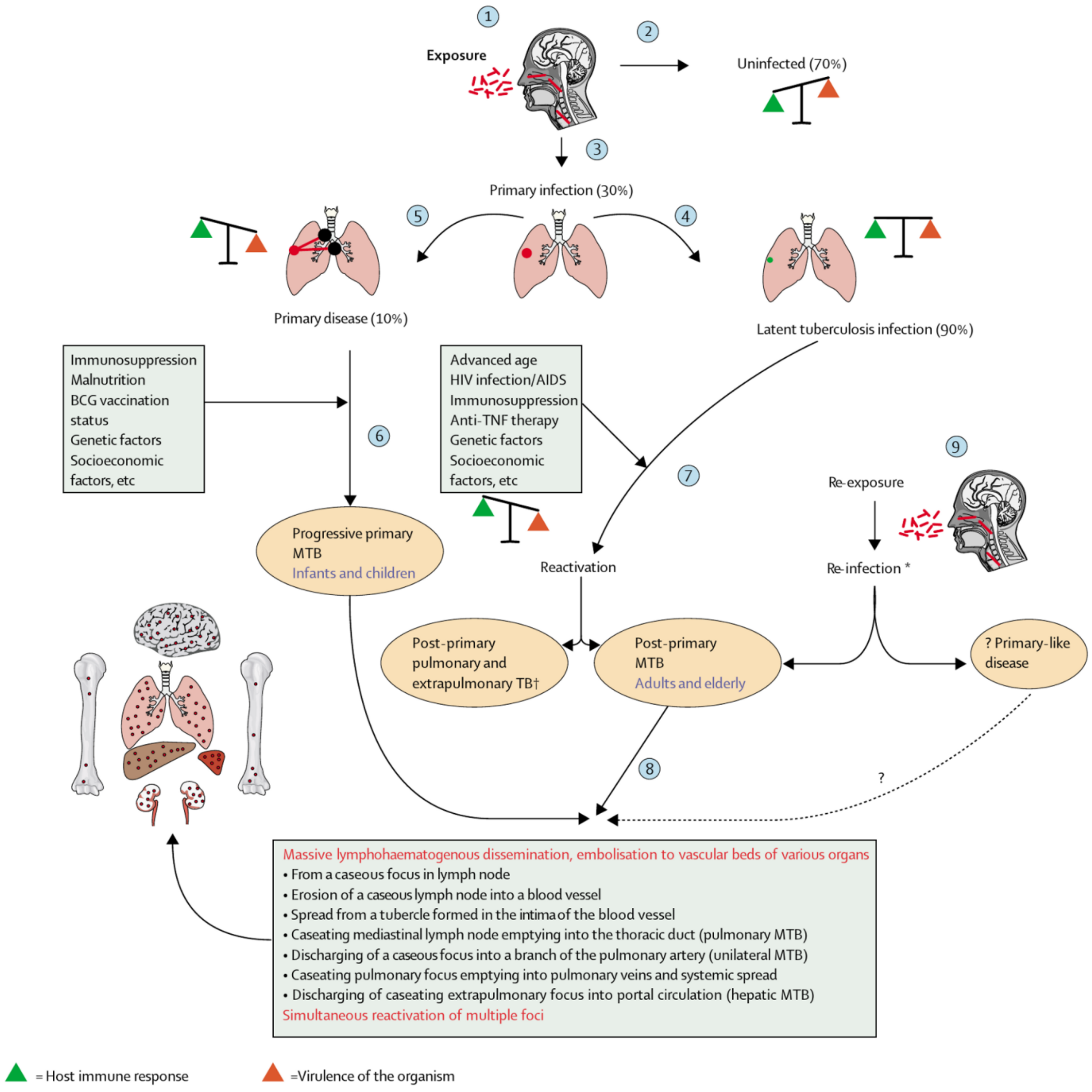

1
Exposure
2
Uninfected (70%)
3
Primary infection (30%)
5
4
Primary disease (10%)
Latent tuberculosis infection (90%)
Immunosuppression
Malnutrition
BCG vaccination status
Genetic factors
Socioeconomic factors, etc
6
Advanced age
HIV infection/AIDS
Immunosuppression
Anti-TNF therapy
Genetic factors
Socioeconomic factors, etc
7
9
Re-exposure
Progressive primary MTB
Infants and children
Reactivation
Re-infection *
Post-primary pulmonary and extrapulmonary TB†
Post-primary MTB
Adults and elderly
? Primary-like disease
8
?
Massive lymphohaematogenous dissemination, embolisation to vascular beds of various organs
• From a caseous focus in lymph node
• Erosion of a caseous lymph node into a blood vessel
• Spread from a tubercle formed in the intima of the blood vessel
• Caseating mediastinal lymph node emptying into the thoracic duct (pulmonary MTB)
• Discharging of a caseous focus into a branch of the pulmonary artery (unilateral MTB)
• Caseating pulmonary focus emptying into pulmonary veins and systemic spread
• Discharging of caseating extrapulmonary focus into portal circulation (hepatic MTB)
Simultaneous reactivation of multiple foci
▲ = Host immune response
▲ =Virulence of the organism

Figure 1.5. The development of tuberculosis.

Small droplet nuclei (1-5 µm) containing *M. tuberculosis* are deposited in the alveoli (1). 70% of individuals exposed are not infected (2) but 30% are infected (3). Infection is contained in 90% of those infected (latent infection) (4). The remaining 10% will develop progressive primary tuberculosis (5). Extensive dissemination (6) to various organs can result in miliary TB. People with latent TB infection have a 10% lifetime risk of re-activation and post-primary tuberculosis (7). 50% of re-activations occur during the first two years of primary infection. The risk of re-activation in HIV-infected individuals is approximately 10%/year. Massive lymphohaematogenous dissemination during re-activation (8) can also result in miliary tuberculosis (progressive post-primary miliary tuberculosis). Re-infection with a new strain of *M. tuberculosis* (9) can occur and the cycle is repeated. *Important in endemic areas. †Organ-restricted tuberculosis with adequate host immunity. MTB=miliary tuberculosis, TB=tuberculosis, TNF=tumour necrosis factor (Sharma et al., 2005).

1.4. TRANSMISSION

The airborne transmission of tuberculosis by droplet nuclei was confirmed by the infection of 71/77 guinea pigs that were housed above a ward for TB patients and exposed to airborne bacilli generated by these patients over a 2-year period. Infection was detected by tuberculin skin testing and confirmed by culture and histological examination of lung, spleen, and lymph node tissue (Riley, 1957).

The infectiousness of a patient with respiratory disease can be determined by the smear microscopy status of the patient. Non-respiratory disease is generally accepted as a form of TB which represents acquisition of *M. tuberculosis* with a low risk of onward transmission to another patient.

The oldest and most widely used technique for primary identification of the presence of acid-fast bacilli is the microscopic examination of a sputum specimen by specific stains, either auramine-phenol for examination of specimens or Ziehl-Neelsen for positive cultures (Ulukanligil *et al.*, 2000;Murray *et al.*, 2003). The presence or absence of acid-fast bacilli in a sputum specimen indicates the degree of infectiousness of the patient. A survey of 1,532 children in Bedfordshire in the UK between 1945 and 1952 showed that 244/374 (65.2%) children in contact with sputum smear positive culture positive patients showed evidence of recent infection using the tuberculin skin test whereas only 61/228 (27.8 %) of children in contact with smear negative culture positive showed evidence of recent infection (Shaw and Wynn-Williams, 1954). Therefore, contacts of patients with smear positive culture positive tuberculosis were at least twice as likely to be infected with tuberculosis when compared to contacts of patients with smear negative culture positive tuberculosis. It was estimated that patients with smear-positive disease expectorate 10^8-10^{10} bacilli daily (Pottenger, 1948), or 10^6-10^7 AFB per ml of sputum, while smear-negative sputum contains $<10^3$ bacilli per ml of sputum (Yeager, Jr. *et al.*, 1967). However, it is not known how many bacilli must be inhaled to contract *M. tuberculosis* as this is very likely to be highly variable between individuals. It is known that the likelihood of infection depends on the intensity, frequency, and duration of exposure (Ewer *et al.*, 2003).

Current guidelines in the UK assess the smear status of an individual patient as part of infection control assessments to determine whether a patient requires hospital isolation in a single bed negative pressure room. Sputum smear status also determines whether enhanced contact screening using Interferon Gamma Release Assays (IGRA) and chest X-rays are required in contact tracing investigations (National Institute for Clinical Excellence, 2011). Within the UK in 2009, a sputum smear microscopy result was known for 2,790/4,851 (58%) pulmonary cases with 1,579/2,790 (57%) smears positive for acid-fast bacilli. Of these 1,579 smear positive cases, 1,389/1,579 (89%) were culture positive (Health Protection Agency, 2010).

1.5. CLINICAL DIAGNOSIS OF ACTIVE DISEASE

The classic signs and symptoms of active TB are malaise, weight loss, fever and night sweats with a cough that may induce haemoptysis in a small proportion of cases (Brandli, 1998). Upon presentation, an initial chest X-ray is taken and if this suggests TB then at least three sputum samples including one early morning sample are obtained and referred for microbiological culture and microscopy. If sputum samples are not spontaneously produced then sputum induction or bronchoscopy and lavage in adults is carried out with safe induction of sputum in children or a gastric lavage. Samples should be obtained before the initiation of therapy if the patient has clinical signs and symptoms conducive with a diagnosis of TB and therapy should be started before microbiological results are known (National Institute for Clinical Excellence, 2011). The confirmation of a clinical case is achieved by identifying *M. tuberculosis* by microbiological culture.

Microbiological diagnosis is based on initial smear microscopy to detect acid fast bacilli (AFB), culture using liquid and solid media, isolate identification, and drug sensitivity testing. There have been various recent developments to improve the microscopic identification of *M. tuberculosis* in the 21st century. Light emitting diodes are now used as alternative light sources to replace expensive mercury lamps which can enable microscopy to be readily applied in high burden areas (Hung *et al.*, 2007). Mobile phone digital cameras have been utilised to enable the transmission of slide images to trained personnel at a central location (Breslauer *et al.*, 2009), and automated slide scanners are being developed by the WHO Foundation for Innovative New Diagnostics.

Culture has traditionally used a solid egg-based medium such as Lowenstein-Jensen (LJ) media which contains a variety of supplements and selective agents to inhibit growth of rapid-growing non-mycobacteria and enhance mycobacterial growth. Culture on LJ slopes can be slow with a recommended incubation time of up to three months for all specimens. To reduce the time taken to obtain a positive culture, more sensitive and rapid radiometric culture systems were introduced. To reduce exposure to hazardous reagents used by the radiometric culture systems, fluorescence-based liquid media culture systems were developed.

A multi-centre study compared radiometric culture (BACTEC 460 TB System), fluorescent detection using Mycobacteria Growth Indicator Tubes (MGIT), and solid media in the isolation of 180 mycobacterial isolates from 1,500 specimens which included 113 *M. tuberculosis* complex and 67 nontuberculous mycobacteria (NTMs) When a combination of liquid and solid media was used, it was shown that fluorescent MGIT plus solid media detected 156/180 (86.7%) of the isolates which was statistically non-inferior to radiometric BACTEC plus solid media which recovered 168/180 (93.3%) of all AFB. A combination of the two liquid media together detected 171/180 (95%) of all isolates which was significantly different when compared to MGIT plus solid media but not when compared to BACTEC plus solid media. The two liquid media systems greatly reduced the time to detection of *M. tuberculosis* complex which was 9.9 days with MGIT, 9.7 days with BACTEC, and 20.2 days with solid media. From this, the MGIT system was shown to be a beneficial alternative to the radiometric cultivation system since it removed the need for the use of hazardous reagents (Pfyffer *et al.*, 1997). However, there are instances where isolates will only be identified on solid LJ media so for optimal rates of isolation both types of culture media still need to be used.

1.5.1. Identification

Mycobacterial isolates were traditionally identified via a series of biochemical tests which examined growth on different substrates and also growth at different temperatures. The *M. tuberculosis* complex can be phenotypically distinguished from NTMs by growth only at 37°C, no growth in the presence of ρ-nitro benzoic acid or thiacetazone, preference for an aerobic atmosphere, lack of pigmentation in light or dark and the ability to hydrolyse Tween. *M. bovis* has similar characteristics except that it cannot hydrolyse Tween (Cowan and Steele, 1993).

Genotypic tests for identification of positive cultures using the Polymerase Chain Reaction (PCR) have been developed. Two commercially available assays are the Innogenetics InnoLipa and Hain LifeScience Genotype assay (Richter *et al.*, 2003;Tortoli *et al.*, 2003). These two assays both employ PCR-based amplification and reverse line hybridisation to detect single nucleotide polymorphisms (SNPs) in the 16S-23S rRNA spacer region or 23S rRNA, *gyrB*, and *M. bovis* BCG RD1 (region of difference 1) respectively. These two tests can identify individual members of the *M. tuberculosis* complex and most of the common NTMs.

1.5.2. Drug Sensitivity Testing

There are three phenotypic methods that are used in Europe which follow guidelines and definitions set out by the WHO: the absolute concentration method; the resistance ratio method; and the proportion method (Canetti *et al.*, 1969). In the absolute concentration method, two dilutions of each drug are used in two separate solid or broth cultures. Susceptibility is defined as the lowest concentration of the drug that inhibits growth and results in less than 20 colonies. This essentially calculates the minimal inhibitory concentration (MIC) for an isolate. Drug concentrations and inoculum sizes must be carefully standardised. The resistance ratio method is a further development of the absolute concentration method where the MIC for a susceptible reference strain is compared to the MIC of the specimen isolate (MIC for test isolate is divided by MIC for reference strain). If the ratio is two or less, the isolate is fully susceptible. If the ratio is eight or more, the isolate is highly resistant. Within the resistance ratio method, low-level or intermediate resistance is difficult to quantify accurately and the inoculum size again needs to be standardised (Drobniewski *et al.*, 2007).

In the proportion method, each isolate is grown on drug-containing and drug-free media and the proportion of drug-resistant mutants is calculated from the total colony count on the drug-free media. This proportion varies with different antibiotics. For both isoniazid and rifampicin, the proportion is 1%. If more than 1% of the total colony count on drug-free media grows on drug-containing media, then the isolate is resistant. The proportions set correlate with an effective clinical outcome. The widely used broth-based BD MGIT system uses a modified proportion method (Drobniewski *et al.*, 2007).

As for culture and isolation, solid media has been traditionally used with liquid media systems recently introduced which have been shown to reduce the turn-around times to obtain antibiotic sensitivity data (Tortoli *et al.*, 2002). The WHO has a global network of 20 supra-national laboratories that provide drug sensitivity testing for all first line agents and second and third line agents as well and these laboratories participate in global surveillance of drug resistance (Wright *et al.*, 2009).

There have been several novel assays recently developed as part of global efforts to produce drug sensitivity testing that can be implemented in high-burden areas. The microscopic-observation drug-susceptibility assay uses a micro-titre plate format to identify rifampicin resistance within seven days (Moore *et al.*, 2006). Phage-based assays have also been developed for the identification of rifampicin resistance as well (Wilson *et al.*, 1997).

1.5.3. Direct identification and Drug Sensitivity Testing from specimens by Nucleic Acid Amplification

It is possible to directly identify *M. tuberculosis* complex DNA present in specimens using PCR-based amplification of species specific DNA sequences. However, a systematic review of rapid diagnostic tests for the detection of tuberculosis infection from 212 studies published between 1975 and 2003 found that TB-PCR is a reliable method for positively identifying tuberculous disease but that sensitivity is too poor to be able to accurately exclude disease, especially in smear-negative disease where clinical diagnosis may be absolutely dependent on microbiological examination (Dinnes *et al.*, 2007).

A recently developed nucleic acid amplification test is the Cepheid Gene Xpert MTB/RIF assay which aims to overcome these technical challenges and reduce the requirement for highly skilled personnel such that this assay can be a rapid on-demand, “near-patient technology”. This assay can simultaneously detect *M. tuberculosis* DNA and mutations which confer rifampicin resistance in a sputum specimen within two hours. It has been shown to exhibit at least 98.2% sensitivity in smear-positive culture-positive cases and at least 72% sensitivity in smear-negative culture-positive cases (Helb *et al.*, 2010;Boehme *et al.*, 2010).

Rapid identification of resistance to rifampicin and isoniazid can also be detected genotypically using nucleic acid amplification tests such as those commercially supplied by InnoLipa and Hain LifeScience (Ling *et al.*, 2008;Hillemann *et al.*, 2005;Morgan *et al.*, 2005;Telenti *et al.*, 1997). Sensitivity and specificity is greatest for detection of mutations conferring rifampicin resistance with reduced values for isoniazid and the two other first-line agents. Mutations conferring resistance to 2nd-line therapeutic agents can also now be detected as well (Hillemann *et al.*, 2009).

1.5.4. Treatment

There are four first-line antibiotics used for the treatment of uncomplicated tuberculosis in the UK: isoniazid, rifampicin, ethambutol, and pyrazinamide. Within the UK, patients with active disease are prescribed six months of isoniazid and rifampicin supplemented in the first two months with pyrazinamide and ethambutol whilst latently infected patients can be treated with three months isoniazid and rifampicin or six months of isoniazid. Treatment of latent infection with six months of isoniazid or rifampicin and isoniazid for three months is considered in patients <35 years old, HIV positive, or Healthcare Workers (National Institute for Clinical Excellence, 2011).

The development of the current first line therapy of four drugs was achieved by studies carried out by the TB Units of the British Medical Research Council (MRC) around the world between 1946 and 1986. There were four milestones between 1970 and 1976 that generated the current standardised therapy of four first line antibiotics: addition of rifampicin or pyrazinamide to the current regimen in 1970 (streptomycin and isoniazid) substantially reduced the subsequent relapse rate; inclusion of rifampicin and pyrazinamide shortened treatment length to six months; and it was shown that the sterilising activity of pyrazinamide was confined to the first two months of treatment whereas rifampicin exhibited sterilising activity throughout the intensive and continuation phase which in turn created the current modern short-course therapy regimen (Fox *et al.*, 1999).

The most important determinant for success of therapy is the development of drug resistance. The known targets for each for each of the four first line drugs and the genetic location of mutations that confer drug resistance are listed in Table 1.2. However, not all mutations that confer resistance have been fully elucidated.

Drug	**Known or probable targets**	**Mutations in genes conferring resistance**	**Function of gene**
Rifampicin	RNA synthesis	*rpoB*	DNA-dependent RNA polymerase (β subunit)
Isoniazid	Mycolic acid biosynthesis	*katG* *inhA*/*mabA* *ahpC*	Catalase/peroxidase Fatty-acid biosynthesis Alkylhydroperoxide C reductase
Ethambutol	Cell wall synthesis	*embA,B,C*	Lipoarabinomannan and arabinogalactan synthesis
Pyrazinamide	Pyrazinamidase	*pncA* + others	Pyrazinamidase

Table 1.2 Molecular detection of drug resistance in *M. tuberculosis*

Table from Drobniewski *et al.*, 2007.

M. tuberculosis isolates that are resistant to both isoniazid and rifampicin are defined as multidrug-resistant TB (MDR-TB). Resistance to isoniazid and rifampicin, plus any fluoroquinolone and at least one of three injectable second-line drugs (amikacin, kanamycin, or capreomycin) is defined as extensively drug resistant TB (XDR-TB) (Centers for Disease Control and Prevention (CDC)., 2006). There were an estimated 440,000 global cases of MDR-TB in 2008 resulting in 150,000 deaths. Of all global incident cases of TB it is estimated that 3.6% are MDR-TB. Almost half of all MDR-TB cases are estimated to be present in China and India (World Health Organisation, 2010a). In the UK in 2008, the level of isonizaid resistance was 6.9% of all new cases while rifampicin mono-resistance and MDR-TB rates were 1.4% and 1.2% respectively (Health Protection Agency, 2010). Eight XDR-TB cases have been reported in the UK between 1995 and 2008 which accounted for 8/678 (0.9%) of all MDR-TB cases reported in this time period (Abubakar *et al.*, 2009).

1.5.5. Immunological diagnosis of latent disease

Identification and prophylactic treatment of latently infected patients before active disease develops is important in "ring-fencing" an outbreak and preventing further transmission. The traditional method of identifying infection with *M. tuberculosis* was via sub-cutaneous injection of purified protein derivative (PPD) as the tuberculin skin test (TST) and analysis of the host response. However, TST has limited specificity as it cannot identify *M. tuberculosis* infection in individuals who have been previously vaccinated with the BCG vaccine or infected with environmental mycobacteria.

Two sensitive enzyme-linked immunosorbent assays (ELISAs) have been developed that can detect T cells present in whole blood specimens with specific antigens for *M. tuberculosis* that are absent from *Mycobacterium bovis* BCG and most environmental mycobacteria. These two tests are the Oxford ImmunoTec T-SPOT®.TB (Ewer *et al.*, 2003;Lalvani *et al.*, 2001) and Cellestis QuantiFERON®-TB Gold (Mazurek *et al.*, 2001).

These two ELISAs detect the production of interferon-gamma in whole blood from patients that is incubated with two antigens expressed by *M. tuberculosis*: early secretory antigenic target-6 (ESAT-6) and culture filtrate protein-10. These two antigens are secreted by all *M. tuberculosis* and *M. bovis* strains but are absent in all *M. bovis* BCG vaccine strains and from most NTMs (Mazurek *et al.*, 2005;Andersen *et al.*, 2000). T-SPOT®.TB enumerates the actual number of IFN-y releasing T-cells after antigen incubation whereas QuantiFERON® Gold measures the amount of released IFN-y. From this, these two ELISAs are termed interferon gamma release assays (IGRA).

1.5.6. Vaccination and screening

The only currently available vaccine against *M. tuberculosis* is the *M. bovis* BCG strain. Albert Calmette and Camille Guérin produced this vaccine by culturing *M. bovis* in bile for 13 years with passaging every two weeks for a total of 230 passages to produce the *M. bovis* BCG vaccine strain in 1921. During this time, the virulence of the passaged strains was measured in several animal species and was found to progressively decrease with continued passaging (Behr, 2002). The BCG vaccine has been administered to more than three billion individuals worldwide and primarily protects children from meningeal and miliary tuberculosis. However, protection of adults from pulmonary tuberculosis varies considerably from 80% in the UK to no protection in South India (Colditz *et al.*, 1994). Notably, the countries with the highest disease burden generally exhibit a low protective efficacy of BCG.

Exposure to environmental mycobacteria may provide natural protection or may impair the protection afforded by BCG (Collins and Kaufmann, 2001). To develop an improved vaccine, there are currently 11 new candidate vaccines in clinical trials (Kaufmann *et al.*, 2010).

All school pupils in the UK were routinely vaccinated with the BCG vaccine until 2005 when the national policy was changed to neonatal BCG vaccination for high risk groups and at pre-employment screening for individuals entering "at risk" occupations. The UK policy was changed as BCG is not very effective in older children and adults with at best 75% protection for 15 years. Repeat vaccinations do not provide any extra benefit. Since 1987 in the UK cases rates in UK-born people aged 15-30 declined to reach less than two cases per 100,000 population per year. This is the sole population group who were protected by routine BCG vaccination in early teenage life. From this, it has been calculated that approximately 10,000 vaccinations were required to prevent a single case of TB with adverse events outweighing the benefits received from vaccination (Davies, 2009).

New immigrants identified with active TB who have recently arrived in the UK do not present immediately after arrival as only 1,039/4,929 (21%) patients were diagnosed within two years of arrival in the UK with 2,209/4,929 (45%) patients presenting within five years; 1,198/4,929 (24%) presented with tuberculosis within five and nine years after entry and 1,522/4,929 (31%) had been in the UK for ten or more years before diagnosis (Figure 1.6). The median duration of stay in the UK before diagnosis of tuberculosis was four years with an interquartile range of 1-13 years (Health Protection Agency, 2010).

The current national UK policy for screening of recently arrived immigrants involves port-of-entry identification and screening with chest X-rays for immigrants who have originated from countries with a TB incidence rate of >40 cases per 100,000 population and intend to stay in the UK for >6 months. The aim of this policy is to detect active pulmonary tuberculosis and not individuals who are latently infected. This policy would need to be altered if identification of latent infection was an aim as well. A recent UK multicentre cohort study of 1,229 immigrants aged 35 years or younger screened for latent tuberculosis infection by the IGRA test from three centres in the UK, showed that the current UK policy would fail to detect 71% of individuals latently infected (Pareek *et al.*, 2011). Two more sensitive and cost effective strategies were proposed which increased the threshold incidence rate for screening to either 150 or 250 cases per 100,000 population. These alternative strategies would have identified 92% of individuals with latent infection.

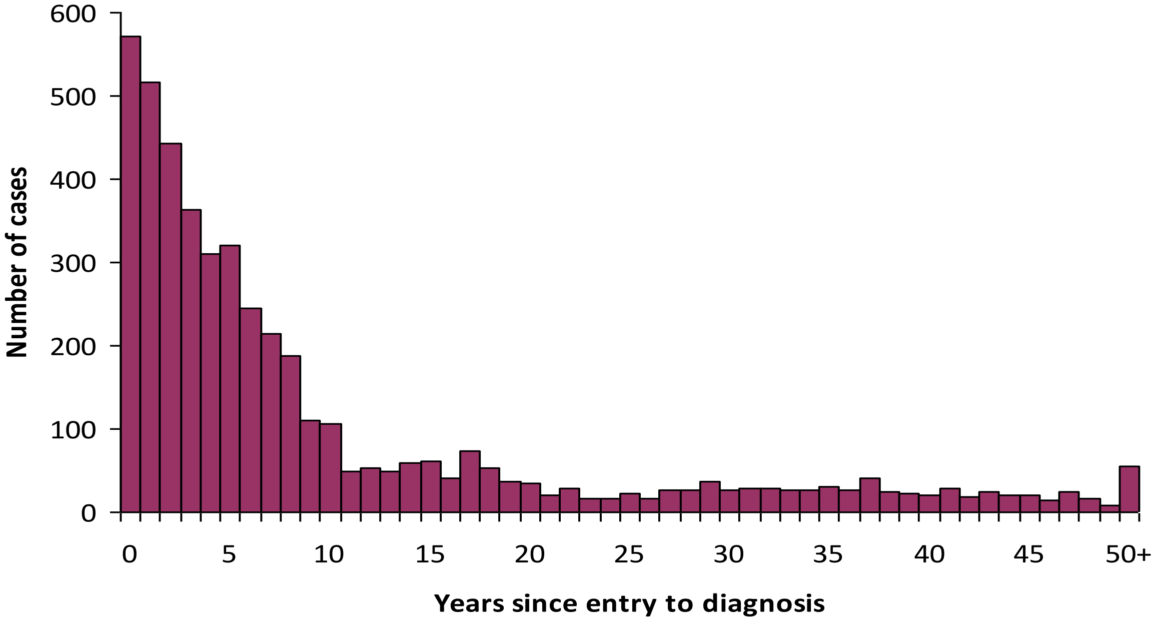


Figure 1.6. Non-UK-born tuberculosis case reports by time since entry to the UK to tuberculosis diagnosis, UK, 2008.

1.6. DNA FINGERPRINTING AND MOLECULAR EPIDEMIOLOGY

DNA fingerprinting is an essential tool in public health control efforts for tuberculosis as identification of indistinguishable strains in ≥2 patients with supporting conventional epidemiological data confirms that transmission has occurred and that a cluster of cases has been detected. Matching strains identified in epidemiologically linked patients reflect recent transmission and rapid progression from infection to active disease (Alland *et al.*, 1994;Small *et al.*, 1994). DNA fingerprinting can also be used to identify laboratory cross-contamination events where bacilli are unintentionally transferred from positive specimens into negative specimens during laboratory investigations and also determine whether two or more episodes of active TB disease in an individual represents endogenous re-activation with the previous strain or re-infection with a new exogenous strain. DNA fingerprinting methods can also provide new insights in the global epidemiology and phylogeny of TB.

1.6.1. Phage typing

The first genotypic typing method for *M. tuberculosis* utilised mycobacteriophages specific for the *M. tuberculosis* complex. There have been over 250 mycobacteriophages identified from many reservoirs including environmental, animal, and human specimens. A standardised methodology for the typing of *M. tuberculosis* using a panel of 12 mycobacteriophages was published by the WHO in 1975 (Rado *et al.*, 1975). Phage typing was a valuable epidemiological tool for cluster investigation with large-scale international surveys identifying that there were striking geographic variations in mycobacteriophage susceptibility between the strain classifications of type A, type B, type C or intermediate. Type A strains predominated in Hong Kong, type B strains were more common in northern Europe, and of the strains from Madras in the south of India, over 40% were of the intermediate type.

Although phage typing provided the first insights into strain transmission and global epidemiology, no further increase in differentiation was achieved and it was eventually superseded by the development of the analysis of the genomic position and copy number of insertion sequences in 1993 (Grange *et al.*, 1976;McNerney, 1999).

1.6.2. IS*6110* Restriction Fragment Length Polymorphism typing

Insertion sequences (IS) are mobile genetic elements (0.8-2.5 kb in length) that contain inverted repeated DNA sequences at each end and a translated transposase which makes the sequence mobile (McAdam *et al.*, 1990;Thierry *et al.*, 1990a). ISs differ from transposons as they only transfer genes necessary for their own replication. More than 16 different ISs in five IS families have been reported in mycobacteria (McAdam *et al.*, 1990). The IS*6110* element is a member of the IS3 family which are the most widely spread group of ISs in bacteria. The IS3 family have been found in more than 24 different genera. IS*6110* is apparently specific to *M. tuberculosis* when analysed by Southern blotting but cross-reactivity between mycobacterial species has been shown when fragments amplified by PCR were compared between mycobacterial species (Kremer *et al.*, 1999;McHugh *et al.*, 1997).

Epidemiologic studies of the *M. tuberculosis* complex were greatly improved by the analysis of IS*6110* by Restriction Fragment Length Polymorphism (RFLP). This is a nucleic acid hybridisation technique that analyses the copy number and location of IS*6110* genomic DNA fragments digested by a restriction enzyme (Figure 1.7). Most strains contain 6-15 copies of IS*6110* with the highest number approaching 30 copies (Cowan and Crawford, 2002). Various groups independently identified IS*6110* with an internationally standardised protocol for RFLP published in 1993 that described the use of a common 245 bp probe that hybridises to

IS*6110* fragments digested with *Pvu*II (Thierry *et al.*, 1990b;Zainuddin and Dale, 1989;Eisenach *et al.*, 1988;van Embden *et al.*, 1993).

The application of IS*6110* RFLP to initial investigations of outbreaks showed that this technique was useful in cluster investigations as it exhibited epidemiologically correct clustering of related isolates and differentiation of non-related isolates (Edlin *et al.*, 1992;van Soolingen *et al.*, 1991a;Hermans *et al.*, 1990). *M. tuberculosis* clinical isolates were highly polymorphic when analysed by IS*6110* RFLP whilst *in vitro* studies demonstrated that transposition of IS*6110* is an extremely rare event (van Soolingen *et al.*, 1991b). In a subsequent study that typed strains from 544 patients with persistent culture positive TB, it was estimated that the half-life of changes in the RFLP banding pattern was at least 3.2 years (95% CI, 2.1-5.0) (de Boer *et al.*, 1999). The half-life was defined as the average time required for 50% of IS*6110* patterns to gain or lose one band. Since the rate of change was not rapid it was accepted that the “molecular clock” of IS*6110* RFLP typing was appropriate as an accurate marker of transmission.

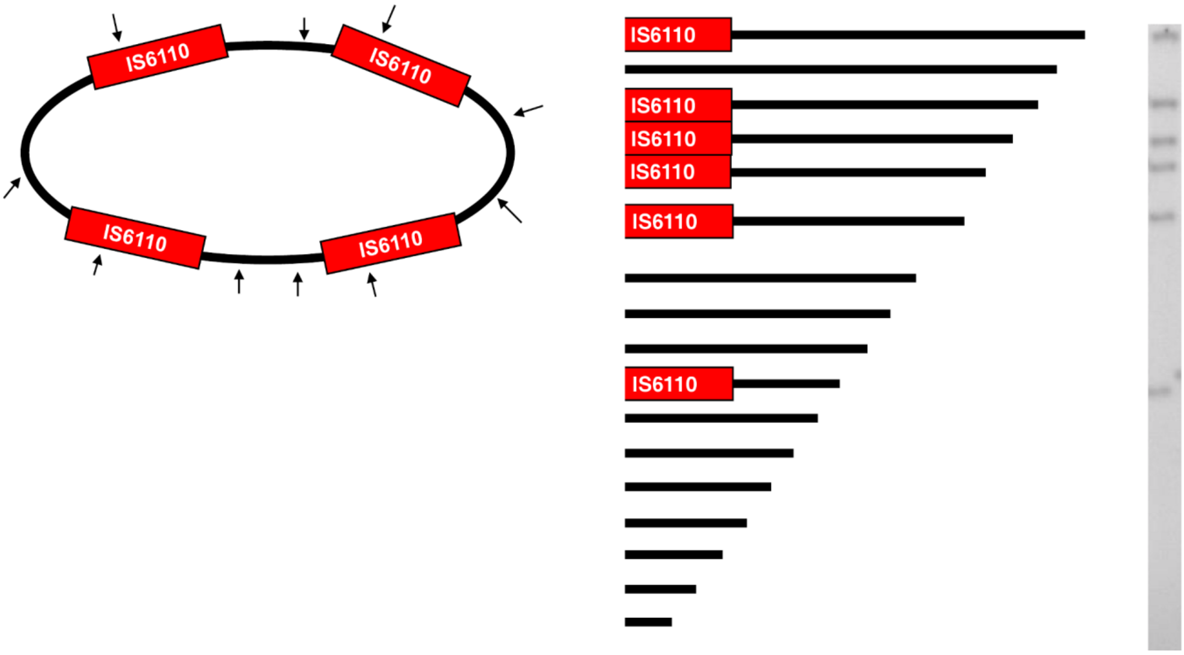


Figure 1.7 IS*6110* RFLP genotyping.

Mycobacterial DNA is digested with the restriction enzyme *Pvu*II. The IS*6110* probe hybridizes to IS*6110* DNA to the right of the *Pvu*II site in IS*6110*. The size of each hybridising fragment depends on the distance from this site to the next *Pvu*II site in adjacent DNA, as reflected by gel electrophoresis of the DNA fragments. The horizontal lines in the middle indicate the extent of the distribution of fragments in the gel, including *Pvu*II fragments that do not contain IS*6110*. A probe specific to IS*6110* is used to detect IS*6110* containing fragments. Figure adapted from Barnes and Cave, 2003.

The mechanism of IS*6110* transposition around the *M. tuberculosis* complex genome is not fully understood. Transposition does involve excision from the current site of integration and one of three possible mechanisms. Transposition can either be conservative where IS*6110* is excised and inserted into a new site or replicative where IS*6110* is duplicated. Homologous recombination between two IS*6110* elements can delete intervening genomic DNA. IS*6110* transposition usually has a negative effect on the destination gene due to the disruption of gene coding or promoter DNA (McEvoy *et al.*, 2007).

IS*6110* RFLP has been utilised in large-scale population based studies on the transmission dynamics of *M. tuberculosis* with the first studies undertaken in New York, Switzerland, San Francisco, and Denmark/Greenland (Yang *et al.*, 1994;Alland *et al.*, 1994;Small *et al.*, 1994;Genewein *et al.*, 1993). The two US studies found that even with apparently efficient tuberculosis-control programmes, recent transmission was unexpectedly high and accounted for 30-40% of all new cases with TB transmission evidently not under control. The European study found that there was significant transmission between human populations resident in Denmark and Greenland.

The country with the greatest experience of IS*6110* RFLP typing is the Netherlands with more than 24,000 strains typed since 1993 (Schurch *et al.*, 2011). The first nationwide study of 4,266 cases from 1993 through 1997 found that 1,493/4,266 (35%) of all cases could be attributed to recent transmission. Demographic risk factors for active transmission of tuberculosis in the Netherlands included male sex, urban residence, Dutch or Surinamese nationality, and long-term residence in the Netherlands (van Soolingen *et al.*, 1999).

Within the UK, there have been two large-scale population-based studies that have utilised IS*6110* RFLP typing. The first study examined 2,042 isolates from Greater London between 1995 and 1997 and estimated the rate of recent transmission within this time period as 14.4%. A later study looked at *M. tuberculosis* isolates from across England in 1998 and estimated an unadjusted rate of recent transmission as 12.2% with an adjusted estimate of 27.6% taking into account missed cases and study duration (Love *et al.*, 2009;Maguire *et al.*, 2002).

1.6.3. Spacer Oligotyping (Spoligotyping)

Spoligotyping as a typing method for *M. tuberculosis* strains was first published in 1997 (Kamerbeek *et al.*, 1997). Spoligotyping was the first genotypic method that enabled the simultaneous detection and strain differentiation of *M. tuberculosis* directly in clinical specimens without the need for culture. This method is based on a DNA polymorphism identified within the Direct Repeat (DR) locus of *M. tuberculosis*. This region contains multiple, conserved 36-bp DRs that have unique, individual spacer sequences ranging from 34-41 bp in length interspersed between each DR. *M. tuberculosis* strains vary in the number and presence or absence of DRs (Figure 1.8) (Groenen *et al.*, 1993;Hermans *et al.*, 1991). This variation is probably caused by homologous recombination between DRs with reassortments caused by transposition of IS*6110* into the DR locus (Groenen *et al.*, 1993). PCR analysis of the DR locus was originally by "Direct Variable Repeat PCR" with the technique enhanced and data analysis simplified by the development of detection via reverse hybridisation to probes against 43 selected individual spacer regions covalently bound to a nylon membrane. The variation in presence or absence of spacer regions provides different hybridisation patterns. Each spacer region has a complementary oligonucleotide probe on the nylon membrane so the technique is called spacer oligotyping or spoligotyping. *M.*

tuberculosis and *M. bovis* exhibit characteristically different spoligopatterns such that spoligotyping can differentiate between the two members of the *M. tuberculosis* complex. *M. bovis* strains characteristically lack the final four spacer sequencers 39-43 (Kamerbeek *et al.*, 1997). When strain patterns obtained from spoligotyping were compared to IS*6110* RFLP patterns in a study of 167 patients from three London hospitals in 1997, it was found that strains clustered by spoligotyping could be distinguished by IS*6110* RFLP, with RFLP demonstrating a generally higher level of differentiation. Spoligotyping is particularly useful for subtyping isolates with low copy numbers of IS*6110* (<5) as these strains exhibit a falsely high rate of clustering by RFLP (Goyal *et al.*, 1997).

The DR locus of *M. tuberculosis* has subsequently been assigned to a common family of repetitive DNA sequences called the clustered regularly interspaced short palindromic repeat (CRISPR) family that are present among archaea and bacteria but are absent in eukaryotes or viruses. There are also CRISPR-associated (*cas*) genes that are located directly adjacent to CRISPR loci. These genes may be involved in DNA metabolism or gene expression (Jansen *et al.*, 2002). Further studies have shown that together the *cas* genes and CRISPR loci may provide defence mechanism against bacteriophages (Barrangou *et al.*, 2007;Sorek *et al.*, 2008).

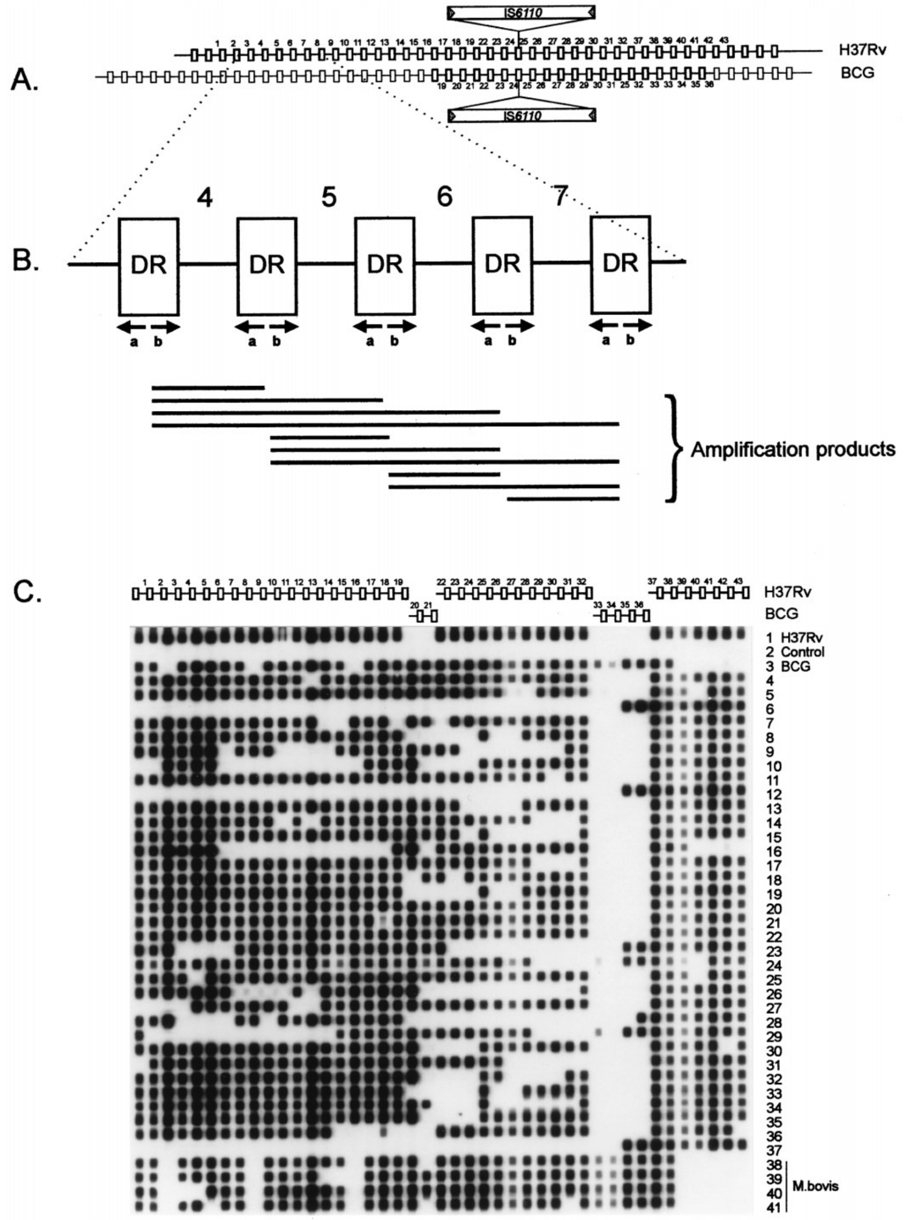
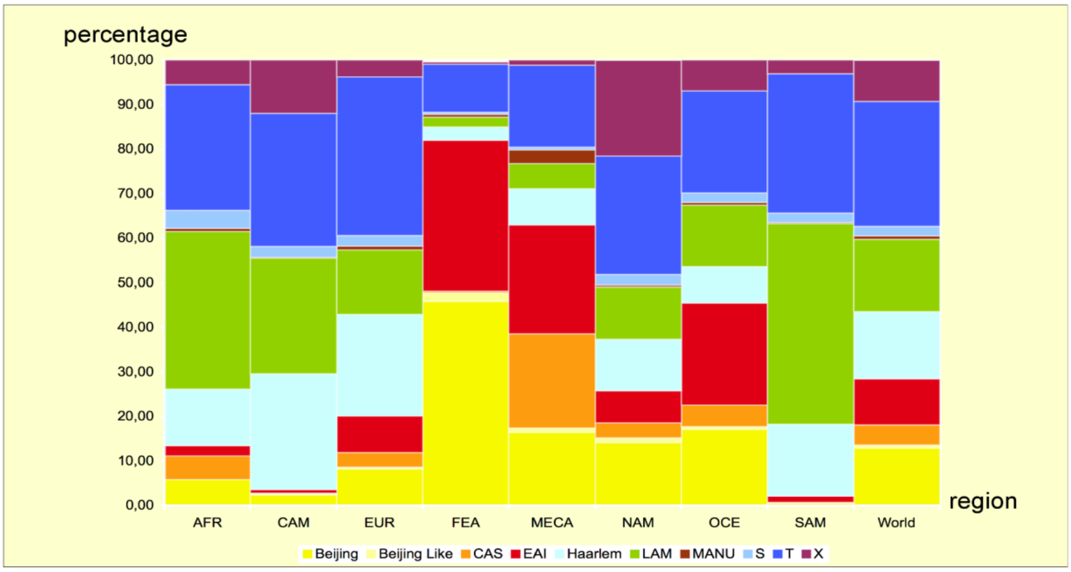


Figure 1.8 Spoligotyping.

The chromosomes of *M. tuberculosis* H37Rv and M. bovis BCG contain 48 and 41 DRs, respectively (depicted as rectangles), which are interspersed with unique spacers varying in length from 35 to 41 bp. The (numbered) spacers used correspond to 37 spacers from *M. tuberculosis* H37Rv and six from *M. bovis* BCG. The site of integration of insertion element IS*6110* is depicted. (B) Principle of in vitro amplification of the DR region by PCR. Any DR in the DR region may serve as a target for these primers; therefore, the amplified DNA is composed of a mixture of a large number of different-size fragments. Shown is the combination of fragments that would be produced by in vitro amplification of a DR target containing only five contiguous DRs. (C) Hybridisation patterns (spoligotypes) of amplified mycobacterial DNAs of 35 *M. tuberculosis* and five *M. bovis* strains (Kamerbeek *et al.*, 1997).

Spoligotyping is not as discriminatory as IS*6110* RFLP and is not used as a single first-line DNA fingerprinting method but as a complementary method to IS*6110* RFLP (Kremer *et al.*, 2005a). Since spoligotyping is a fast and robust genotyping technique it has been applied in multiple studies of *M. tuberculosis* complex strains around the world. Spoligotyping data from many studies have been consolidated into the "fourth international spoligotyping database (SpolDB4) for classification, population genetics and epidemiology" (Brudey *et al.*, 2006). This database contains spoligotyping data from 39,295 strains in 122 countries. There are a total of 5,309 individual spoligopatterns with 1,939/5,309 shared by more than one isolate and 3,370 orphan types. Using a mixed expert-based and bioinformatic approach, 62 global clades have been identified. This database has been used to construct a global phylogenetic relationship for *M. tuberculosis* and has led to insights into global strain transmission and migration.

There have been eight major *M. tuberculosis* spoligotyping clades identified: Beijing; Central Asian Strain (CAS); East-African Indian (EAI); Haarlem; Latin American and Mediterranean (LAM); X; and T. Each clade appears to be associated with prevalence in specific geographical continents (Figure 1.9): the Beijing strain represents about 50% of the strains in Far-East Asia and is emerging in Russia; CAS in the Middle-East/South Asia and specifically in India; EAI in Central/South-East/Far-East-Asia, the Middle East, and Oceania (Figure 1.10). The presence of Haarlem in Europe and also Central America/Caribbean suggests that there is a link with historical European colonisation. LAM is present in South America, coastal regions in the Mediterranean and the Caribbean (Figure 1.11). The X family in North America and Central America could be linked to Anglo-Saxon ancestry but is also currently correlated with African-Americans. T "ill-defined" T family is found on all continents as strains are designated as "T" by default (Brudey *et al.*, 2006).



Abbreviations AFR = Africa, CAM = Central America, EUR = Europe, FEA = Far-East Asia, MECA = Middle-East and Central Asia, NAM = North America, OCE = Oceania, SAM = South America

Figure 1.9. Percentage of main spoligotyping-defined MTBC genotype families within SpolDB4.

Region abbreviations: AFR = Africa; CAM = Central America; EUR = Europe; FEA = Far-East Asia; MECA = Middle-East and Central Asia; NAM = North America; OCE = Oceania; SAM = South America. Genotype family abbreviations: CAS = Central Asian Strain; EAI = East African Indian; LAM = Latin American Mediterranean (Brudey *et al.*, 2006).

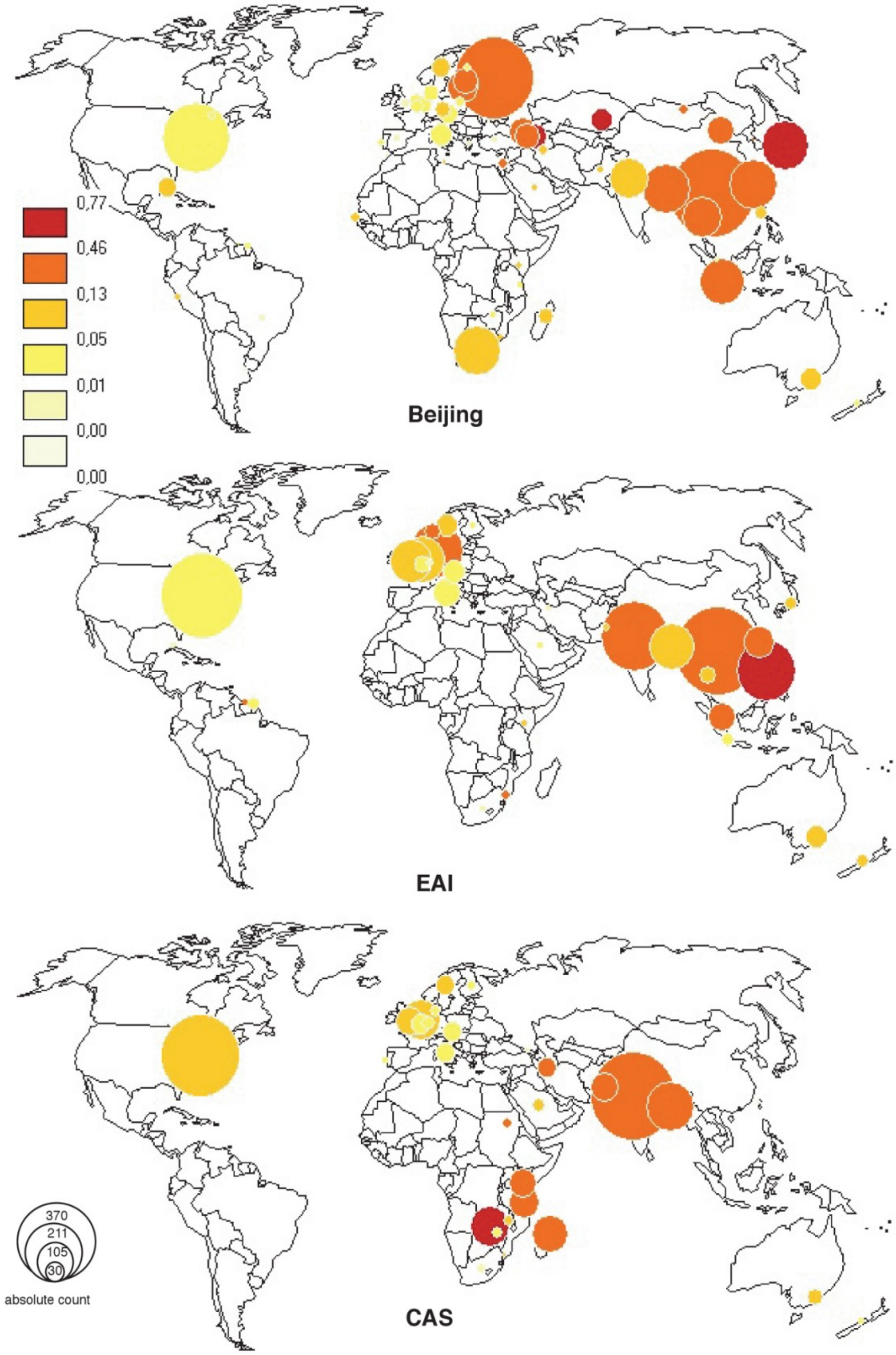


Figure 1.10. World maps showing the global distribution of Beijing, EAI, and CAS.

The diameter of each circle shows the absolute numbers and the colour represents proportions of total strain numbers within each country. The maps were built using an updated SpolDB4 on 14th September 2005, on clusters of the 50 most frequent shared types for a total of n = 17,212 isolates (Beijing n = 4,042, EAI n = 1,684, CAS n = 1,022) (Brudey *et al.*, 2006).

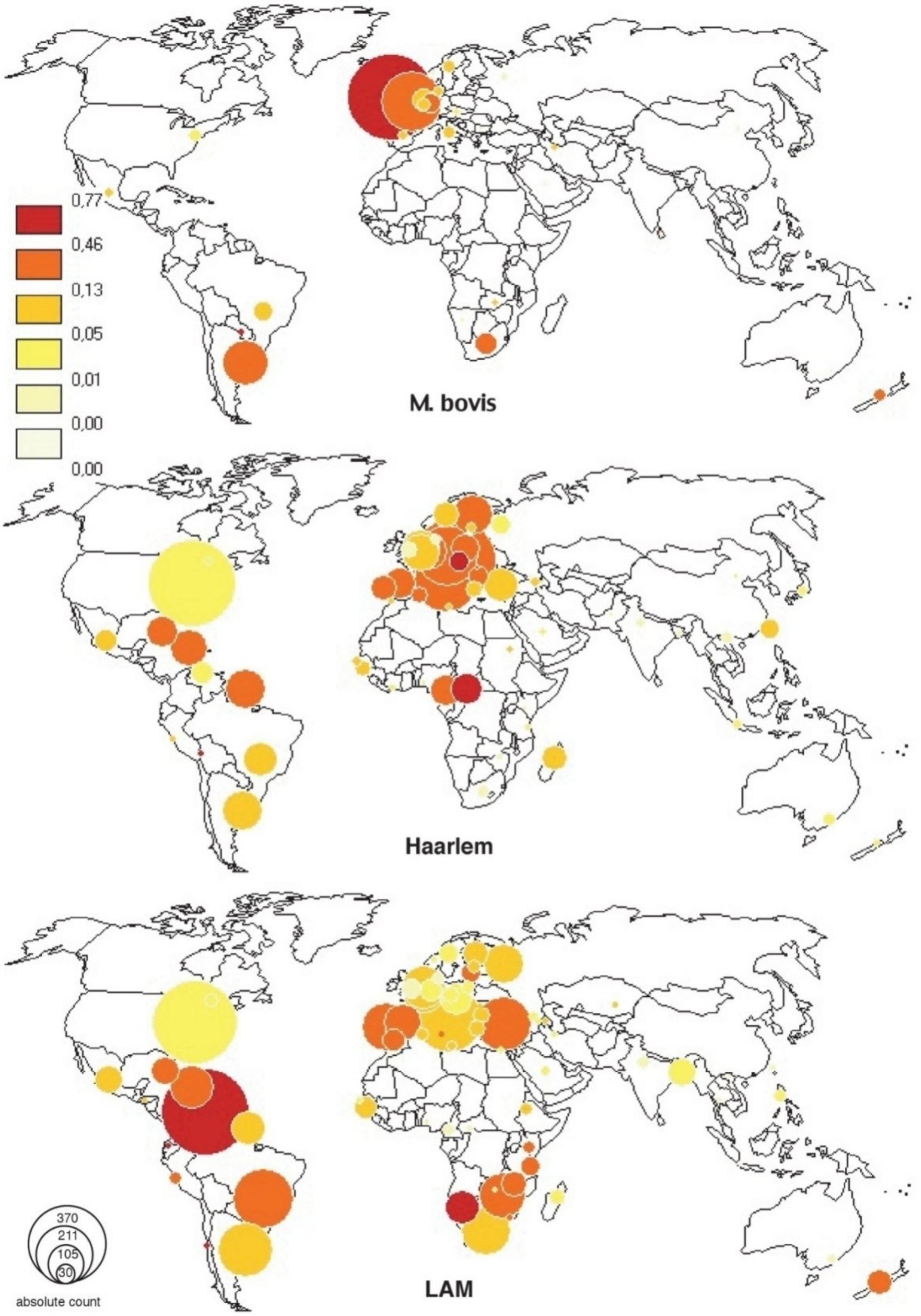


Figure 1.11. World maps showing the global distribution of *M. bovis*, Haarlem, and LAM.

The diameter of each circle shows the absolute numbers and the colour represents proportions of total strain numbers within each country. The maps were built using an updated SpolDB4 on 14th September 2005, on clusters of the 50 most frequent shared types for a total of n = 17,212 isolates (*M. bovis* n = 3888, LAM n = 3400, Haarlem n = 3176) (Brudey *et al.*, 2006).

Spoligotyping can be transferred to a high-throughput automated suspension array system (Luminex) by coupling the spacer oligonucleotides to 43 different microspheres which each contain slightly different proportional mixtures of two fluorochromes that enable the multiplex analysis of all 43 spacer oligonucleotides in a single well. Two lasers are used for detection: the first laser excites the fluorochromes within the microsphere which identifies the exact microsphere type and then a second laser excites a reporter molecule bound to the hybridised PCR product which enables quantification of the exact number of hybridised microspheres within seconds. The automated system definitively separates positive (present) and negative (absent) spacer regions without any overlap. The US CDC now routinely use the Luminex system to spoligotype all *M. tuberculosis* strains referred in the US (Cowan *et al.*, 2004). The multiplex capacity of the Luminex system has been utilised to develop an extended spoligotyping spacer set containing an additional 25 spacers to make a total of 68 spacer regions analysed. This enhanced spoligotyping panel has been used to further reconstruct and refine the global phylogenetic scenario for *M. tuberculosis* (Zhang *et al.*, 2010). Matrix-assisted laser desorption ionisation-time of flight mass spectrometry (MALDI-TOF MS) has also been used to increase the through-put of spoligotyping (Honisch *et al.*, 2010).

1.6.4. MIRU-VNTR typing

Variable number of tandem repeat (VNTR) sequences are multiple independent genetic loci found in at least 40 locations across the *M. tuberculosis* genome. Each locus contains repetitive DNA sequences that can differ in the number of repeats between strains (Figure 1.12 and Figure 1.13) (Supply *et al.*, 2000).

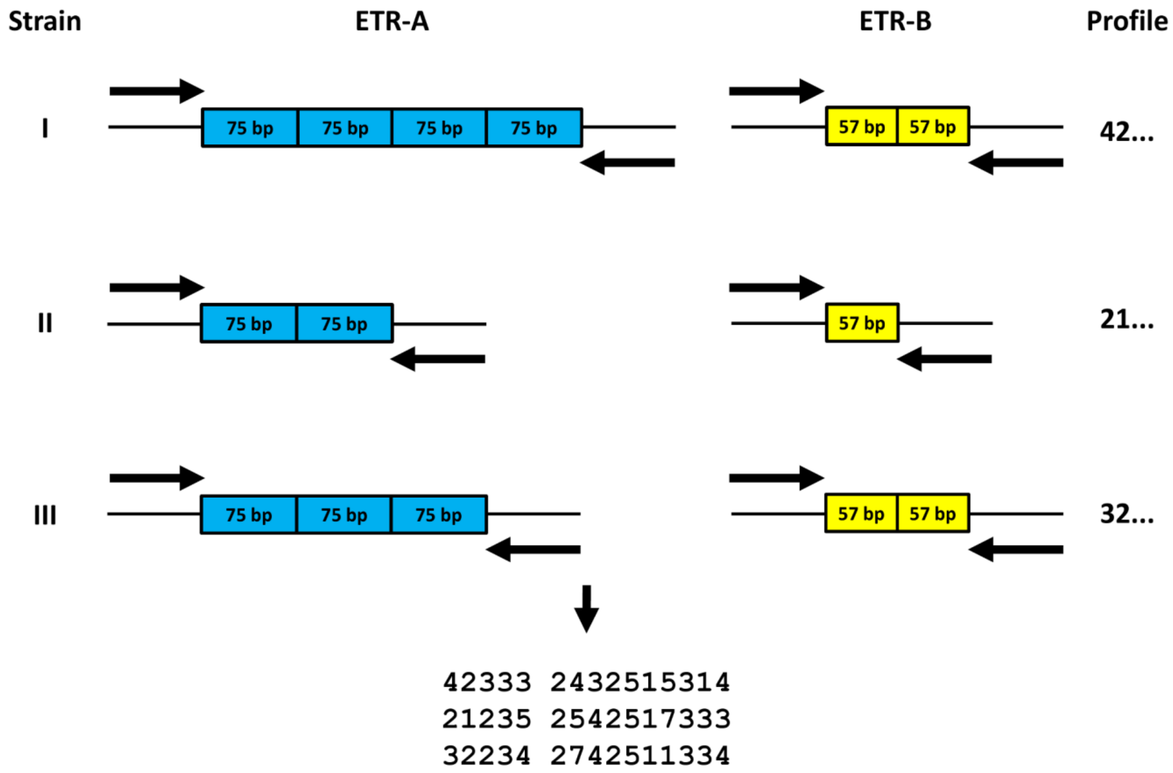


Figure 1.12. MIRU-VNTR typing.

Two example MIRU-VNTR loci are shown (ETR-A and ETR-B) in three example strains. Each locus is amplified by PCR using primers specific for the flanking region up- and downstream of the repetitive DNA sequences. Variation in the number of repeats results in varying sizes of PCR amplicons. From the sized PCR amplicon, the number of repeats at each locus can be calculated and concatenated to form the complete MIRU-VNTR profile.

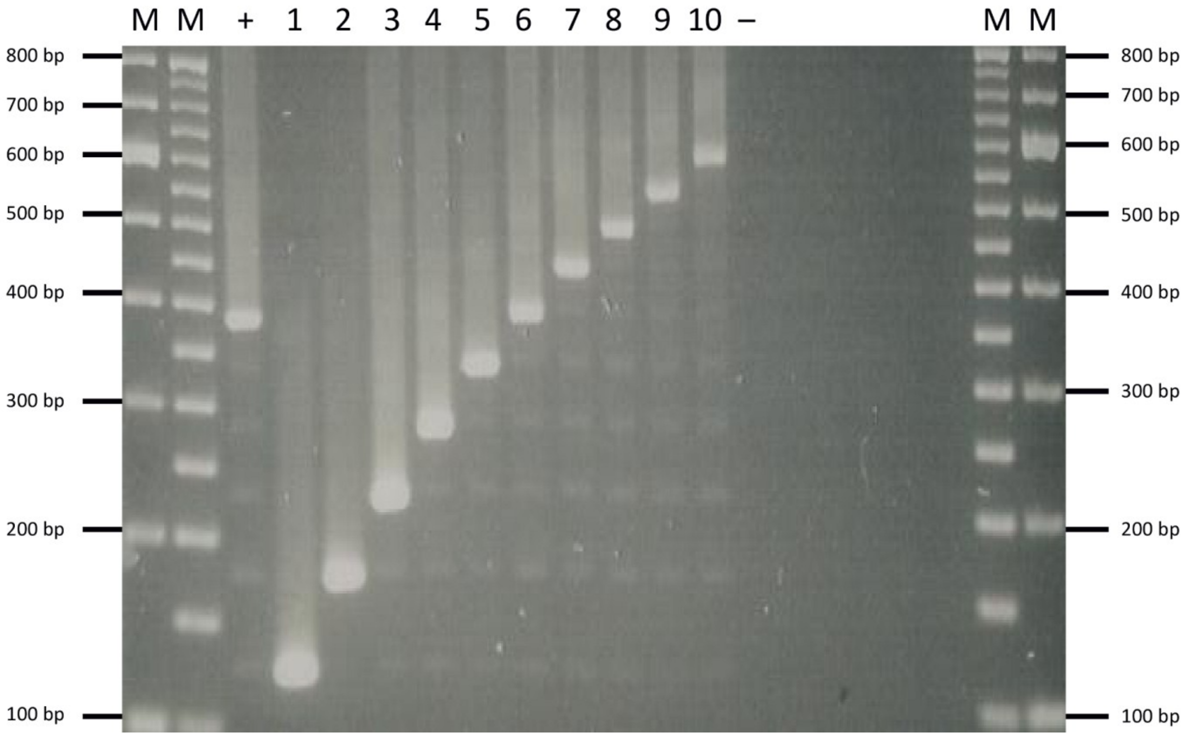


Figure 1.13 MIRU-23 alleles analysed by agarose gel electrophoresis.

The MIRU-23 locus in ten different *M. tuberculosis* strains was amplified by PCR and separated by agarose gel electrophoresis. Each strain had a different allele value which ranged from 1-10 repeats. The two markers used were a 50 bp and 100 bp molecular markers. The positive control was H37Rv.

VNTRs were first described in *M. tuberculosis* in 1994 through determination of the mechanism of isoniazid resistance caused by variations in the *katG* gene. Variation in the *katG* gene in both isoniazid sensitive and resistant strains was found to be caused by varying numbers of a 75-bp repeat sequence (Zhang and Young, 1994). VNTR loci were first used in human gene mapping and identification, forensic analysis and paternity testing (Jeffreys *et al.*, 1985). VNTRs have been described in or applied to other members of the *M. tuberculosis* complex including: *M. africanum* (Frothingham *et al.*, 1999); *M. bovis* (Skuce *et al.*, 2002); *M. canettii* (Fabre *et al.*, 2004); *M. caprae* (Prodinger *et al.*, 2005); and *M. pinnipedii* (Moser *et al.*, 2008). VNTRs have also been identified in NTMs including: *Mycobacterium avium* (Bull *et al.*, 2003); *Mycobacterium intracellulare* (Ichikawa *et al.*, 2010); *Mycobacterium leprae* (Groathouse *et al.*, 2004); and *Mycobacterium marinum/Mycobacterium ulcerans* (Stragier *et al.*, 2005). Other bacterial species that also possess VNTRs including include: *Bacillus anthracis* (Jackson *et al.*, 1997), *Staphylococcus aureus* (Hardy *et al.*, 2004), and *Clostridium difficile* (Marsh *et al.*, 2006). The number of repeats in each VNTR locus is calculated by PCR and fragment sizing. Each allele is then concatenated to form a VNTR profile (i.e. 42235) which can then be compared to other profiles to identify matches (Figure 1.12).

The first analysis of available *M. tuberculosis* H37Rv genome sequences before publication of the complete genome sequence identified two types of VNTR loci in a total of 11 tandem repeat loci. There were five major polymorphic tandem repeat (MPTR) loci that contained 15-bp repeats with substantial sequence variation in adjacent copies and six exact tandem repeat (ETR) loci that contained largely identical DNA repeats. Seven loci (one MPTR locus and all six ETR loci) exhibited length polymorphisms that were caused by insertions or deletions of tandem repeats and so could be analysed by agarose gel electrophoresis without requirement for elucidation of the full DNA sequence (Frothingham and Meeker-O'Connell, 1998). VNTR typing by PCR is a rapid and reproducible method that can be performed on crude DNA extracts and the data produced is digital which simplifies the comparison of large numbers of strains. From this it was proposed that VNTR typing in *M. tuberculosis* could be a useful alternative fingerprinting method to IS*6110* RFLP. However, a subsequent study showed that the discriminatory power of the ETR loci did not match that exhibited by IS*6110* RFLP (Kremer *et al.*, 1999). The publication of the whole genome of *M. tuberculosis* H37Rv (Cole *et al.*, 1998) enabled further genomic studies which identified additional mycobacterial interspersed repetitive units (MIRU) loci containing VNTRs (MIRU-VNTR) that increased the discriminatory power of VNTR typing in *M. tuberculosis* such that it began to approach that exhibited by IS*6110* RFLP (Mazars *et al.*, 2001;Smittipat and Palittapongarnpim, 2000).

MIRU-VNTR typing using 12 loci was shown to be very stable and reproducible. In a study of 123 serial isolates from 56 patients over six years, 55/56 (98.2%) patients had identical 12 locus MIRU-VNTR profiles whereas 11/56 (19.6%) of patients had varying RFLP profiles (Savine *et al.*, 2002). Two studies of historically distant BCG strains from around the world showed that only one locus (ETR-D/MIRU-04) exhibited wide variation between BCG strains (Supply *et al.*, 2000).

Variations in VNTR loci may be caused by an increase or reduction in the number of tandem repeats by slipped-strand mispairing (Levinson and Gutman, 1987). It is thought that slippage during DNA replication in the template DNA strand reduces the number of repeats and increases repeat number if slippage occurs in the replicated strand. A recent study that analysed the number of repeats per locus in >400 isolates indicated that there appears to be a general global trend towards reduction in the number of repeats in modern *M. tuberculosis* strains (Arnold *et al.*, 2006). Although in a closed population in South Africa where multiple MIRU-VNTR profiles arose from a single ancestral genotype, MIRU-VNTR variation was hypothesised to have occurred in both directions which resulted in an increase or decrease in the number of repeats (Warren *et al.*, 2004).

MIRU-VNTR typing has been used to: identify a variant Beijing strain in New York (Bifani *et al.*, 1999); identify laboratory contamination (Gascoyne-Binzi *et al.*, 2001); subdivide clusters containing isolates with few copies of IS*6110* (Barlow *et al.*, 2001); and identify mixed infections (Shamputa *et al.*, 2006).

The Beijing strain is one of the most prevalent global clades of *M. tuberculosis* and it was initially thought that the development of a rapid highly discriminatory method such as MIRU-VNTR in combination with spoligotyping would greatly enhance epidemiological studies of this strain (Sola *et al.*, 2003;van Soolingen *et al.*, 1995). However, subsequent studies have shown that the internationally recommended VNTR loci set did not provide sufficient differentiation of the Beijing strain. Additional VNTR loci can provide further differentiation (Nikolayevskyy *et al.*, 2006). Many of these additional hypervariable VNTR loci have been excluded from other studies as their stability and reproducibility have been questioned (Supply *et al.*, 2006). The true epidemiological significance of the apparent enhanced discrimination offered by additional VNTR loci has not been fully evaluated as many of these studies have focussed solely on strain differentiation and lack the associated epidemiological data that could confirm the relevance of the enhanced differentiation (Mokrousov *et al.*, 2009;Dou *et al.*, 2009;Hanekom *et al.*, 2008;Wada *et al.*, 2007).

The exact function of VNTR loci in *M. tuberculosis* is not well understood as most loci are located in non-coding regions. For those VNTR loci that are located within an open reading frame (ORF), several proposed possible functions include: a source of antigenic variation; regulation of gene expression; differential translation of genes; or acting as structural components in genome organisation (Supply *et al.*, 1997).

VNTR loci in *M. tuberculosis* are similar to short sequence repeats in other bacteria which can modulate gene expression (van Belkum, 1999). In a comparison of the differences between protein expression between H37Rv and CDC1551 grown *in vitro*, differences in protein expression were found in only 15/1,750 (0.85%) of all soluble protein fractions identified by 2D electrophoresis (Betts *et al.*, 2000). Although *in vivo* gene expression may vary between strains, this remarkable similarity suggested that the patterns of gene regulation in *M. tuberculosis* are highly conserved (Brosch *et al.*, 2001).

Of the current recommended set of 24 MIRU-VNTR loci, 11 have repeat units which are a multiple of three basepairs. Therefore, variations in these loci will not disrupt ORFs. Most loci (15) are located within putative genes but only four loci are located in genes with a defined function (Le Fleche *et al.*, 2002). The *leuA* gene in *M. tuberculosis* encodes alpha-isopropylmalate synthase (α-IPMS) which is a key enzyme in the first step of the leucine biosynthetic pathway. The primary substrates for α-IPMS are alpha-ketoisovalerate and acetyl CoA. The *leuA* gene contains a VNTR locus with a 57 bp repeat motif. A recent study on the effect of low (2 repeats) or high (14 repeats) allele number in the *leuA* gene found that this variation in repeat number affected protein structure and function. The strain with 14 repeats tolerated wider pH and temperature ranges and exhibited a higher affinity for alpha-ketoisovalerate and acetyl CoA (Yindeeyoungyeon *et al.*, 2009). A different study on the formamidopyrimidine (faPy)-DNA glycosylase (Mtb-fpg1) that corrects oxidative damage in purines and pyrimidines showed that an increased number of tandem repeats in an upstream region upregulated the expression of the *Mtb-fpg1* gene (Olsen *et al.*, 2009). A similar relationship was identified with VNTR3690 and the *lpdA* gene (Akhtar *et al.*, 2009).

1.6.5. SNPs, Regions of Difference, and Whole Genome Analysis

The first differences in the genome content of different members of the *M. tuberculosis* complex, *M. bovis* BCG, and several NTMs were initially identified by subtractive hybridisation. This showed that there were three large chromosomal regions present in virulent *M. bovis* that were absent in a *M. bovis* BCG vaccine strain. Southern hybridisation identified the presence of all three regions in virulent *M. tuberculosis* H37Rv and the *M. tuberculosis* Erdman strain but these regions could not be detected in *M. avium* or *M. smegmatis*. These large chromosomal regions (9.3-10.7-kb) were termed genomic regions of difference (RD) (Mahairas *et al.*, 1996).

1.6.6. Identification of three Principal Genetic Groups based on SNPs in *katG* and *gyrA*

The first differences due to SNP variations in *M. tuberculosis* were identified by analysis of a total of two megabases of DNA sequence data in 26 regions which identified three principal genetic groups (PGG) based on two SNPs in genes encoding for catalase-peroxidase (*katG*) and subunit A of the DNA gyrase gene (*gyrA*) (Figure 1.14). Group 1 and 3 organisms were found to be evolutionarily distinct with group 1 or 2 strains but not group 3 organisms associated with large clusters of cases. This study suggested that *M. tuberculosis* is evolving toward a state of reduced transmissibility or virulence (Sreevatsan *et al.*, 1997).

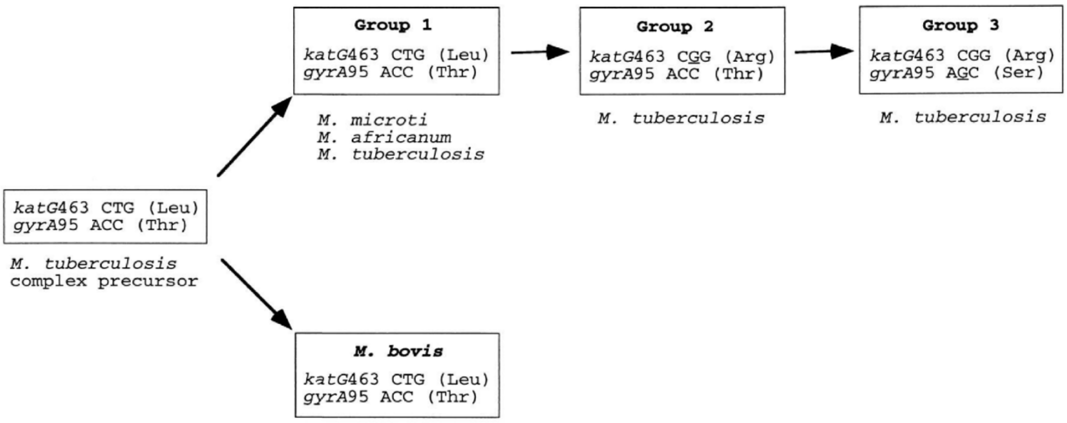


Figure 1.14. Initial evolutionary scenario for *M. tuberculosis* complex organisms based on three SNP genotypes.

The precursor of *M. tuberculosis* complex organisms was characterised by SNPs present in *katG* codon 463 (Leu) and *gyrA* codon 95 (Thr) (Sreevatsan *et al.*, 1997).

1.6.7. The complete genome sequence of *M. tuberculosis* and other mycobacteria

Whole genome analysis of *M. tuberculosis* was greatly enhanced by the publication of the complete genome sequence of *M. tuberculosis* laboratory strain H37Rv in 1998. It was found that the entire genome is 4,411,529 bp long, contains about 4,000 genes, and has a very high GC content (Cole *et al.*, 1998). Comparison of H37Rv and complete genome sequences of clinical *M. tuberculosis* strains that were subsequently characterised have shown that SNPs are rare and insertion or deletion events are the principal source of genomic variation. Insertion and deletions are caused by insertion sequences, expansion or reduction in repetitive DNA sequences or replication errors (Fleischmann *et al.*, 2002).

The complete genome sequence of *M. marinum* was published in 2008. This mycobacterial species has a broad host range and is a ubiquitous pathogen of fish and amphibians and is related to *M. tuberculosis*. The chromosomal genome of *M. marinum* was found to be 6,636,827-bp long with a 23-kb plasmid present also. *M. marinum* has been widely used as a model to study *M. tuberculosis* pathogenesis and comparison of these two complete genomes identified an average amino acid homology of 85% which indicated that both species have a recent common genetic ancestor. It was hypothesised that *M. marinum* has retained more of the genetic content of the common ancestor compared to the reductive evolution and adaptation undergone by *M. tuberculosis* within the human host (Stinear *et al.*, 2008). *M. leprae* has undergone even further genomic reduction when compared to the *M. tuberculosis* genome (4.41 Mb) as the genome is only 3.27 Mb long with less than half of the *M. leprae* genome containing functional genes (Cole *et al.*, 2001).

Within the *M. tuberculosis* complex, various insights have been achieved by the analysis of complete genome sequences. *M. tuberculosis* H37 was first isolated in 1905 and was noted for its virulence in the guinea pig model which was used in the classification of "human tuberculosis" in the early 20th century. Serial passage of H37 through media at different pH levels dissociated this strain into two variants that have different colony morphologies: virulent H37Rv (rough) and attenuated H37Ra (smooth) (Steenken *et al.*, 1934;Bifani *et al.*, 2000). These two strains have been widely used in studies of mycobacterial virulence. However, it is only relatively recently that the genetic basis for the phenotypic difference between H37Rv and H37Ra laboratory strains has been elucidated (Brosch *et al.*, 1999;Lee *et al.*, 2008;Zheng *et al.*, 2008). Complete genome sequences have also enabled the differentiation of the original and current modern *M. bovis* BCG vaccine strains (Seki *et al.*, 2009;Brosch *et al.*, 2007;Pym *et al.*, 2002), and the construction of a new evolutionary scenario for the appearance of *M. tuberculosis* and *M. bovis*.

1.6.8. *M. tuberculosis* can be divided into ancestral and modern lineages

A study of 20 variable RDs generated from insertion/deletion events were analysed in 100 strains of the *M. tuberculosis* complex including *M. tuberculosis*, *M. africanum*, *M. canettii*, *M. microti*, and *M. bovis* (Figure 1.15). RDs are generated primarily by IS*6110* insertion, subsequent homologous recombination and excision of intermediary DNA that occur in a unidirectional manner. It was discovered that these events had occurred in common progenitor strains defined by the presence or absence of a *M. tuberculosis* specific deletion called "TbD1". This single region was used to divide the *M. tuberculosis* complex into ancestral (TbD1 is present) and modern (TbD1 absent) strains. Modern strains include major global clades such as the Beijing and Haarlem strains. Deletion of RD9 and other subsequent deletions occurred in a lineage now containing *M. africanum*, *M. microti*, and *M. bovis* that occurred before divergence from *M. tuberculosis* and before alterations in TbD1 occurred. This study contradicted the previous hypothesis that *M. tuberculosis* had evolved from *M. bovis*. The genomes of *M. canettii* and TbD1+ (ancestral) *M. tuberculosis* genomes were the most intact and had not undergone as much genetic reduction through loss of RD loci and appeared to be closest to a most recent common ancestor for the *M. tuberculosis* complex

(Brosch *et al.*, 2002). Subsequent studies that combined RD9, TbD1 and PGG SNP data identified that RD9 was deleted in the ancestral PGG1 group but not in the modern PGG2/3 group whereas TbD1 was present in ancestral strains but deleted in modern strains. The deletion of RD9 and presence of TbD1 in ancestral strains indicated that deletion of RD1 occurred before deletion of TbD1 and the generation of modern lineages (Huard *et al.*, 2006).

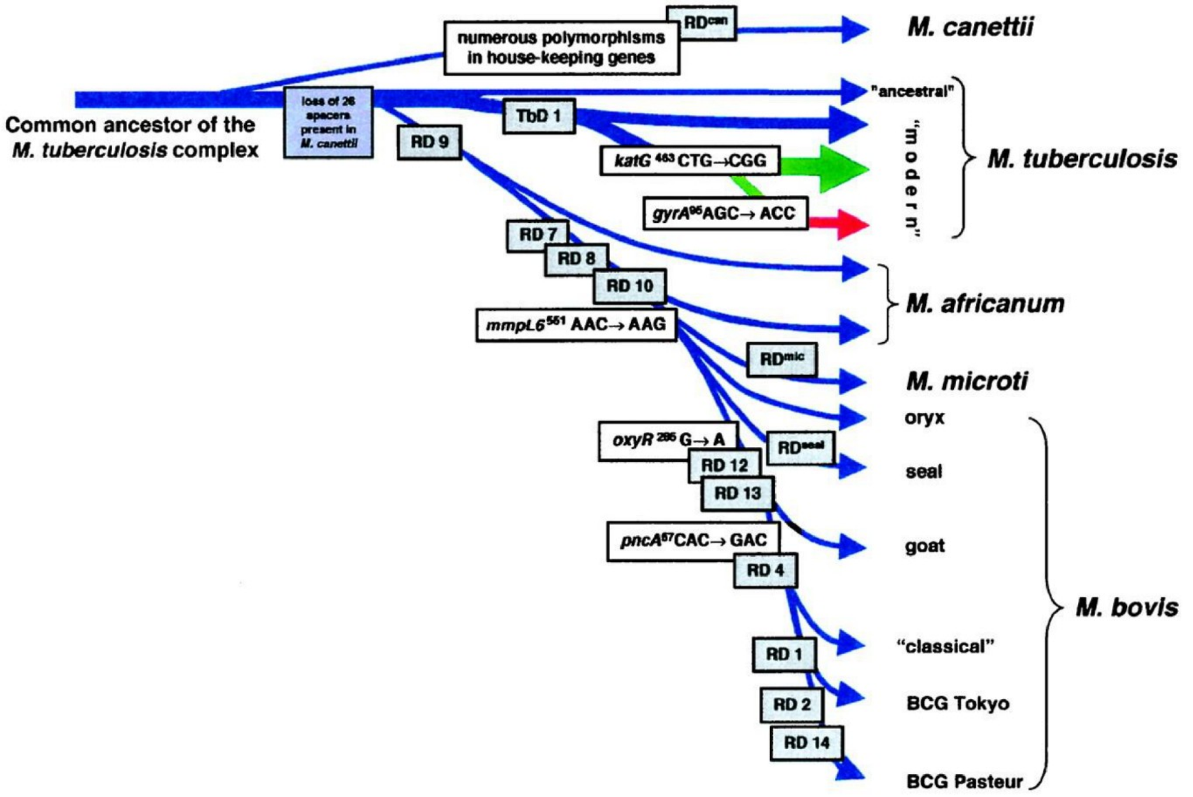


Figure 1.15. Scheme of the proposed evolutionary pathway of the tubercle bacilli illustrating successive loss of DNA in certain lineages.

The scheme is based on the presence or absence of conserved deleted regions and on sequence polymorphisms in five selected genes. Note that the distances between certain branches may not correspond to actual phylogenetic differences calculated by other methods. Blue arrows indicate that strains are characterised by *katG* 463. CTG (Leu), *gyrA* 95 ACC (Thr), typical for group 1 organisms. Green arrows indicate that strains belong to group 2 characterised by *katG* 463 CGG (Arg), *gyrA* 95 ACC (Thr). The red arrow indicates that strains belong to group 3, characterised by *katG* 463 CGG (Arg), *gyrA* 95 AGC (Ser) (Brosch et al., 2002).

1.6.9. The origin and ancestry of *M. tuberculosis*

M. canettii represents the most well-defined member of the smooth tubercle bacilli in the *M. tuberculosis* complex. After it was identified as an ancestral lineage to the rest of the *M. tuberculosis* complex by analysis of RDs, an analysis of the variation in six housekeeping genes (16S rRNA, *katG*, *gyrA*, *gyrB*, *hsp65*, *rpoB*, and *sodA* genes) of 37 smooth tubercle bacilli predominantly from Djibouti in East Africa compared to nine other members of the MTBC was undertaken (Gutierrez *et al.*, 2005). It was found that the genetic diversity of the smooth bacilli was far greater than other members of the *M. tuberculosis* complex. Combined with the RD distribution, this placed the appearance of the smooth bacilli before the rest of the *M. tuberculosis* complex. This meant that at some point in time an evolutionary bottleneck had occurred where the diversity of the mycobacterial population was previously far greater than that of current modern *M. tuberculosis* strains which now represent a small proportion of the total diversity once exhibited by the ancestral strains of *M. tuberculosis*. *M. prototuberculosis* was proposed as the name for this new highly diverse species that is a progenitor of *M. tuberculosis*.

The combined sequences of the six housekeeping genes were compared within the smooth bacilli and to the rest of the *M. tuberculosis* complex. The *M. prototuberculosis* housekeeping genes appear to be composed of a mosaic of gene sequences that originated from different eight subtypes of *M. canettii* whereas the other members of the *M. tuberculosis* complex had completely homologous gene sequences without any variation. This suggested that *M. prototuberculosis* had undergone multiple horizontal gene transfer events, a function that is absent in modern *M. tuberculosis*. (Gutierrez *et al.*, 2005). Also, there is no convincing evidence to date for the presence of plasmids in *M. tuberculosis*. Extrachromosomal DNA which was presumed to be a plasmid was visualised by agarose gel electrophoresis but this has never been fully isolated or characterised (Crawford and Bates, 1979;Zainuddin and Dale, 1990). The transfer of genomic DNA in *M. tuberculosis* under natural conditions has not yet been demonstrated (Balganesh *et al.*, 2010).

The increased nucleotide diversity in the ancestral strains also greatly increased the estimate of when a progenitor of *M. tuberculosis* originally appeared from 35,000 years to 2.6-2.8 million years (Hughes *et al.*, 2002). The new evolutionary age of 2.6-2.8 million years and the geographical restriction of smooth tubercle to East Africa where early human ancestors were present three million years ago has led to the hypothesis that both humans and tubercle bacilli emerged in Africa (Semaw *et al.*, 2005).

Further studies that examined 89 genes in 108 strains that originated from around the world found that the level of genetic diversity in *M. tuberculosis* was higher than previously thought (Hershberg *et al.*, 2008). This diversity was linked to the emergence of ancient human populations in Africa about 40,000 years ago and subsequent migration. A second study identified that “modern” strains of *M. tuberculosis* emerged about 10-20,000 years later (Wirth *et al.*, 2008).

1.6.10. Clinical application of genome-informed assays

Large sequence polymorphisms (LSP) are deletion events like RDs but they encompass all other deleted regions that are not part of the well-defined RD set (Mahairas *et al.*, 1996). The analysis and identification of LSPs has been utilised to identify clinically important strains of *M. tuberculosis*. Analysis of five LSPs enabled the genome level-informed identification by

PCR of an outbreak strain in Leicester with a different set of LSPs bound to a nylon membrane in reverse-line hybridisation assay which could identify deletions in 43 genomic regions in up to 40 strains at one time (Rajakumar *et al.*, 2004;Goguet de la Salmoniere YO *et al.*, 2004; Cardoso Oelemann, 2011). More recently, a prevalent strain in Rio de Janeiro was identified through the detection of a solitary 26 kb deletion event (Lazzarini *et al.*, 2007).

1.6.11. High-throughput Whole Genome Sequencing

New sequencing technologies have been developed using a variety of strategies that generate DNA sequence data using massively parallel reactions that generate millions of bases from a single experiment at a reduced cost (Pallen *et al.*, 2010). High-throughput whole genome sequencing (WGS) is starting to be applied to *M. tuberculosis* strains to examine strain transmission, drug resistance, global phylogeny of the Beijing strain, and mutation rates in latent infection. WGS has been applied to the first and last isolates of one small cluster of five patients over a 12 year period and one large cluster in the Netherlands and one outbreak in British Columbia. In the small Dutch cluster, it was found that a total of six polymorphisms were present which consisted of four SNPs, a variation in a VNTR locus (not one that is internationally analysed), and a known transposition of IS*6110*. Surprisingly, 5/6 polymorphisms arose first in a single patient who had been non-compliant. Within this cluster, *M. tuberculosis* had been relatively stable with a burst of mutation (Schurch *et al.*, 2010b). Three isolates from a total of 104 in one of the largest clusters in the Netherlands (Harlingen) were selected for WGS which identified eight SNPs. Traditional contact tracing had identified two routes of transmission in this outbreak whereas application of WGS provided greater resolution with five SNP subclusters identified (Schurch *et al.*, 2010a). In an earlier study, two apparently clonal Beijing strains from Uzbekistan were sequenced and considerable diversity was identified with 130 SNPs present. However, these two strains did not share the exact same fingerprinting pattern as 23/24 MIRU-VNTR loci were indistinguishable (Niemann *et al.*, 2009). It may be that this variation is indicative of the genetic distance between strains that are closely related.

WGS was combined with social network analysis and applied to 32 isolates in an outbreak in British Columbia that had an indistinguishable MIRU-VNTR and RFLP genotype. WGS and social network analysis identified two lineages and clarified the potential routes of transmission (Gardy *et al.*, 2011). In two independent studies in South Africa, WGS was used to examine the distribution of mutations in the KwaZulu-Natal (KZN) XDR and Beijing strains. WGS showed that MDR and XDR strains did not share the same mutations conferring resistance to rifampicin and pyrazinamide but XDR strains from various locations had identical drug resistance mutations. Therefore, MDR strains and XDR had acquired mutations independently which excluded gradual increment in drug resistance within each patient from MDR to XDR but rather that a single XDR clone had emerged and expanded (Ioerger *et al.*, 2009). Conversely, WGS of 14 XDR Beijing strains from two separate clusters identified that resistance mutations appeared multiple times independently within each of the two clusters. Therefore, drug resistance in these two clusters had been acquired within each patient and were not a result of increased rate of mutation and clonal expansion of a single drug-resistant strain (Ioerger *et al.*, 2010). WGS of six Beijing strains from different countries combined with RFLP patterns identified that about 80% of Beijing strains were closely related and represented a monophyletic lineage that has recently appeared and undergone significant clonal expansion to become one of the most prevalent strains in the world today (Schurch *et al.*, 2011).

The accumulation of mutations in active and latent disease was compared by sequencing 33 *M. tuberculosis* isolates from nine experimentally infected macaques. It was found that there was a similar mutation rate during latency as during active disease or in a log phase culture. From the distribution of polymorphisms, it was deduced that the mutational changes were caused by oxidative damage in the absence of mycobacterial replication (Ford *et al.*, 2011). This finding has several implications including the use of WGS in strain transmission studies as a strain infecting an individual, not presenting as primary disease initially, and eventually reactivating could possess a different SNP pattern to other strains within the same cluster.

1.6.12. Clinical utility of DNA Fingerprinting

DNA fingerprinting is an essential part of a TB control programme as it can estimate transmission in specific populations groups, focus interventions by identifying factors for transmission and can be used to evaluate the effectiveness of targeted interventions to prevent transmission.

The Netherlands have undertaken universal prospective DNA fingerprinting of all *M. tuberculosis* isolates since 1993. This enabled the identification of transmission after migration in patients recently arrived in the Netherlands with 30-40% of cases in Turkish, Moroccan, and Somali patients attributed to recent transmission. (Borgdorff *et al.*, 2001).

Analysis of six years worth of data obtained from nationwide tuberculosis contact investigation and DNA fingerprinting surveillance in the Netherlands showed that in 2,206 clustered cases, DNA fingerprinting increased the number of epidemiologic links from 462 before DNA fingerprinting data to 1,002 epidemiologic established links after cluster investigation which involved the combination of molecular and epidemiological data. DNA fingerprinting did not extend contact tracing investigations. Cluster monitoring did enable the identification of transmission events not detected by contact investigations, the development of focused interventions, and the evaluation of regional tuberculosis eradication programmes (Lambregts-van Weezenbeek *et al.*, 2003).

A third study in the Netherlands identified an association between rapidly expanding clusters that expanded from two patients to five or more patients within two years and patient epidemiological factors. If the first two patients in a cluster were identified within three months of each other, one or both were <35 years old, and both patients resided in an urban area and originated from Sub-Saharan Africa, there was a more than five times increased probability that a strain identified in an initial cluster of two paired patients would be identified in five or more patients within two years (Kik *et al.*, 2008).

A study in the US state of Maryland showed that cluster investigation which involved the combination of conventional and molecular epidemiological data increased the number of clusters with known links between patients by 61% from 70 using conventional epidemiological data alone to 113. This system also detected that 29% of case pairs defined by conventional epidemiological data were not actually the same strain (McNabb *et al.*, 2004).

1.7. BACKGROUND TO THE PROJECT

The rates and case numbers of *M. tuberculosis* in the Midlands have been slowly and consistently increasing since the late 1980's. Conventional epidemiological data for TB cases in the Midlands are collected and analysed on an annual basis (Health Protection Agency, 2010). However, the molecular epidemiology of *M. tuberculosis* in the Midlands has yet to be fully elucidated. Examining the molecular epidemiology will enable prevalent global lineages and specific strains to be identified within the Midlands as well as providing an insight into transmission routes.

1.8. AIMS OF THESIS

The principal aim of this thesis was to understand the epidemiology and transmission of *M. tuberculosis* in the Midlands.

The specific objectives were to:

- Evaluate IS*6110* RFLP data in comparison to MIRU-VNTR data to identify if MIRU-VNTR typing could be a rapid alternative method to IS*6110* RFLP.
- Develop automated analysis of MIRU-VNTR loci using non-denaturing High Performance Liquid Chromatography (non-dHPLC) which would enable the analysis of all *M. tuberculosis* strains in the Midlands and enhance the knowledge and understanding of the epidemiology and transmission of *M. tuberculosis*.
- Identify the prevalent global lineages present in the Midlands and the continental origin of infected patients using non-dHPLC, MIRU-VNTR, and a novel computer software programme called Origins.
- Identify, confirm, and analyse the distribution of the single most prevalent MIRU-VNTR profile in the Midlands.
- Undertake a whole-genome analysis of prevalent strains in the Midlands using DNA-DNA microarray hybridisation to identify properties unique to these strains and the genetic origin of these strains.

2. MATERIALS AND METHODS

2.1. REAGENTS AND SUPPLIERS

Reagents were purchased from Sigma-Aldrich (Gillingham, Dorset, UK) unless otherwise stated. Nucleotides for PCR were purchased from Eurofins MWG Operon (Ebersberg, Germany). For each DNA polymerase used, the provided PCR buffer and $MgCl_2$ was also used. PCR amplifications were carried using MultiBlock System Thermal Cyclers from Thermo Fisher Scientific (Loughborough, Leicestershire, UK). A list of solutions frequently used in this thesis can be found in Appendix I.

2.2. STUDY POPULATION

This PhD will focus on *M. tuberculosis* strains isolated from across the West and East Midlands. The Midlands region of the UK encompasses the West and East Midlands (Figure 1.3) which had a total population of 5.4 and 4.4 million people respectively in 2009 (Office for National Statistics, 2010). There are 82 counties in England with six counties in the West Midlands region (Herefordshire, Shropshire, Staffordshire, Warwickshire, West Midlands county, Worcestershire) and five counties in the East Midlands (Derbyshire, Leicestershire, Lincolnshire, Northamptonshire, and Nottinghamshire) (Figure 2.1.). The West Midlands exists on two geographical levels as a region but also as a county. Community health care in England is currently organised and provided by 151 Primary Care Trusts (PCT) which work with local authorities (LA) to provide health and social care in local communities. There are 17 PCTs in the West Midlands and nine PCTs in the East Midlands (Figure 2.2.A) with nine PCTs located in the county of the West Midlands (Figure 2.2.B).

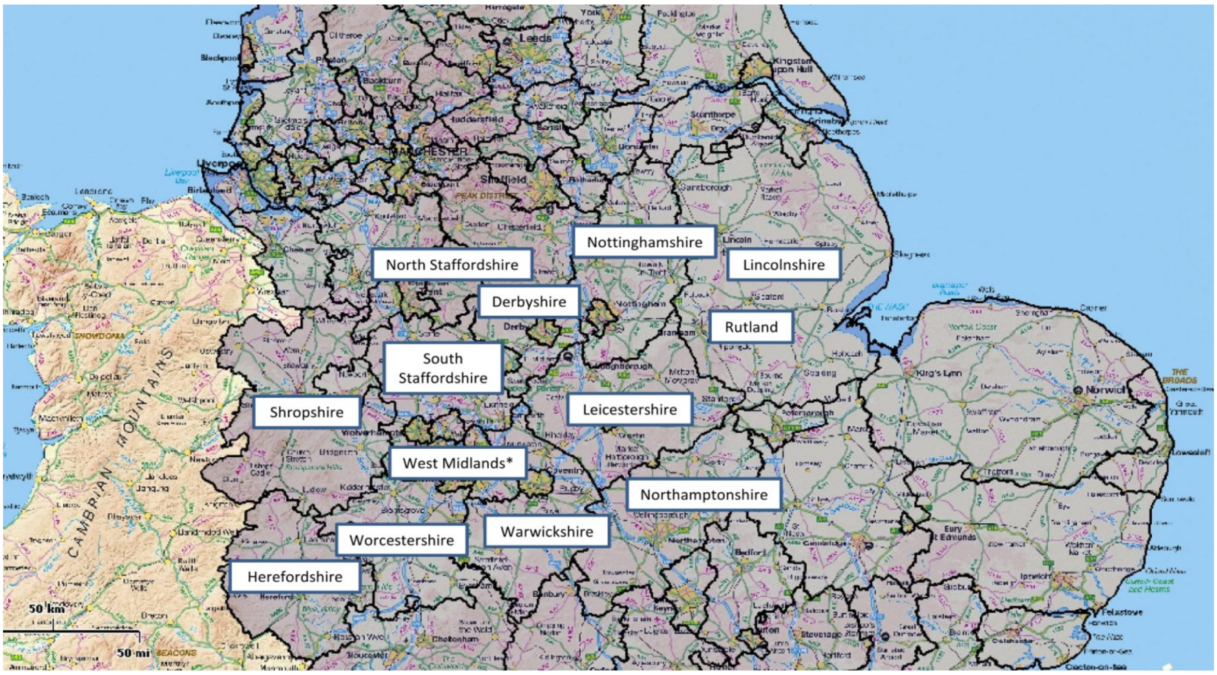
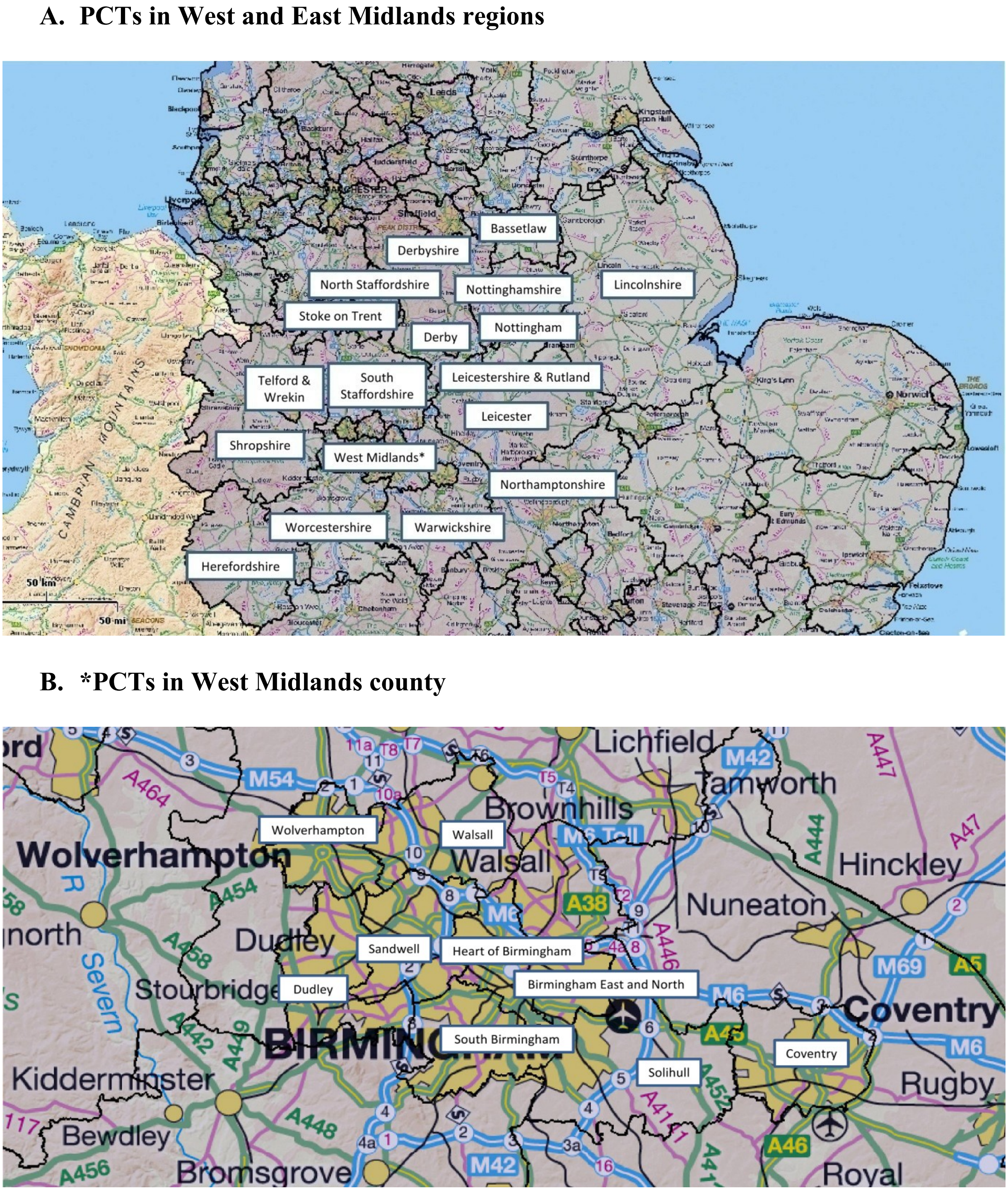


Figure 2.1. Counties in the West and East Midlands.

This map was generated using the HPA Geographic Information System Intranet Application and under license by © Crown copyright and database right 2010. All rights reserved. Ordnance Survey Licence number 100016969



A. PCTs in West and East Midlands regions

B. *PCTs in West Midlands county

Figure 2.2. Primary care organisation boundaries in the West and East Midlands.

This map was produced under license by ©Crown Copyright and database right 2010 using the HPA Geographic Information System Mapping System. Ordnance Survey License 100016969/100022432.

2.3. INITIAL SPECIMEN COLLECTION, PROCESSING, AND RESULT REPORTING

Specimens received at the MRCM were processed as described in the HPA National Standard Method BSOP40 (Evaluations and Standards Laboratory, 2006).

2.3.1. Mycobacterial culture using solid media

For each isolate, an approximate 20 ml universal containing a slope of LJ Medium (Media for Mycobacteriology, Penarth, South Glamorgan, UK) was inoculated. The formula for 600 ml was as follows: 2.5 g monopotassium phosphate; 0.24 g magnesium sulphate; 0.6 g sodium citrate; 3.6 g L-asparagine; 30 g potato flour; 0.4 g malachite green; 12 ml glycerol. The final step of preparation was the addition of 1,000 ml whole egg to the rest of the 600 ml.

Egg-based media support the growth of a wide variety of mycobacteria. L-asparagine and potato flour provide nitrogen and vitamins, monopotassium phosphate and magnesium sulphate ac as buffers, glycerol and the egg suspension provide fatty acids and proteins required for mycobacterial metabolism, and sodium citrate and malachite green act as selective agents to prevent the growth of non-mycobacterial contaminants.

2.3.2. Mycobacterial culture using liquid media

Each MGIT (Becton-Dickinson, Oxford, UK) contained 7 ml of modified Middlebrook 7H9 broth base. The complete medium included: 7H9; oleic acid, albumin, dextrose, catalase (OADC); and polymixin B, amphotericin B, nalidixic acid, trimethoprim, and azlocillin (PANTA). Each MGIT contained 110 µl of fluorescent indicator (Tris 4,7-diphenyl-1,10-phenanthroline ruthenium chloride pentahydrate in a silicone rubber base) and 7 ml of 7H9 broth (5.9 g Modified Middlebrook 7H9 Broth base and 1.25 g Casein peptone per 1 l of purified H_2O). For a batch of 18 MGITs, 15 ml of Middlebrook OADC enrichment was added to a lyophilised mixture of antimicrobial agents (PANTA) and 0.8 ml of the reconstituted OADC-PANTA supplement was then added to each individual MGIT. The OADC supplement contained: 0.1 g oleic acid; 50 g bovine albumin; 20 g dextrose; 0.03 g catalase; and 1.1 g polyoxyethylene stearate per l of purified H_2O. The PANTA supplement contained: 6,000 U polymyxin B; 600 µg amphotericin B; 2,400 µg nalidixic acid; 600 µg trimethoprim; and 600 µg azlocillin.

For DNA microarray analysis, Dubos broth was used. For 900 ml of sterile distilled H_2O, this contained approximately: 0.5 g pancreatic digest of casein; 2.0 g L-asparagine; 1.0 g monopotassium phosphate; 2.5 g disodium phosphate; 0.05 g ferric ammonium citrate; 0.01 g magnesium sulphate; 0.2 g polysorbate 80; 0.5 mg calcium chloride; 0.1 mg zinc sulphate; 0.1 mg copper sulphate; and 100 ml bovine albumin.

Enriched Dubos broth contains an enzymatic digest of casein and L-asparagine as nutrient sources. A variety of inorganic salts provide ions required for mycobacterial metabolism. Polysorbate 80, an oleic acid ester, supplies essential fatty acids for mycobacterial replication. Bovine albumin acts as a protective agent by binding free fatty acids that may be toxic to mycobacteria. The albumin is heat-treated to inactivate lipase, which may release fatty acids from the polysorbate 80. Phosphate buffers maintain the pH of the medium. The particular value of Dubos broth is that it provides enhanced dispersed growth, free of excessive clumps.

2.3.3. DNA extraction for PCR

Either a 1 µl loopful of mycobacteria from solid culture was suspended in 0.5 ml sterile distilled H_2O or 0.5 ml of liquid culture was pipetted into a 1.5 ml screw-capped micro centrifuge tube. The mycobacteria were centrifuged for 5 minutes at 8,000 g using an aerosol-containment rotor. The supernatant was discarded and the pellet was resuspended in 300 µl sterile distilled H_2O and vortexed for 5 seconds. The suspension was heated at 100°C for 30 minutes and then sonicated for 15 minutes at room temperature. The samples were then centrifuged for 5 minutes at 13,000 g and the supernatant was transferred to a clean labelled 1.5 ml microcentrifuge tube. DNA extracts were then stored at $^-$20ºC until analysis.

2.4. IDENTIFICATION OF MYCOBACTERIAL ISOLATES.

All positive cultures in the HPA Midlands Region Centre for Mycobacteriology were identified as specific members of the *M. tuberculosis* complex with the GenoType® MTBC test (HAIN Lifescience, Nehren, Germany). This assay distinguishes between: *M. tuberculosis*; *M. africanum*; *M. microti*; *M. bovis* ssp. bovis; *M. bovis* ssp. BCG; and *M. bovis* ssp. *caprae* by PCR amplification and reverse line hybridisation of three target regions (23S rDNA, *gyrB*, and RD1 in BCG) (Richter *et al.*, 2004).

Hybridisation included the following steps: chemical denaturation of the amplification products; hybridisation of the single-stranded, biotin-labelled amplicons to membrane-bound probes; stringent washing; addition of a streptavidin/alkaline phosphatase (AP) conjugate; and an AP mediated staining reaction. Hybridisation can be automated using the GT-Blot 48 System.

2.4.1. HAIN GenoType MTBC: PCR amplification

For each individual sample, the PCR amplification mix (45 µl) was prepared in a DNA-free room. The PCR mastermix contained: 35 µl PNM; 5 µl 10X polymerase incubation buffer; 4 µl $MgCl_2$ solution; 1.5 U HotStarTaq DNA Polymerase (Qiagen, Crawley, West Sussex, UK); and 0.7 µl H_2O. In a separate laboratory, 5 µl of template DNA was added to make a final volume of 50 µl with one negative control containing 5 µl of sterile distilled H_2O. An initial denaturation at 95°C for 5 minutes was followed by 10 cycles of denaturation at 95°C for 30 seconds and annealing at 58°C for 2 minutes. This was then followed by 20 cycles of 95°C for 25 seconds, annealing at 43°C for 40 seconds, and 70°C for 40 seconds. A final extension step at 70°C for 8 minutes concluded the reaction program. The amplified regions have a length of

215 bp (Universal Control), 152 bp and 203 bp (*gyrB*). BCG will have an additional amplicon of 117 bp. The amplicons were then hybridised using the provided reagents in the HAIN MTBC kit on a BeeBlot system (Bee Robotics, Caernarfon, Gwynedd, UK).

2.4.2. HAIN GenoType MTBC: Evaluation and interpretation of results

Hybridised strips were pasted onto a provided evaluation sheet, positive signals were noted, and species were determined using a provided interpretation chart. Each strip has a total of 13 reaction zones including two control zones (conjugate and universal). The conjugate control identifies the efficiency of conjugate binding and the substrate reaction. A line should always be present in the conjugate control zone. The universal control detects all known mycobacteria and members of gram positive bacteria with a high GC content. The MTBC zone hybridises with amplicons generated from all members of the *M. tuberculosis* complex. There are 10 other zones with specific probes for each member of the *M. tuberculosis* complex. Only those bands whose intensities that are at least equivalent to the universal control zone were considered.

2.4.3. Agarose gel electrophoresis

For each agarose gel, the appropriate weight of Agarose-1000 (Invitrogen, Paisley, Renfrewshire, UK) was weighed, added to the appropriate volume of 0.5X tris borate EDTA (TBE), and left to hydrate for 15 minutes. The agarose-TBE mixture was then heated in a microwave until boiling, removed and mixed. If the agarose percentage was greater than 1%, the gel was heated again until boiling. If the agarose percentage was less than 1%, the gel was heated to boiling once again. The agarose was then left to cool for 5 minutes at room temperature. For all gels except those for IS*6110* RFLP Southern Hybridisation, a 12 x 14 cm gel casting tray was used. The agarose was poured into the gel tray, 1 mm width 20 lane gel combs inserted, left to set for 30 minutes at room temperature and then at least 30 minutes (preferably overnight) at 4ºC. For gels that analysed PCR amplicons, 10 µl 5X loading buffer was added to each PCR sample and mixed with a pipette. The gel was then inserted into the gel tank, TBE was added such that there was at least 0.5 cm TBE buffer above the gel, and gel combs were removed. Either 50 bp or 100 bp ladder was added to first, middle, and last wells and 12 µl of each amplicon was added to the corresponding well. The gel tank lid was placed and connected to a power supply unit. Voltage and time was set to the desired values. Gels for PCR amplicon detection were stained in ethidium bromide for 15 minutes and destained in 1X TBE for 15 minutes before an image was taken using an ImageMaster Gel Documentation System (GE Healthcare, Chalfont St Giles, Buckinghamshire, UK).

2.5. IS*6110* RFLP

2.5.1. Isolation of high molecular weight genomic DNA from mycobacteria

All steps were performed in a Class I Biological Safety Cabinet within Containment Level III. One LJ slope was inoculated with the mycobacterial strain of interest and incubated at 37ºC for at least four weeks or until growth was clearly visible. At least two loops of mycobacteria were transferred into a 1.5 ml screw-capped microcentrifuge tube containing 400 µL 1X tris ethylenediaminetetraacetic acid (TE) buffer and the cells were then heated at 80ºC for 20 minutes. To each strain, 50 µL lysozyme (10 mg/ml) was added; tubes were vortexed and incubated at 37°C overnight. The next day, 70 µL 10% sodium dodecyl sulphate (SDS) and 5 µl proteinase K (10 mg/ml) was added, tubes were vortexed, and incubated at 65°C for 30 minutes. After this, 100 µL 5 M NaCl and 100 µL cetyl trimethylammonium bromide (CTAB)/NaCl solution were added, tubes were vortexed, and incubated at 65°C for 10 minutes. To separate DNA from the cell suspension, 750 µL chloroform/isoamyl alcohol (24:1) was added, tubes were inverted 10 times, and centrifuged at 12,000 g for 8 minutes. The aqueous supernatant was transferred to a fresh microcentrifuge tube and 450 µL isopropanol stored at -20ºC was added. Nucleic acids were precipitated by slowly inverting the tube a few times and the tube was incubated at -20°C for at least 30 minutes. DNA was pelleted by centrifugation at 12,000 g for 15 minutes. Most of the supernatant was removed and 1 ml 70% ethanol stored at -20ºC was added and centrifuged at 12,000 g for 5 minutes. Most of the supernatant was again removed and an additional centrifugation of at 12,000 g for 1 minute was done. The rest of the supernatant was removed, dried for 15 minutes at room temperature, and resuspended in 1X TE buffer. If there was no precipitate of nucleic acids visible, the pellet was dissolved in 20 µL 1X TE buffer. If there was a small precipitate visible, then the pellet was dissolved in 35 µL. Medium and large precipitates required 50 µL and 80 µL, respectively. The DNA extract was incubated at 4ºC overnight to redissolve and until needed.

2.5.2. Estimation of the DNA concentration before digestion using a GeneQuant Spectrophotometer

DNA samples were vortexed briefly and 2 µl of each DNA sample was added to 198 µL 1X TE. A GeneQuant™ Spectrophotometer (GE Healthcare, Little Chalfont, Buckinghamshire, UK) was used and the mode was selected for double stranded DNA quantification. A blank cuvette containing 100 µL 1X TE was pipetted into a clean cuvette and a reading was taken as the reference by using the "SET REF" function. For each sample, the absorbance of 100 µl of the diluted solution was analysed at 260 nm and noted.

The concentration of DNA was calculated using the following formula:
DNA concentration (mg/ml) = 50 mg/ml x Measured A_{260} x dilution factor

The amount of DNA required for *Pvu*II digestion (4.5 µg) was calculated with the formula:
4500/DNA concentration (µg/ml) = µL DNA required for digestion.

2.5.3. Analysis of the DNA quality

A 0.8% agarose gel was prepared and 1 µl extracted DNA was added to 9 µL gel loading buffer. All (10 µL) of each sample was loaded into separate wells and 5 µl HyperLadder IV (Bioline, London, UK) was loaded into adjacent lanes. The samples were electrophoresed at 110 V until the dye front had migrated to approximately 5 mm from the end of the gel and then visualised using a Gel Documentation System. The quality of the DNA samples was assessed by the presence of unsheared whole genomic DNA or sheared DNA. If a DNA extract had a significant amount of sheared DNA then the extraction process would be repeated.

2.5.4. Digestion of chromosomal DNA by *Pvu*II

For the digestion of each DNA extract, the following mastermix was used: 2 µl digestion buffer M; 1 µl *Pvu*II restriction enzyme (10 U/µl) (Roche Applied Science, Burgess Hill, West Sussex, UK); volume of DNA for 4.5 µg; and sterile distilled H_2O to make a total volume of 20 µl. The digestion reaction was vortexed briefly and centrifuged at 11,000 g for 5 seconds. Sufficient control DNA (*M. tuberculosis* MT14323) for three reactions was also digested. The digestion reaction was vortexed briefly, centrifuged at 11,000 g for 5 seconds, and incubated at 37°C for 3 hours in a thermal cycler. The enzyme was inactivated by incubation at 65°C for 15 minutes.

2.5.5. Estimation of the DNA concentration after digestion

A small (14 x 20 cm) 0.8% agarose gel was prepared. For 2 µl digested DNA, 18 µl of gel loading buffer was added and 10 µl of each sample was then loaded into individual lanes on the gel. In the first and last lanes, 5 µl Molecular Weight Marker IV (Roche Applied Science, Burgess Hill, West Sussex, UK) was added. The samples were electrophoresed at 150V for 1 hour and visualised using a Gel Documentation System. The intensity of digested DNA in each lane was analysed and the volume of DNA required to get an equal concentration in each lane on the Southern blot gel was estimated.

2.5.6. Agarose gel electrophoresis

A large (28 x 20 cm) 0.8% agarose gel was prepared and 5 µL Molecular Weight Marker IV was loaded in the first and last lanes. In the remaining lanes, the volume of each DNA estimated previously was loaded including MT14323 in the first, middle, and last lanes. The gel was then subjected to electrophoresis at 100 V for 1 hour until the DNA samples had migrated out of the wells and into the gel. The voltage was then reduced to 60 V and electrophoresed for 20 hours. The migration of DNA in the gel was checked on a UV transilluminator and electrophoresed until the 2.0 kb fragment of the DNA standard had migrated at least 7 cm from the wells. The gel was then prepared for Southern blotting as follows.

2.5.7. Gel pre-treatment before Southern blotting

All pre-treatment steps were carried out with constant slow agitation. The gel was depurinated in 250 ml 0.25M hydrochloric acid twice for 15 minutes and then the gel was rinsed in sterile distilled H_2O. The gel was then denatured in 250 ml 0.5M sodium hydroxide twice for 20 minutes, and then 250 ml 1.5 M NaCl twice for 20 minutes and then rinsed in sterile distilled H_2O. The gel was then neutralised in 250 ml 3M NaCl, 0.5M Tris (pH 7.0) twice for 20 minutes each time.

2.5.8. Transfer of DNA onto a membrane by vacuum blotting

Once the gel pre-treatment was complete, a Hybaid Vacu-Aid (Thermo Fisher Scientific, Loughborough, Leicestershire, UK) was assembled to perform the vacuum blot. An aperture was cut in a rubber mask which was 5-10 mm smaller than the gel. This was used to form a seal around the gel when vacuum is applied. The wells in the gel were excluded by cutting them off the gel which ensured that the wells would not rupture and 'short circuit' the DNA transfer. The Vacu-aid was assembled without locating the rubber mask or top manifold and a cross lattice support and porous screen was inserted. A piece of 3MM paper (Whatman, Maidstone, Kent, UK) 2-4 cm larger than the gel to be blotted was cut, wetted with 2X saline sodium citrate (SSC), and positioned onto the porous screen. The Hybond™-N+ (GE Healthcare, Chalfont St Giles, Buckinghamshire, UK) transfer membrane was cut slightly longer than the gel to be blotted and placed on top of the 3MM paper. The rubber mask was carefully placed over the membrane and 3MM paper so that the aperture was exactly over the membrane. The top manifold and clamp were inserted on top of the assembly. The aperture in the rubber mask was flooded with a small quantity of transfer buffer. Ensuring the rubber mask was lying flat; the agarose gel was positioned exactly over the aperture. All air bubbles were excluded from between the gel and the membrane using a gloved finger. Once a satisfactory seal was created, a vacuum was applied and transfer buffer was applied to the top of the gel so that all of the gel was immersed. Blotting was carried out for 90 minutes. When the transfer was completed, the vacuum was turned off and the transfer buffer and gel was removed. The membrane blot was removed and DNA was fixed by UV crosslinking.

2.5.9. Preparation of IS*6110* probe

DNA from *M. tuberculosis* MT14323 was extracted by the CTAB method as described previously. Multiple PCR reactions were prepared using the following PCR mastermix for each reaction: 200 μM each dNTP; 1 μM each oligonucleotide (INS-1 5'-CGT GAG GGC ATC GAG GTG GC-3' and INS-2 5'-GCG TAG GCG TCG GTG ACA AA-3'); 5 μl 10X PCR Buffer and 1.25 U Super Taq DNA polymerase (Sphaero Q, Gorinchem, Netherlands); 5 μl template DNA; and H_2O to a make a total reaction volume of 50 μl. A negative control of 5 µl H_2O instead of template DNA was added to one reaction. An initial denaturation at 96°C for 3 minutes was followed by 30 cycles of denaturation at 96°C for 1 minute, annealing at 65°C for 1 min, and extension at 72°C for 2 minutes. A final extension step of 6 minutes at 74°C concluded the reaction program.

A 1% agarose gel was prepared using tris acetate EDTA (TAE) buffer. The PCR products were pooled, 2 µL gel loading buffer was added to 10 µL of the pooled PCR products and loaded onto the agarose gel. In the first and last lanes, 5 µL 50 bp electrophoresis marker (Promega, Southampton, Hampshire, UK) was loaded and the gel was electrophoresed at 100 V for 1 hour and visualised with the ImageMaster Gel Documentation system. A 245 bp PCR fragment was observed and excised using a clean, sterile scalpel in approximately 300 µl (300 mg) agarose, transferred to a 1.5 ml microcentrifuge tube and incubated at 70°C until the agarose was completely melted. Each melted agarose slice was purified using the Wizard® PCR Preps DNA Purification Kit (Promega, Southampton, Hampshire, UK). The purified 245 bp amplicon was then labelled with fluorescein using the Gene Images Random Prime Labelling Kit (GE Healthcare, Chalfont St Giles, Buckinghamshire, UK) and incorporation of fluorescein was confirmed by comparison with provided standards.

2.5.10. Hybridisation and detection

Hybridisation and detection was carried out using the Enhanced Chemiluminescence (ECL) Direct Nucleic Acid Labelling and Detection Kit (GE Healthcare, Chalfont St Giles, Buckinghamshire, UK). The required volume of hybridisation buffer (0.3 ml/cm^2) was pre-heated to 60°C and the blot was wetted in 5X SSC. The blot was placed in hybridisation buffer and pre-hybridised at 60°C for 30 minutes with constant, gentle rotation in a hybridisation oven (Model 400 Hybridisation Incubator, SciGene, Sunnyvale, California, USA). The required amount of IS*6110* probe (10 ng/ml of hybridisation buffer) was removed from -20ºC to a clean microcentrifuge tube, denatured by heating to 100°C, and snap cooled on ice. The denatured probe was centrifuged briefly and a small aliquot of the hybridisation buffer was removed from the blot, mixed with the probe, and returned to the bulk of the buffer. The blot was then hybridised at 60°C overnight with gentle rotation. Stringency wash solution 1 was prepared (5 ml per cm^2 of membrane) and pre-heated to 60°C. The blot was transferred to this solution and washed at 60°C for 15 minutes with gentle rotation. The blot was then washed with pre-heated stringency wash solution 2 (5 ml per cm^2 of membrane) and washed at 60°C for 15 minutes with gentle rotation

2.5.11. Blocking, antibody incubation and washes

The following steps were all performed at room temperature and all the incubations required constant agitation of the blot. All containers were rinsed with 70 % ethanol before use to remove any bacterial AP contamination. Following the stringency washes, the blot was incubated with gentle agitation for 1 hour in a 1:10 dilution of liquid blocking agent in hybridisation buffer (1.0 ml/cm^2 membrane). The anti-fluorescein-AP conjugate was diluted 1:5000 in freshly prepared 0.5 % (w/v) bovine serum albumin in hybridisation buffer and the blots were incubated in diluted conjugate (0.3 ml/cm^2 of membrane) with gentle agitation for 1 hour. Unbound conjugate was removed by washing three times for 10 minutes each time in 0.3 % (w/v) Tween 20 in hybridisation buffer (5 ml/cm^2 membrane) at room temperature with gentle agitation.

2.5.12. Signal generation and detection

After completion of the antibody incubation stage and subsequent stringency washes, any excess wash buffer was drained from the blots. The blot was placed sample side up on a sheet of SaranWrap on a flat surface. An aliquot of detection reagent (40 µl/cm^2 membrane) was removed to a universal, pipetted on to the blots for 5 minutes and then drained off. The blot was then transferred to a supplied developing bag and the outside was wiped dry and then transferred to a darkroom. The blot was exposed for 5 minutes using one sheet of Hyperfilm-MP (GE Healthcare, Chalfont St Giles, Buckinghamshire, UK) on top of the blot contained in a film cassette. After 5 minutes, the film was removed and placed in developer solution until all expected fragments in MT14323 were observed, rinsed in H_2O, placed in fixer solution until the film was transparent, and then dried overnight.

2.5.13. Cluster analysis

Using a Gel Documentation System, an image of the developed blot was saved a Tagged Image File Format (TIFF) file. This file was imported into BioNumerics (Applied Maths, Sint-Martens-Latem, Belgium) as a 2D TIFF file with 8 bit optical density depth (256 grey scale). Each lane was identified and selected for analysis and a spectral analysis was carried out that removed background images and identified the fragments with lowest intensity that will be analysed. The gel was then normalised using the three external MT14323 reference standard lanes. Fragments containing IS*6110* were then identified automatically with manual confirmation. A similarity dendrogram was constructed by using the Dice coefficient, and results were displayed via the unweighted pair group method using arithmetic averages (UPGMA). Band tolerance levels were set at 1%.

2.6. MIRU-VNTR TYPING USING AGAROSE GEL ELECTROPHORESIS

2.6.1. MIRU-VNTR typing using the originally described oligonucleotides.

VNTR analysis was performed by using the oligonucleotides for the five loci ETR-A to ETR-E as originally described (Frothingham and Meeker-O'Connell, 1998) (Table 2.1). A total PCR volume of 25 µl was used for each reaction which contained: 400 nM each primer; 2.5 µl of PCR Gold Buffer; 1.5 mM $MgCl_2$; 200 µM each of the four dNTPs; 4% (vol/vol) dimethyl sulphoxide; 0.5 U of Amplitaq Gold (Applied Biosystems, Warrington, UK); and 2 ng of DNA. An initial denaturation at 95°C for 7 minutes was followed by 35 cycles of denaturation at 95°C for 30 seconds, annealing at 60°C for 1 minute, and extension at 72°C for 1 minute. A final extension step at 72°C for 5 minutes concluded the reaction program. The amplicons were sized using a 2% (wt/vol) agarose-1000 gel for 2.5 h at 150 V with 50-bp and 100 bp ladder size standards (Table 2.2).

MIRU analysis was performed using the primers for the 12 MIRU loci as originally described (Supply *et al.*, 2001) (Table 2.1). Four different concentrations of $MgCl_2$ were used. $MgCl_2$ was added to the reaction mixture at 1.5 mM for MIRU loci 20, 24, and 27. For MIRU loci 10, 16, and 31, 2 mM $MgCl_2$ was used. For MIRU loci 2, 23, and 39, 2.5 mM $MgCl_2$ was used. For MIRU loci 4, 26, and 40, 3 mM $MgCl_2$ was used. Each PCR mixture contained: 4% (vol/vol) dimethyl sulphoxide; 200 µM each of the four deoxynucleoside triphosphates; 2.5 µl PCR Gold buffer; 400 nM each primer; 0.25 U of AmpliTaq Gold; 2 ng of template DNA; and H_2O to a final volume of 25 µl. An initial denaturation at 95°C for 7 minutes was followed by 40 cycles of denaturation at 95°C for 1 minute, annealing at 59°C for 1 minute, and 72°C for 90 seconds. A final extension step at 72°C for 10 minutes concluded the reaction program. The amplicons were sized using a 2% (wt/vol) agarose-1000 gel for 2.5 h at 150 V with 50 bp and 100 bp ladder size standards (Table 2.2).

No.	VNTR	Locus Name	Alias	GenBank Accession No.	Direction	Oligonucleotide sequence (5’-3’)
1	2165	ETR-A		BX842578	Forward	AAATCGGTCCCATCACCTTCTTAT
					Reverse	CGAAGCCTGGGGTGCCCGCGATT
2	2461	ETR-B		BX842579	Forward	GCGAACACCAGGACAGCATCATG
					Reverse	GGCATGCCGGTGATCGAGTGG
3	577	ETR-C		BX842573	Forward	GTGAGTCGCTGCAGAACCTGCAG
					Reverse	GGCGTCTTGACCTCCACGAGT
4	580	ETR-D		BX842573	Forward	CAGGTCACAACGAGAGGAAGAGC
					Reverse	GCGGATCGGCCAGCGACTCCTC
5	3192	ETR-E		BX842581	Forward	CTTCGGCGTCGAAGAGAGCCTC
					Reverse	CGGAACGCTGGTCACCACCTAAG
6	154	MIRU-02		BX842572	Forward	TGGACTTGCAGCAATGGACCAACT
					Reverse	TACTCGGACGCCGGCTCAAAAT
7	960	MIRU-10		BX842574	Forward	GTTCTTGACCAACTGCAGTCGTCC
					Reverse	GCCACCTTGGTGATCAGCTACCT
8	1644	MIRU-16		BX842576	Forward	TCGGTGATCGGGTCCAGTCCAAGTA
					Reverse	CCCGTCGTGCAGCCCTGGTAC
9	2059	MIRU-20		BX842577	Forward	TCGGAGAGATGCCCTTCGAGTTAG
					Reverse	GGAGACCGCGACCAGGTACTTGTA
10	2531	MIRU-23		BX842579	Forward	CTGTCGATGGCCGCAACAAAACG
					Reverse	AGCTCAACGGGTTCGCCCTTTTGTC
11	2687	MIRU-24		BX842579	Forward	CGACCAAGATGTGCAGGAATACAT
					Reverse	GGGCGAGTTGAGCTCACAGAA
12	2996	MIRU-26		BX842580	Forward	TAGGTCTACCGTCGAAATCTGTGAC
					Reverse	CATAGGCGACCAGGCGAATAG
13	3007	MIRU-27	QUB-5	BX842580	Forward	TCGAAAGCCTCTGCGTGCCAGTAA
					Reverse	GCGATGTGAGCGTGCCACTCAA
14	4348	MIRU-39		BX842584	Forward	CGCATCGACAAACTGGAGCCAAAC
					Reverse	CGGAAACGTCTACGCCCCACACAT
15	802	MIRU-40		BX842574	Forward	GGGTTGCTGGATGACAACGTGT
					Reverse	GGGTGATCTCGGCGAAATCAGATA
16	424	VNTR0424	Mtub04	BX842573	Forward	CTTGGCCGGCATCAAGCGCATTATT
					Reverse	GGCAGCAGAGCCCGGGA
17	1955	VNTR1955	Mtub21	BX842577	Forward	AGATCCCAGTTGTCGTCGTC
					Reverse	CAACATCGCCTGGTTCTGTA
18	2163b	VNTR2163b	QUB-11b	BX842578	Forward	CGTAAGGGGGATGCGGGAAATAGG
					Reverse	CGAAGTGAATGGTGGCAT
19	2347	VNTR2347	Mtub29	BX842578	Forward	GCCAGCCGCCGTGCATAAACCT
					Reverse	AGCCACCCGGTGTGCCTTGTATGAC
20	2401	VNTR2401	Mtub30	BX842578	Forward	CTTGAAGCCCCGGTCTCATCTGT
					Reverse	ACTTGAACCCCCACGCCCATTAGTA
21	3171	VNTR3171	Mtub34	BX842581	Forward	GGTGCGCACCTGCTCCAGATAA
					Reverse	GGCTCTCATTGCTGGAGGGTTGTAC
22	3690	VNTR3690	Mtub39	BX842582	Forward	CGGTGGAGGCGATGAACGTCTTC
					Reverse	TAGAGCGGCACGGGGGAAAGCTTAG
23	4052	VNTR4052		BX842583	Forward	AACGCTCAGCTGTCGGAT
					Reverse	CGGCCGTGCCGGCCAGGTCCTTCCCGAT
24	4156	VNTR4156		BX842583	Forward	TGACCACGGATTGCTCTAGT
					Reverse	GCCGGCGTCCATGTT

Table 2.1. DNA sequences of original published oligonucleotides used in agarose gel electrophoresis.

The oligonucleotides for the 5 ETR loci were first published by Frothingham in 1998, the 10 MIRU loci by Supply in 2001, and the additional nine loci were first described together by Supply in 2006.

Locus	Repeat Length (bp)	H37Rv Allele	Repeat number and expected size (bp)										
			0	1	2	3	4	5	6	7	8	9	10
A	75	3	195	270	345	420	495	570	645	720	795	870	945
B	57	3	121	178	235	292	349	406	463	520	577	634	691
C	58	4	44	102	160	218	276	334	392	450	508	566	624
D	77	4’	56	133	210	287	364	441	518	595	672	749	826
E	53	3	65	118	171	224	277	330	383	436	489	542	595
M02	53	2	402	455	508	561	614	667	720	773	826	879	932
M04	77	3’	133	210	287	364	441	518	595	672	749	826	903
M10	53	3	482	535	588	641	694	747	800	853	906	959	1,012
M16	53	2	565	618	671	724	777	830	883	936	989	1,042	1,095
M20	77	2	437	514	591	668	745	822	899	976	1,053	1,130	1,207
M23	53	6	558	611	664	717	770	823	876	929	982	1,035	1,088
M24	54	1	395	449	503	557	611	665	719	773	827	881	935
M26	51	3	461	512	563	614	665	716	767	818	869	920	971
M27	53	3	498	551	604	657	710	763	816	869	922	975	1,028
M31	53	3	65	118	171	224	277	330	383	436	489	542	595
M39	53	2	540	593	646	699	752	805	858	911	964	1,017	1,070
M40	54	1	354	408	462	516	570	624	678	732	786	840	894
424	51	2	537	588	639	690	741	792	843	894	945	996	1,047
1955	57	2	92	149	206	263	320	377	434	491	548	605	662
2163b	69	5	67	136	205	274	343	412	481	550	619	688	757
2347	57	4	335	392	449	506	563	620	677	734	791	848	905
2401	58	2	247	305	363	421	479	537	595	653	711	769	827
3171	54	3	326	380	434	488	542	596	650	704	758	812	866
3690	58	5	272	330	388	446	504	562	620	678	736	794	852
4052	111	5	163	274	385	496	607	718	829	940	1,051	1,162	1,273
4156	59	2	563	622	681	740	799	858	917	976	1,035	1,094	1,153

Table 2.2. Sizing table for original published oligonucleotides using agarose gel electrophoresis.

MIRU-04 contains an additional invariable 53 bp repeat at the 3’ of the 77-bp VNTR units in nearly all clinical isolates. *M. bovis* BCG, H37Rv, H37Ra, and <1 % of all *M. tuberculosis* strains do not possess this 3’-terminal bp MIRU unit and are designated with a prime symbol (i.e. 310 bp equates to 3’ for H37Rv at MIRU-04. ETR-D and ETR-E are the same loci as MIRU-4 and MIRU-31 respectively. Assigning alleles using ETR-D counts all repeats present whereas MIRU-04 takes account of the 3’ 53 bp repeat but does not count it in clinical strains. ETR-E and MIRU-04 count the same repeats. The alleles for H37Rv are from the published sequenced strain (GenBank Accession Number: AL123456.2).

2.6.2. Extra Nine Typing Loci using originally described oligonucleotides and agarose gel electrophoresis.

To complete the optimal 24 loci set, an additional nine MIRU-VNTR loci were amplified. The universal PCR mastermix for these nine loci contained: 30.80 µl H_2O; 2 µl DMSO (100%); 200 µM each dNTP; 5 µl 10X PCR Buffer; 0.5 µM each primer; 3.0 mM $MgCl_2$; and 1 U AmpliTaq Gold. 1 µl of template DNA was added to each 50 µl mastermix. An initial denaturation at 95°C for 10 minutes was followed by 40 cycles of denaturation at 95°C for 1 minute, annealing at 63°C for 1 minute, and extension at 72°C for 2 minutes. A final extension step of 10 minutes at 72°C concluded the reaction program. 10 µl of each sample was electrophoresed for 2 hours 30 minutes at 150 V on a 2% agarose gel.

2.7. MIRU-VNTR TYPING USING NON-DHPLC

2.7.1. MIRU-VNTR typing using non-dHPLC: Primer design

New primers for all 24 loci were designed with shorter primer flanking regions so that amplicons could be reliably sized on a WAVE® System (Table 2.3). Primers were designed using the computer program Primer3 (Rozen and Skaletsky, 2000). To amplify all DNA templates using the same PCR mastermix reaction conditions, primer length and Tm were designed to be as similar as possible. A common magnesium concentration and annealing temperature was identified using serial titrations.

No.	VNTR	Locus	Alias	GenBank Accession No.	Direction	Oligonucleotide sequence (5'-3')
1	2165	ETR-A		BX842578	Forward	TCGGTCCCATCACCTTCTTA
					Reverse	GGATTGAGGGGATCGTGATT
2	2461	ETR-B		BX842579	Forward	CGAACACCAGGACAGCATC
					Reverse	GGTGATCGAGTGGCTATACG
3	577	ETR-C		BX842573	Forward	CAGGCCTTCGCTCACTTAC
					Reverse	CTTGACCTCCACGAGTGCTA
4	580	ETR-D		BX842573	Forward	GTTGATCGAGGCCTATCACG
					Reverse	CTCGTCCTCCACAATCAACA
5	3192	ETR-E		BX842581	Forward	CCACAGCCTTCTCCATTTTC
					Reverse	ACCACCTAAGGGGACTACGC
6	154	MIRU-02		BX842572	Forward	CAGGTGCCCTATCTGCTGAC
					Reverse	GTGTCCGACCGAGTCATAGG
7	960	MIRU-10		BX842574	Forward	ACCGTCTTATCGGACTGCAC
					Reverse	CCCACGACCGATAATGGAG
8	1644	MIRU-16		BX842576	Forward	CCCGTATCGCTTACATACCC
					Reverse	ACGCCTACGCTGATTCCAC
9	2059	MIRU-20		BX842577	Forward	AGGTGCAAGTGCCGACAT
					Reverse	ACTAACGGTGGCGGGTATG
10	2531	MIRU-23		BX842579	Forward	TCTTCGGTGGTCTCGAGTG
					Reverse	CACCGTCTGACTCATGGTGT
11	2687	MIRU-24		BX842579	Forward	CTGGCCAAGACCGAATGC
					Reverse	GGTGAGGACGAGCTGAGG
12	2996	MIRU-26		BX842580	Forward	CGGATAGGTCTACCGTCGAA
					Reverse	CAACTGCCTCGCGGAATAG
13	3007	MIRU-27	QUB-5	BX842580	Forward	GGTGACCAACGTCAGATTCA
					Reverse	GGAGCGGATCAAGACGTTAC
14	4348	MIRU-39		BX842584	Forward	CCGGTCAACAGACCACTAGA
					Reverse	GTCCGTACTTCCGGTTCAG
15	802	MIRU-40		BX842574	Forward	AAGCGCAAGAGCACCAAG
					Reverse	TCTTTCTCTCACGCTCTCGTC
16	424	VNTR0424	Mtub04	BX842573	Forward	CCTGGTCGTCTGGAAACC
					Reverse	GGCATCCTCAACAACGGTAG
17	1955	VNTR1955	Mtub21	BX842577	Forward	AGATCCCAGTTGTCGTCGTC
					Reverse	GCCAATAGCACAGCACCAG
18	2163b	VNTR2163b	QUB-11b	BX842578	Forward	GTTAATCGTAAGGGGGATGC
					Reverse	AGCGTCGAAGTGAATGGTG
19	2347	VNTR2347	Mtub29	BX842578	Forward	AACCCATGTCAGCCAGGTTA
					Reverse	AGAACGAGCGGAACCACAT
20	2401	VNTR2401	Mtub30	BX842578	Forward	GTCGCCGAGCTGGATTTG
					Reverse	GCAGCTAAGGCTATCGGATT
21	3171	VNTR3171	Mtub34	BX842581	Forward	TGCTCCAGATAAGCCGTCAG
					Reverse	TCACCGATTGGGAGAGGATA
22	3690	VNTR3690	Mtub39	BX842582	Forward	GATCACGATGCGGGTCAC
					Reverse	CACGGGGGAAAGCTTAGAC
23	4052	VNTR4052		BX842583	Forward	CTGGAAAGTCCAGGTTACCG
					Reverse	CTACCGGTCGTTGGTCTAGC
24	4156	VNTR4156		BX842583	Forward	ACATCACCTGGTCGCTACG
					Reverse	GACCAGACCGCCGATCAT

Table 2.3. Oligonucleotide sequences designed for analysis by non-dHPLC on a Transgenomic WAVE® system.

2.7.2. MIRU-VNTR typing using non-dHPLC: Analysis gradient design

The parameters for analysis of a PCR amplicon on a Transgenomic WAVE® System (Glasgow, UK) can be altered by the user to enable identification and sizing of specific sizes of PCR amplicons using a non-denaturing sizing programme type which analyses PCR amplicons at 50ºC. Variation in retention time is caused by differences in the sizes of PCR amplicon fragments and not in variation of the actual DNA sequence amplified as occurs at higher temperatures for mutation detection. The aim of designing a fragment sizing analysis gradient was to be able to size PCR amplicons as quickly as possible with a level of accuracy that was sufficient for differentiation between PCR fragments that differ in size by a minimum of 50 bp. After serial amendment of analysis parameters, the following WAVE® analysis parameters were selected: 0.5 minutes per 100 bp; 20 gradient segments; 20 base pair minimum; 900 base pair maximum; clean duration time of 0.1 minutes; equilibration duration time of 0.1 minutes; fast wash at the end of each injection; and a continuous buffer flow rate of 0.9 ml per minute. This resulted in an analysis time of 7.4 minutes for each single injection (Table 2.4).

Time (minutes)	% Buffer A	% Buffer B
0.0	71.0	29.0
0.5	66.0	34.0
0.7	54.8	45.2
1.0	48.9	51.1
1.2	45.2	54.8
1.4	42.7	57.3
1.7	40.9	59.1
1.9	39.5	60.5
2.1	38.4	61.6
2.3	37.5	62.5
2.6	36.8	63.2
2.8	36.2	63.8
3.0	35.7	64.3
3.3	35.2	64.8
3.5	34.9	65.1
3.7	34.5	65.5
4.0	34.2	65.8
4.2	34.0	66.0
4.4	33.7	66.3
4.6	33.5	66.5
4.9	33.3	66.7
5.0	0.0	100.0
5.5	0.0	100.0
5.6	71.0	29.0
6.5	71.0	29.0

Table 2.4. non-dHPLC analysis conditions.

Buffer A acts as the binding agent between the cartridge and DNA with Buffer B eluting DNA from the separation cartridge dependent on the length of the DNA sequence.

2.7.3. MIRU-VNTR typing using non-dHPLC: Creation of customised standard

To improve the accuracy of MIRU-VNTR fragment sizing when using a Transgenomic WAVE® System, an allele ladder was constructed using MIRU-23. *M. tuberculosis* strains that possessed 1-10 repeats at MIRU-23 were selected and amplified using the same PCR conditions as for MIRU-VNTR typing. The PCR amplicons from the 10 strains were then quantified by analysis of 5 µl on the WAVE® System. To obtain an allele ladder with 10 fragments of equivalent intensity and height, the quantity of each amplicon was standardised by comparison with the other nine amplified alleles. Amplicons with higher intensity required less volume. The 10 PCR amplicons were then combined in a single tube and 5 µl was injected before and after each MIRU-VNTR locus.

2.7.4. MIRU-VNTR Typing using non-dHPLC: MIRU-VNTR PCR set-up for nine and 24 loci.

In a PCR clean room separate from areas containing amplified DNA, PCR reactions for each of the MIRU-VNTR analysis loci was set up using the oligonucleotides for each of the 15 or 24 loci listed in Table 2.3 in a single reaction well and not in a multiplex reaction. A total PCR volume of 50 µl was used for each reaction. Each PCR reaction contained: 400 nM each primer; 5 µl 10X PCR Gold Reaction Buffer; 1.5 mM $MgCl_2$; 200 µM each dNTPs; 4% (vol/vol) dimethyl sulphoxide; 1 U Amplitaq Gold; and 2 µl template DNA. An initial denaturation 95°C for 5 minutes was followed by 40 cycles of denaturation at 95°C for 1 minute, annealing at 63°C for 1 minute, and extension at 72°C for 2 minutes. A final extension step at 72°C for 5 minutes concluded the reaction program. One sample of *M. tuberculosis* H37Rv was included in each batch. For 15 or nine loci analysis, batch sizes were 16 DNA extracts which included: 14 clinical DNA extracts; one H37Rv extract; and one negative containing sterile distilled H_20. For 24 loci analysis, batch sizes were eight DNA extracts which included: six clinical DNA extracts; one H37Rv extract; and one negative containing sterile distilled H_20. Five microlitres of each amplified PCR product were injected onto the WAVE® System

2.7.5. MIRU-VNTR typing using non-dHPLC: Calculation of a standard curve

All single peaks on the WAVE® chromatogram between 2.5 and 6.2 minutes and more than 1 mV in height were selected for repeat allele calculation using a Microsoft Excel file. Retention times for each of the standard peaks were entered into the Excel file which then calculates a standard curve. The retention times for each of the sample PCR amplicon peaks were then imported and compared against the standard curve to obtain a predicted size and nearest whole number of repeats based on the following equation (Table 2.5):

$$\text{Number of repeats} = \frac{\text{PCR amplicon size (bp) - VNTR primer flanking region (bp)}}{\text{Expected size of one VNTR repeat (bp)}}$$

Any MIRU-VNTR alleles that were not amplified or were greater than +/- 0.25 repeats from the nearest whole repeat were repeated. Each MIRU-VNTR allele was concatenated together to obtain a five (ETR), 10 (MIRU), 17 (ETR+MIRU), or 24 digit profile and transferred to a Microsoft Access database for reporting and analysis via a linked BioNumerics database and the local Laboratory Information Management System.

Locus Name	Repeat Length (bp)	H37Rv Allele	Repeat number and expected size (bp)										
			0	1	2	3	4	5	6	7	8	9	10
A	75	3	117	192	267	342	417	492	567	642	717	792	867
B	57	3	112	169	226	283	340	397	454	511	568	625	682
C	58	4	39	97	155	213	271	329	387	445	503	561	619
D	77	4’	102	179	256	333	410	487	564	641	718	795	872
E	53	3	77	130	183	236	289	342	395	448	501	554	607
M02	53	2	121	174	227	280	333	386	439	492	545	598	651
M10	53	3	65	118	171	224	277	330	383	436	489	542	595
M16	53	2	85	138	191	244	297	350	403	456	509	562	615
M20	77	2	59	136	213	290	367	444	521	598	675	752	829
M23	53	6	74	127	180	233	286	339	392	445	498	551	604
M24	54	1	131	185	239	293	347	401	455	509	563	617	671
M26	51	3	64	115	166	217	268	319	370	421	472	523	574
M27	53	3	108	161	214	267	320	373	426	479	532	585	638
M39	53	2	105	158	211	264	317	370	423	476	529	582	635
M40	54	1	110	164	218	272	326	380	434	488	542	596	650
424	51	2	80	131	182	233	284	335	386	437	488	539	590
1955	57	2	48	105	162	219	276	333	390	447	504	561	618
2163b	69	5	78	147	216	285	354	423	492	561	630	699	768
2347	57	2	107	164	221	278	335	392	449	506	563	620	677
2401	58	2	64	122	180	238	296	354	412	470	528	586	644
3171	54	3	86	140	194	248	302	356	410	464	518	572	626
3690	58	5	110	168	226	284	342	400	458	516	574	632	690
4052	111	5	90	201	312	423	534	645	756	867	978	1,089	1,200
4156	59	2	213	272	331	390	449	508	567	626	685	744	803

Table 2.5. Allele calling table for the 24 MIRU-VNTR loci using oligonucleotides designed for analysis using non-dHPLC.

2.8. SPOLIGOTYPING USING A SUSPENSION ARRAY SYSTEM

2.8.1. Oligonucleotide-microsphere coupling protocol

Spoligotyping was carried out using a BioPlex 200 Suspension Array System (Bio-Rad, Hemel Hempstead, Hertfordshire, UK) as previously described (Cowan *et al.*, 2004), using oligonucleotides previously designed for use in membrane-based hybridisation (Table 2.6) (Kamerbeek *et al.*, 1997). During the coupling procedure, exposure of the microspheres to light was minimised as much as possible. A fresh aliquot of EDC (1-ethyl-3-[3-dimethylaminopropyl]carbodiimide hydrochloride) powder (Thermo Fisher Scientific, Loughborough, Leicestershire, UK) was warmed to room temperature. Each of the 43 amino-linked spacer oligonucleotides with a six carbon spacer were resuspended to a concentration of 0.1 mM in 0.1M morpholineethanesulfonic acid (MES) (pH 4.5). All of the 43 microsphere sets were pelleted by centrifugation at 8,000 g for 1 minute, sonicated for 15 seconds, and vortexed for 15 seconds. From each of the 43 stock solutions, 200 µl of microspheres were transferred to a 1.5 ml microtube and pelleted by centrifugation at 8,000 g for 1 minute. For each of the 43 bead sets, beads were then resuspended in 25 µl 0.1 M MES (pH 4.5) and 2 µl of the appropriate 0.1 mM amino linked oligonucleotide was added. To each of the bead sets, 2.5 µl EDC (10 mg/ml) was added and sonicated for 15 seconds, vortexed gently for 15 seconds, and then incubated for 30 minutes at room temperature protected from light. The addition of EDC was then repeated once. Each of the 43 bead sets were then washed twice. For the first wash, 0.9 ml 0.02% Tween-20 was added to the beads, vortexed for a few seconds, and then centrifuged at 8,000 g for 1 minute. The supernatant was then removed without disturbing the pellet. For the second wash, 0.9 ml 0.1% SDS was added to each bead set and vortexed for a few seconds. The beads were centrifuged at 8,000 g for 1 minute and the supernatant was removed. Each bead set was then resuspended in 50 µl 1X TE, pH 8.0

and pelleted by centrifugation at 8,000 g for 1 minute. The beads were then stored at 4ºC in the dark.

Name	Bio-Rad Cat. No.	Oligonucleotide sequence (5’-3’)
sp01-011	171-506011	ATAGAGGGTCGCCGGTTCTGGATCA
sp02-017	171-506017	CCTCATAATTGGGCGACAGCTTTTG
sp03-018	171-506018	CCGTGCTTCCAGTGATCGCCTTCTA
sp04-019	171-506019	ACGTCATACGCCGACCAATCATCAG
sp05-020	171-506020	TTTTCTGACCACTTGTGCGGGATTA
sp06-021	171-506021	CGTCGTCATTTCCGGCTTCAATTTC
sp07-024	171-506024	GAGGAGAGCGAGTACTCGGGGCTGC
sp08-025	171-506025	CGTGAAACCGCCCCCAGCCTCGCCG
sp09-026	171-506026	ACTCGGAATCCCATGTGCTGACAGC
sp10-027	171-506027	TCGACACCCGCTCTAGTTGACTTCC
sp11-028	171-506028	GTGAGCAACGGCGGCGGCAACCTGG
sp12-029	171-506029	ATATCTGCTGCCCGCCCGGGGAGAT
sp13-032	171-506032	GACCATCATTGCCATTCCCTCTCCC
sp14-033	171-506033	GGTGTGATGCGGATGGTCGGCTCGG
sp15-034	171-506034	CTTGAATAACGCGCAGTGAATTTCG
sp16-035	171-506035	CGAGTTCCCGTCAGCGTCGTAAATC
sp17-036	171-506036	GCGCCGGCCCGCGCGGATGACTCCG
sp18-037	171-506037	CATGGACCCGGGCGAGCTGCAGATG
sp19-038	171-506038	TAACTGGCTTGGCGCTGATCCTGGT
sp20-039	171-506039	TTGACCTCGCCAGGAGAGAAGATCA
sp21-040	171-506040	TCGATGTCGATGTCCCAATCGTCGA
sp22-041	171-506041	ACCGCAGACGGCACGATTGAGACAA
sp23-042	171-506042	AGCATCGCTGATGCGGTCCAGCTCG
sp24-046	171-506046	CCGCCTGCTGGGTGAGACGTGCTCG
sp25-047	171-506047	GATCAGCGACCACCGCACCCTGTCA
sp26-051	171-506051	CTTCAGCACCACCATCATCCGGCGC
sp27-052	171-506052	GGATTCGTGATCTCTTCCCGCGGAT
sp28-053	171-506053	TGCCCCGGCGTTTAGCGATCACAAC
sp29-054	171-506054	AAATACAGGCTCCACGACACGACCA
sp30-055	171-506055	GGTTGCCCCGCGCCCTTTTCCAGCC
sp31-056	171-506056	TCAGACAGGTTCGCGTCGATCAAGT
sp32-061	171-506061	GACCAAATAGGTATCGGCGTGTTCA
sp33-062	171-506062	GACATGACGGCGGTGCCGCACTTGA
sp34-063	171-506063	AAGTCACCTCGCCCACACCGTCGAA
sp35-064	171-506064	TCCGTACGCTCGAAACGCTTCCAAC
sp36-065	171-506065	CGAAATCCAGCACCACATCCGCAGC
sp37-066	171-506066	CGCGAACTCGTCCACAGTCCCCCTT
sp38-072	171-506072	CGTGGATGGCGGATGCGTTGTGCGC
sp39-073	171-506073	GACGATGGCCAGTAAATCGGCGTGG
sp40-074	171-506074	CGCCATCTGTGCCTCATACAGGTCC
sp41-075	171-506075	GGAGCTTTCCGGCTTCTATCAGGTA
sp42-076	171-506076	ATGGTGGGACATGGACGAGCGCGAC
sp43-077	171-506077	CGCAGAATCGCACCGGGTGCGGGAG

Table 2.6. Sequences of the oligonucleotides used for spoligotyping.

2.8.2. PCR amplification of the DR locus

For 28 isolates and 4 controls, the following PCR mastermix was prepared for a total volume of 25 µl per reaction: 75 µl 10X PCR Buffer; 15 µl dNTP mix; 60 µl Dra primer (5’-GGTTTTGGGTCTGACGAC-3’, 5’ biotinylated) (5 pmoles/µl); 60 µl DRb primer (5’-CCGAGAGGGGACGGAAAC-3’) (5 pmoles/µl); 462 µl H_2O; 3µl Super Tth DNA polymerase (Sphaero Q, Gorinchem, Netherlands); and 2.5 µl template DNA. *M. tuberculosis* H37Rv and *M. bovis* were used as controls with sterile distilled H_2O added to the final two reactions. The PCR reactions were transferred to a thermal cycler and the following amplification programme was carried out: 3 minutes at 94°C; 20 cycles of 1 minute at 94°C, 1 minute at 55°C, 30 seconds at 72°C; with a terminal extension step of 5 minutes at 72 °C.

2.8.3. Hybridisation and detection

The coupled spoligotyping microspheres solutions were resuspended by sonication for 10 seconds and vortexing 10 times to ensure complete dispersion of the microsphere pellet. The working microsphere solution was prepared (990 µl 1.5X tetramethylammonium chloride (TMAC) hybridisation buffer and 128 µl bead mix) and gently vortexed for 15 second. In a 96-well hybridisation plate, 33 µl of the working microsphere solution and 17 µl of each PCR reaction were added to 32 wells. The plate was sealed with plastic film. The beads and PCR amplicons were hybridised at 94ºC for 10 minutes followed by 52ºC for 30 minutes with continued heating at 52ºC until required. The hybridisation reaction plate was centrifuged at 2,000 rpm for 3 minutes and the supernatant was then removed without disturbing the pellet. The detection reagent was then prepared (10 µl streptavidin-phycoerythrin conjugate (Invitrogen, Paisley, Renfrewshire, UK) and 2.5 ml 1X TMAC hybridisation buffer pre-warmed at 52°C) and thoroughly vortexed. To each hybridisation reaction, 75 µl of detection reagent was added and beads were resuspended by pipetting. The hybridisation reactions were then incubated for 5 minutes at 52°C.

2.8.4. Microsphere Suspension Array analysis on the BioPlex and data analysis

The hybridisation plate was loaded and the protocol settings were set so that 100 beads per region were analysed and the sample size was 50 µl. Data obtained was then exported to Microsoft Excel. Using an Excel file available from the CDC, the median rfu values for each bead set were analysed and used to construct the spoligotype for each strain.

2.9. DATA AND STATISTICAL ANALYSIS

2.9.1. MIRU-VNTR *M. tuberculosis* Strain clustering analysis.

MIRU-VNTR data were entered into a Microsoft Access database linked to BioNumerics. Individual Locus data was imported as character data and clusters were identified by using the Categorical Co-efficient and the UPGMA method for displaying the similarity dendrogram. Two or more strains with indistinguishable alleles at all loci were considered a cluster.

2.9.2. Data Analysis: Assigning Cultural, Ethnic, and Linguistic groups.

Experian Origins software v6.3.0.7 (Nottingham, UK) was used to assign (Cultural, Ethnic, and Linguistic) CEL groups (Webber, 2007). Data was imported as a Comma Separate Variable file, strings such as Jr, Mr, Mrs were removed and the "England, AAA" data processing type was selected. All specific CEL groups identified were accepted and used.

2.9.3. n-1 clustering method

Within a cluster, it can be assumed that the earliest case is the source case and subsequent patients were recently infected from contact with the source case and not reactivation of latent infection (Small *et al.*, 1994;Godfrey-Faussett *et al.*, 2000). A minimum estimate of the proportion of tuberculosis caused by transmission within each cluster can therefore be calculated as:

$$\frac{\text{Number of clustered patients - number of clusters}}{\text{Total number of patients}}$$

2.9.4. Hunter-Gaston Discrimination Index

The discrimination and diversity of typing systems and mycobacterial populations was calculated using the Hunter-Gaston Discrimination Index (HGDI) (Hunter and Gaston, 1988):

$$D = 1 - \frac{1}{N(N-1)} \sum_{j=1}^{s} x_j (x_j - 1)$$

Where *D* is the index of discriminatory power (0.000-1.000), *N* the number of unrelated strains tested, *s* the number of different types, and *xj* the number of strains belonging to the *j*th type. A *D* value of 1.000 indicates that a typing method was able to distinguish every member of a strain population from all other members of that population. Conversely, an index of 0.000 would indicate that all members of a strain population were of an indistinguishable type. An index of 0.500 would mean that if one strain was chosen at random from a strain population, then there would be a 50% probability that the next strain chosen at random would be indistinguishable from the first (Bikandi *et al.*, 2004). The HGDI value was calculated using the HPA VNTR Diversity and Confidence Extractor online tool (Platt, 2011).

2.9.5. Available online databases of *M. tuberculosis* Complex strains

2.9.6. SpolDB4

SpolDB4 is the fourth version of an international global spoligotyping database that contains at least 39,925 strains from 122 countries classified into 62 clades or lineages (Brudey *et al.*, 2006). A list of all known shared types is available as supplementary data with the publication by Brudey *et al* in 2006 or the dataset can be queried online at http://www.pasteur-guadeloupe.fr:8081/SITVITDemo/index.jsp.

2.9.7. MIRU-VNTR*plus*

MIRU-VNTR*plus* is an online database that contains genotypic data from five distinct methods (MIRU-VNTR, Spoligotyping, LSPs, SNPs, and IS*6110* RFLP) for 186 isolates that represent the major branches of the principal global *M. tuberculosis* complex lineages as defined by LSP and spoligotyping analysis. This database can be queried using data generated for any or all of the 5 genotyping methods and a strain lineage can be subsequently assigned. The MIRU-VNTR*plus* database (http://www.miru-vntrplus.org) was developed by D. Harmsen, S. Niemann, P. Supply and T. Weniger (Allix-Beguec *et al.*, 2008b).

2.9.8. Epidemiological data collection.

Information from contact tracing and screening was collated by the Consultant in Communicable Disease Control (CCDC) responsible for the investigation of each described cluster. This information was then considered together with the results of molecular typing by the microbiologists in order to confirm or refute the clinical hypothesis of source and secondary cases, together with linkage.

2.9.9. Epidemiological data analysis

Pearson’s chi-squared test with Fisher's exact test was used where necessary. Univariate and multivariate logistic regression modelling was used to test the significance of odds ratios in Stata v10 (Stata, College Station, Texas, USA). The multivariable model was assembled by adding covariates individually in decreasing order of significance and the "goodness of fit" of each model was assessed using the likelihood ratio test. All cases with missing values for the variables examined were excluded from multivariate models. A univariate analysis of the epidemiological investigation of patients resident in Wolverhampton was undertaken using EpiData Analysis v2.2 (EpiData Association, Odense, Denmark). The extent of any association was expressed as an odds ratio (OR) with 95% confidence intervals.

2.10. DNA MICROARRAYS

Whole genome DNA microarrays of *M. tuberculosis* were kindly provided by the BμG@S group at St. George's Hospital Medical School, London. They were constructed by spotting PCR amplicons from partial sequences of the 3,924 predicted ORFs of the sequenced strain *M. tuberculosis* H37Rv onto poly-L-lysine-coated glass microscope slides. Two hybridisations were done for each strain using a different dye (Cy3 or Cy5) each time.

2.10.1. Preparation of genomic DNA from mycobacteria

M. tuberculosis strains were grown in 100 ml Dubos medium for at least four weeks at 37ºC. The day before cell extraction was started; 10 ml 2 M glycine was added. Mycobacterial cells were harvested by centrifugation for 20 minutes at 7,000 g and washed in 30 ml SET solution (0.3 M sucrose, 50 mM Tris-Cl (pH 8.0), 10 mM EDTA). Cells were resuspended in 2 ml SET solution also containing 2 mg/ml lysozyme and 2 mg/ml lipase and incubated for 60 minutes at 37ºC. To the resuspended cells, 8 ml GSE Solution (6 M guanidinium chloride, 1% sarkosyl, and 20 mM EDTA) was added and incubation was continued at 37ºC for a further 2 hours. To separate DNA, 10 ml chloroform:isoamyl alcohol (24:1) was added, inverted, and centrifuged at 7,000 g for 20 min. The supernatant was removed, 2.5 vol ethanol was added and incubated overnight at -20 °C. The precipitate was centrifuged at 7,000 g for 20 minutes and the ethanol was removed. The pellet was resuspended in 2 ml TE (containing 0.5mg/ml of RNase, 0.5mg/ml of Proteinase K, and 0.5% SDS) and incubated at 37ºC overnight. The chloroform and ethanol steps were repeated using 2 ml and 2.5 vol respectively and the DNA was finally resuspended in 100 μl TE buffer (Davis *et al.*, 1991).

2.10.2. DNA Microarray analysis: Preparation of Cy3/Cy5 labelled DNA

One Cy3 and one Cy5 labelled DNA sample was prepared per microarray slide using the following mastermix (GE Healthcare, Chalfont St Giles, Buckinghamshire, UK): 2-5 µg DNA; 1 µl random primers (3 µg/µl); and H_2O to a total volume of 41.5 µl. The mastermix was heated at 95ºC for 5 minutes, snap cooled on ice, and briefly centrifuged. To each mastermix was added: 5 µl 10X React2 Buffer; 1 µl dNTPs (5 mM dA/G/TTP, 2 mM dCTP); 1.5 µl Cy3 or Cy5 dCTP; and 1 µl Klenow enzyme (3-9U/µl). This mastermix was then incubated in the dark for 90 min.

2.10.3. Slide prehybridisation

The prehybridisation solution for a maximum of four slides was prepared as follows: 8.75 ml 20X SSC; 250 µl 20% SDS; 5 ml BSA (100 mg/ml); and H_2O to a total volume of 50 ml. This solution was pre-warmed to 65ºC in a Coplin staining jar. Up to four slides in one Coplin jar were incubated at 65ºC for 20 minutes, rinsed in 400 ml sterile distilled H_2O for 1 minute, rinsed in 400 ml isopropanol for 1 minute, and each slide was centrifuged at 1,500 rpm for 5 minutes and then stored in a dark, dust-free box until hybridisation.

2.10.4. Wash preparation

Wash A was prepared as follows: 20 ml 20X SSC; 1 ml 20% SDS; and sterile distilled H_2O to a total volume of 400 ml. The wash buffer was pre-warmed to 65ºC.

2.10.5. Slide hybridisation with Cy3/Cy5 labelled DNA

Cy3 and Cy5 labelled DNA samples were purified using the MinElute Purification kit (Qiagen, Crawley, West Sussex, UK). The purified Cy3/Cy5 labelled DNA was mixed with hybridisation solution for one 22 x 22 mm LifterSlip slide (Thermo Fisher Scientific, Loughborough, Leicestershire, UK) as follows: 14.9 µl labelled DNA sample; 4.6 µl 20X SSC; and 3.5 µl 2% SDS. The prehybridised microarray slide was placed in the hybridisation cassette and two 15 µl aliquots H_2O was added to the wells in the cassette. The labelled DNA and hybridisation solutions was heated at 95ºC for 2 minutes and allowed to cool slightly. A LifterSlip was placed carefully over the array area on the slide and the hybridisation solution was pipette under one corner of the LifterSlip allowing capillary action to draw the solution completely across the array. The hybridisation cassette was sealed and submerged in a waterbath at 65ºC in the dark for 16-20 hours.

2.10.6. Slide washing

Wash A buffer pre-heated to 65ºC was added to a pre-heated staining trough at 65ºC. The microarray slide was removed from the hybridisation cassette and washed in Buffer A to remove the LifterSlip. After this, the slide was then washed in Buffer A for 2 minutes. Wash B buffer (1.2 ml 20X SSC and 398.8 ml H_2O) was prepared. The slide was then washed in Buffer B for 2 minutes and then transferred to fresh Buffer B for a second wash of 2 minutes. The slide was then centrifuged at 1,500 rpm for 5 minutes to dry. The slide was then scanned using a GenePix 4000A scanner (Molecular Devices, Wokingham, Berkshire, UK), analysed by GenePix v.3.0 software (Molecular Devices, Wokingham, Berkshire, UK) and normalized using GeneSpring v.6 (Silicon Genetics, Stockport, Cheshire, UK). The following normalizations were undertaken: after dye swap of the slides, the data were normalized per spot dividing by control channel using a cut off 0.01 and per chip normalizing to the 50th percentile. Genes were only considered to be deleted when the p-value was <0.05.

2.10.7. Confirmation of deleted regions by PCR and sequencing.

PCR and DNA sequencing was used to confirm deletions which included: Rv1519; Rv3516-17; Rv3738c-39c; Rv1917c; plcC-cut1; Rv3017c-Rv3022c; and Rv3135. Primers were designed using Primer 3 with an annealing temperature between 58 and 60ºC (Rozen and Skaletsky, 2000). AmpliTaq Gold polymerase or the Expand Long Template PCR System (Roche Applied Science, Burgess Hill, West Sussex, UK) was used depending on the genomic region and length of predicted amplicon. When AmpliTaq Gold was used, 1 mM of primers and 1.5 mM of $MgCl_2$ was used. The following thermal cycler conditions were used: an initial denaturation of 95°C for 1 minute and then 35 cycles of 95ºC for 1 minute, 58ºC for 1 minute, and 72ºC for 1 minute, followed by a terminal extension step at 72ºC for 10 minutes. When the Expand Long Template PCR System was used, the PCR conditions were carried out using 1 mM of primers and 2 mM of $MgCl_2$. The PCR cycling programme was as follows: 94ºC for 2 minutes; 10 cycles of 95ºC for 30 seconds, 58ºC for 30 seconds and 68ºC for 5-8 minutes (2 minutes for up to 3 kb, 4 minutes for 6 kb, 8 minutes for 10 kb) followed by 15 cycles of 95ºC for 30 seconds, 58ºC for 30 seconds and 68ºC for 5-8 minutes (2 minutes for up to 3 kb, 4 minutes for 6 kb, 8 minutes for 10 kb) plus 20 seconds cycle elongation for each successive cycle. A terminal extension step at 68ºC for 7 minutes completed the amplification. Sequencing reactions were performed with BigDye Terminator v 1.1 (Applied Biosystems, Warrington, UK) and purified with the DyeEx 2.0 Spin Kit (Qiagen, Crawley, West Sussex, UK). The reactions were purified and analysed by capillary sequencing using a MegaBACE DNA Analysis System (GE Healthcare, Chalfont St Giles, Buckinghamshire, UK). The sequences obtained were compared to TubercuList (http://genolist.pasteur.fr/TubercuList/index.html) and NCBI (http://www.ncbi.nlm.nih.gov).

2.11. ETHICAL CONSIDERATIONS

This thesis details the development of DNA fingerprinting and analysis of subsequent data generated which has been undertaken as part of normal public health practice by microbiologists, respiratory physicians, and public health teams in the West and East Midlands. Therefore, specific ethical approval was not required. The HPA has Patient Information Advisory Group permission under the Health and Social Care Act 2001 to collect and analyse such data for public health purposes.

3. A COMPARISON OF MIRU-VNTR AND IS*6110* RFLP MOLECULAR TYPING METHODS AND RELATIONSHIP TO CONVENTIONAL EPIDEMIOLOGICAL DATA.

3.1. INTRODUCTION

3.1.1. IS*6110* RFLP is the historical gold standard for DNA fingerprinting of *M. tuberculosis*

Initial studies of the IS*6110* element showed that isolates from epidemiologically linked patients generally possessed identical IS*6110* RFLP patterns which equates to a transmission cluster. One of the initial studies that examined the association between epidemiological links and linkage of strains by IS*6110* RFLP showed that two epidemiologically linked clusters of nine and 23 isolates were also clustered by IS*6110* RFLP and that six epidemiologically distinct strains were distinct by RFLP (Hermans *et al.*, 1990). A case-control study of 18 AIDS patients with a drug resistant TB clone and fully sensitive control patients with different distinct strains in a New York Hospital showed that patients with drug-resistant TB had a significantly higher risk of hospitalisation and admittance to the same ward as an infectious patient with drug resistant TB than patients with drug sensitive TB (Edlin *et al.*, 1992). One of the first reports of the use of RFLP in a public health investigation in the UK was in the Midlands. The isolates from one patient who had died of tuberculous meningitis and their neighbour were analysed by IS*6110* RFLP to determine if transmission had occurred. Ten isolates from the same city isolated three months before the death of the patient were also typed. The patient with TB meningitis and neighbour had identical fingerprints whereas the 10 control strains were all distinct. The authors concluded that this data was consistent with, but did not absolutely prove, transmission from the neighbour to the deceased patient (Godfrey-Faussett *et al.*, 1992).

Since these early studies, IS*6110* RFLP has been applied to many different epidemiological which have included regional epidemiological studies of drug-resistant strains, point source outbreaks, and laboratory cross-contamination.

DNA fingerprinting of 253 strains in New York City showed transmission of a drug-resistant strain within New York but also spread of this clone to four other US cities (Bifani *et al.*, 1996). RFLP typing was used to investigate nosocomial transmission of MDR-TB in a HIV positive cohort. Isolates from 6/8 patients were identical by IS*6110* RFLP. Patients who had been hospitalised for more than days within three rooms of an infectious MDR-TB patient had a significantly higher risk of acquiring the MDR-TB strain which highlighted the lack of adequate infection control procedures (Coronado *et al.*, 1993).

Two clusters highlighted the utility of DNA fingerprinting in identifying links that were not originally apparent when conventional epidemiological data was analysed. An epidemiological study of 48 individuals infected with the same strain in Houston showed that individuals who had visited multiple public bars in the same neighbourhood each night had few social links. Each individual was linked to no more than one or two other individuals infected with the same strain. This indicated that there was a complex transmission scenario where limited contact within a large cohort was enough to generate widespread transmission (Yaganehdoost *et al.*, 1999). A retrospective epidemiological study in Berne, Switzerland identified an RFLP cluster that encompassed two apparently discrete social groups. The first group contained 22 patients who were all known to be drug addicts with the second apparently separate group centred on a restaurant. Further investigation revealed that a regular patron of the restaurant had a cough with bloody expectoration for six months before diagnosis with active tuberculosis. This patient was in close contact with a second patient, who was living in a facility for drug addicts where six other people had tuberculosis (Genewein *et al.*, 1993).

As part of a regional epidemiological study in San Francisco, two clusters of six patients each were caused by laboratory cross-contamination. Only analysis of DNA fingerprinting and laboratory records identified these patients and the authors estimated that the cost of this series of false-positive cultures was approximately $15,000 (Small *et al.*, 1993).

3.1.2. Shortcomings of IS*6110* RFLP Typing

Although IS*6110* RFLP had the best correlation with epidemiological data of any typing method developed before 2000, this method does have several shortcomings. Technically, IS*6110* RFLP is labour-intensive and the data obtained requires intensive normalisation using software to overcome interpretability and inter-laboratory transfer issues (Collyns *et al.*, 2002;Kremer *et al.*, 1999;van Soolingen, 2001). The timescale for obtaining IS*6110* RFLP data is at least four weeks from first isolation of a positive culture which limits the opportunity for early and effective public health interventions to prevent further transmission. IS*6110* RFLP cannot be used to assign a genotype to all *M. tuberculosis* strains as not all strains possess IS*6110* elements and IS*6110* RFLP exhibits an artificially increased level of clustering in strains which harbour less than five copies of IS*6110* (van Soolingen *et al.*, 1993). There are specific regions around the world where strains that have low copy number strains are prevalent (Fomukong *et al.*, 1994). In Southern India, it was found that 50/80 (63%) of typed strains did not possess IS*6110* or only had one copy which essentially meant that the majority of strains in this region could not be adequately differentiated by IS*6110* RFLP (Radhakrishnan *et al.*, 2001).

3.1.3. Other alternative methods to IS*6110* RFLP

Molecular solutions for the problems of IS*6110* RFLP have focused on PCR-based methodologies for characterizing IS*6110* polymorphism, such as mixed linker typing or fluorescent amplified fragment length polymorphism (fAFLP) (Haas *et al.*, 1993;Thorne *et al.*, 2007a). Although these methods are attractive, the reproducibility of mixed-linker PCR is sometimes a problem, and the fundamental problems of zero and low copy number isolates are still evident in fAFLP. Multilocus sequence typing (MLST) has been used to produce a discriminatory, digital, PCR-based typing method in a range of bacteria, including methicillin-resistant *S. aureus* and *Neisseria meningitidis* (Maiden *et al.*, 1998). Initial studies of a limited number of genes in *M. tuberculosis* identified few polymorphisms, which limited the utility of this method (Sreevatsan *et al.*, 1997). Subsequent studies analysed an increasing number of target sequences, which in turn has substantially increased the range of diversity identified and improved our understanding of the global phylogeny of *M. tuberculosis* clades (Baker *et al.*, 2004;dos Vultos *et al.*, 2008). However, MLST has not replaced IS*6110* RFLP as a DNA fingerprinting method for public health investigations. Spoligotyping is useful as a second-line alternative when required but does not generally provide sufficient discrimination as a single method for outbreak investigation (Kremer *et al.*, 1999).

3.1.4. MIRU-VNTR typing is a possible alternative to IS*6110* RFLP

MIRU-VNTR typing is being used in an increasing number of studies to overcome the problems of IS*6110* RFLP typing. MIRU-VNTR typing is faster than IS*6110* RFLP as it is a

PCR-based method and genotypes can be obtained from crude extracts after 1-2 weeks growth rather than four weeks growth required for IS*6110* RFLP. MIRU-VNTR results are in a digital format which makes data exchange easier. Preliminary studies showed that MIRU-VNTR typing is highly reproducible and discriminatory (Mazars *et al.*, 2001;Savine *et al.*, 2002;Supply *et al.*, 2001). A study of 93 isolates from Tanzania showed that VNTR typing can help to overcome the issues caused by strains with low copy numbers of IS*6110*. ETR typing reduced the number of isolates clustered by RFLP from 33/48 low copy number strains to 24/48 which increased the HGDI from 0.892 to 0.957. The addition of five locus ETR typing reduced the proportion of low-copy-number isolates that would have been clustered by IS*6110* RFLP alone by 27% (Barlow *et al.*, 2001).

3.1.5. The discriminatory power of MIRU-VNTR typing has been increased by successive increments in the number of MIRU-VNTR loci analysed

MIRU-VNTR typing achieves a high level of discrimination by combining genotyping data from multiple independent loci. There have been three widely used collections of MIRU-VNTR loci published since 1998, each with an increased number of loci and associated increased discriminatory power and utility. The first collection was published in 1998 and used five ETR loci (Frothingham and Meeker-O'Connell, 1998). A second collection of MIRU-VNTR loci was published in 1998 and used 12 MIRU-VNTR loci (Supply *et al.*, 2001). The most recent "3rd generation" of MIRU-VNTR loci uses 24 loci that includes a subset of 15 highly discriminatory (HD) loci selected from multiple studies that examined other potentially useful loci (Table 3.1) (Supply *et al.*, 2006). The highest level of performance exhibited by MIRU-VNTR typing was achieved when IS*6110* RFLP clusters with no apparent epidemiological links were differentiated by the 24 MIRU-VNTR loci.

3.1.6. MIRU-VNTR loci set terminology

Within this chapter, five sets of MIRU-VNTR loci that vary in locus composition and number will be analysed (Table 3.1): the five ETR loci that will be called “ETR typing”; the 12 MIRU loci will be called “MIRU typing”; the five ETR and 12 MIRU loci combined will be called “ETR+MIRU typing”; the recently internationally optimised set of 24 MIRU-VNTR loci (ETR, MIRU, and VNTR+ sets) will be called “24 loci typing” with a subset of these 24 loci that contains 15 highly discriminatory loci will be called “HD15 typing”.

Generation	Loci Set name	VNTR	Locus name	Alias	ETR	MIRU	24 loci	HD 15	Reference
1st	ETR	2165	ETR-A		✓		✓	✓	Frothingham, 1998
		2461	ETR-B		✓		✓		
		577	ETR-C		✓		✓	✓	
		580	ETR-D*		✓		✓	✓	
		3192	ETR-E*		✓		✓	✓	
2nd	MIRU	154	MIRU-02			✓	✓		Supply, 2001
		580	MIRU-04*			✓	✓		
		960	MIRU-10			✓	✓	✓	
		1644	MIRU-16			✓	✓	✓	
		2059	MIRU-20			✓	✓		
		2531	MIRU-23			✓	✓		
		2687	MIRU-24			✓	✓		
		2996	MIRU-26			✓	✓	✓	
		3007	MIRU-27	QUB-5		✓	✓		
		3192	MIRU-31*			✓	✓		
		4348	MIRU-39			✓	✓		
		802	MIRU-40			✓	✓	✓	
3rd	VNTR+	0424	VNTR0424	Mtub04			✓	✓	Supply, 2006
		1955	VNTR1955	Mtub21			✓	✓	
		2163b	VNTR2163b	QUB11b			✓	✓	
		2347	VNTR2347	Mtub29			✓		
		2401	VNTR2401	Mtub30			✓	✓	
		3171	VNTR3171	Mtub34			✓		
		3690	VNTR3690	Mtub39			✓	✓	
		4052	VNTR4052				✓	✓	
		4156	VNTR4156				✓	✓	

Table 3.1. Composition of three published MIRU-VNTR loci sets.

*There are two MIRU-VNTR loci that are present in both the ETR and MIRU loci sets but are the same locus. The two loci are ETR-D/MIRU-04 and ETR-E/MIRU-31.

3.2. AIMS

This work presented in this chapter details two studies that evaluated the utility and precision of MIRU-VNTR typing.

The aim of this chapter was to compare molecular genotyping data obtained from the various VNTR sets and IS*6110* RFLP versus known conventional epidemiological data gathered from outbreak investigation records.

We hypothesise that MIRU-VNTR typing can provide equivalent DNA fingerprinting results to that obtained by IS*6110* RFLP across four criteria which can be used to assess the utility of a DNA fingerprinting method: typing ability, discriminatory power, reproducibility, and applicability (Hawkey and Kerr, 2003). The four criteria will be measured as follows: typing ability will be measured by the proportion of all isolates that have typing data; discriminatory power will be measured by calculation of the highest HGDI (See Methods Section 2.9.4) and lowest clustering rates (Methods Section 2.9.3); reproducibility will be assessed by variability between epidemiologically linked patients; and applicability will be assessed by the potential ability for universal prospective DNA fingerprinting on a regional scale.

3.2.1. Collection 1: Comparison of IS*6110* RFLP and 15 locus MIRU-VNTR data.

The first study undertaken in 2003 assessed whether MIRU-VNTR typing could be used as a rapid alternative to IS*6110* RFLP typing by comparing IS*6110* RFLP data to 1st (ETR) and 2nd (MIRU12) generation MIRU-VNTR loci data in four clusters (Collection 1) with defined epidemiological links.

3.2.2. Collection 2: Comparison of 15 and 24 locus MIRU-VNTR typing

In 2009, to evaluate the increase in discrimination offered by the 3rd generation of internationally optimized 24 loci, we carried out an evaluation of 3rd generation MIRU-VNTR data in seven clusters (Collection 2) defined by 1st and 2nd generation MIRU-VNTR loci with a spectrum of epidemiological links. The epidemiological evidence for the significance of the cluster ranged from complete linkage to no evidence of social linkage. Collection 2 compared 24 and 15 loci in clusters with varying degrees of epidemiological linkage. Collection 2 was used to assess the improvement in correlation between conventional epidemiological and molecular data provided by analysing 24 loci.

The technical properties of MIRU-VNTR typing could enable a large-scale prospective population based study where every isolate of *M. tuberculosis* is typed which in turn would improve our understanding of the epidemiology and transmission of *M. tuberculosis* in the Midlands.

3.3. METHODS

Primary specimens were processed as described in Methods Section 2.3 and cultures were grown on LJ slopes for at least four weeks (Section 2.3.1). For collection 1, DNA was extracted using the CTAB method (Section 2.5.1) and DNA was extracted using a heating and sonication method for strains in Collection 2 (Section 2.3.3). IS*6110* RFLP was carried out using the recommended international method (Section 2.5). For MIRU-VNTR typing of strains in Collection 1, the five ETR and 12 MIRU loci were analysed using the primers and PCR conditions described by Frothingham and Supply. PCR amplicons were analysed by agarose gel electrophoresis (Section 2.6.1). For Collection 2, the classical 15 MIRU-VNTR loci were analysed using non-dHPLC (Section 2.7) with the additional nine loci amplified as described by Supply using agarose gel electrophoresis (Section 2.6.2). Epidemiological data for each cluster was collected (Section 2.9.8) and compared to molecular data to assess the degree of concordance.

3.4. RESULTS

3.4.1. Collection 1: Discriminatory power of MIRU-VNTR and IS*6110* RFLP typing in clusters with defined epidemiological links.

The proportion of clustering and discriminatory power exhibited by each of the three sets of VNTR loci and the gold standard for DNA fingerprinting (IS*6110* RFLP) was assessed (Table 3.2). 1st generation ETR typing using five loci exhibited the highest rate of clustering (75%) and the lowest HGDI value (0.75, 95% CI 0.682-0.818). Results obtained with the 2nd generation 12 MIRU loci and combining the five ETR loci and 12 MIRU loci (ETR+MIRU12) were equivalent as the less discriminatory ETR loci did not split any 12 locus MIRU profiles in this study. ETR+MIRU12 typing did produce a slightly lower rate of clustering (32%) than IS*6110* RFLP (40%) with one more cluster and three fewer clustered isolates across the four clusters with similar levels of diversity as determined by the HGDI. There were three instances of indistinguishable strains identified in two or more investigations. MIRU-VNTR profiles 94265 254326223513 (isolate 78) and 71466 254326153622 (isolate 295) were present as single strains in Investigations A and C were clustered by IS*6110* RFLP as both strains possessed a single copy of IS*6110* in a similar genomic location. MIRU-VNTR profile 32433 224325153324 was the major strain in Investigation B and possessed only two copies of IS*6110*. A single strain in Investigation C also possessed two copies of IS*6110* and had a similar but different MIRU-VNTR profile (32433 224326153324). No specific epidemiological links were identified between any of the investigations.

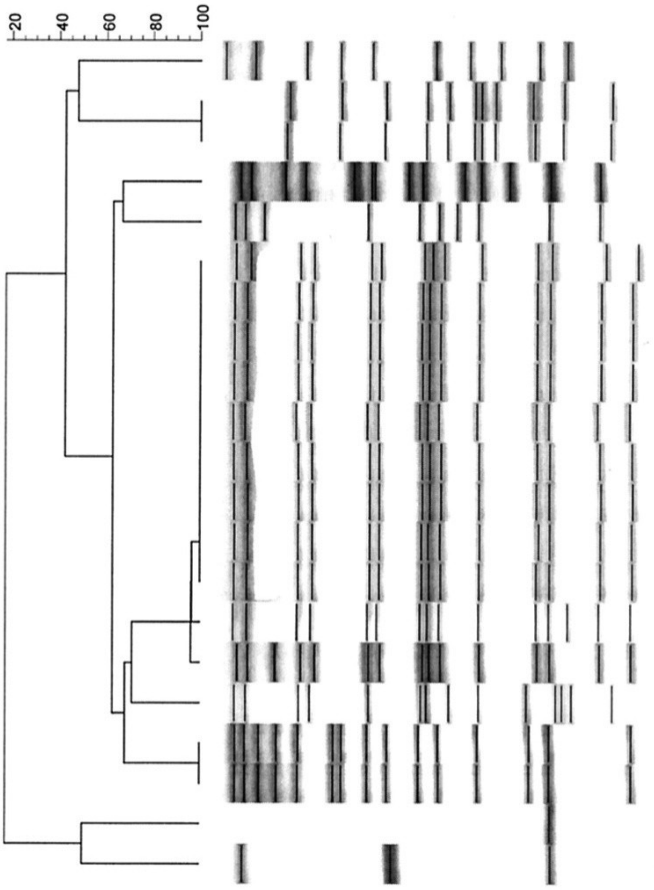
Typing method	**ETR (5 loci)**	**MIRU (12 loci)**	**ETR+MIRU (17 loci)**	**IS*6110* RFLP**
No. of clusters	9	21	21	20
No. of clustered isolates	49	38	38	41
No. of unique isolates	4	15	15	12
Total isolates	53	53	53	53
Clustering (%)	75	32	32	40
HGDI (95% CI)	0.750 (0.682-0.818)	0.881 (0.830-0.932)	0.881 (0.830-0.932)	0.89 (0.845--0.936)
Concordance with RFLP in isolates part of main clusters	30/32 (94%)	31/32 (97%)	31/32 (97%)	-
Concordance with RFLP in isolates outside main clusters	12/21 (57%)	20/21 (95%)	20/21 (95%)	-

Table 3.2. Discriminatory power of ETR, MIRU12, ETR+MIRU, and IS*6110* RFLP in Collection 1.

3.4.2. Collection 1: Correlation between MIRU-VNTR and IS*6110* data in clusters with defined epidemiological links.

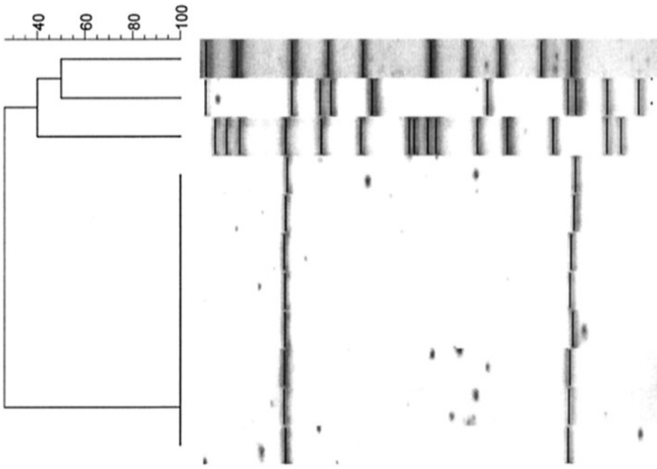
The clustering of isolates by IS*6110* RFLP is shown in Figure 3.1. The ETR and MIRU profiles are shown adjacent to each isolate number. Inspection of the data set reveals a very strong correlation between IS*6110* RFLP type and both the ETR and the MIRU type. The three different sets of MIRU-VNTR loci were compared to RFLP fingerprints within each investigation based on concordance in isolates that were assigned to the major epidemiological cluster by RFLP typing and those that were excluded from the major epidemiological cluster. ETR, MIRU, and ETR+MIRU concorded with RFLP data in at least 94% isolates within and outwith major epidemiological clusters. The only exception was five locus ETR typing in strains that had been excluded from epidemiological clusters by RFLP as ETR typing clustered more strains than RFLP in this group (57% vs 14% respectively) (Table 3.2). The ETR profiles identified the major clusters in investigations B and C and identified 15/20 (75%) of the unrelated (by IS*6110* RFLP analysis) isolates as unrelated by ETR typing. In investigation A, the predominant group of indistinguishable isolates (by IS*6110* RFLP analysis) belonged to ETR type 42235, and another six unrelated isolates showed the same ETR code. Isolates 09 and 92 were less than 100% related by IS*6110* RFLP (each had an additional band) but had identical MIRU profiles. These patients were considered to be epidemiologically part of the main cluster. The same was true to a lesser extent of investigation D, in which isolates 17, 130, and 221, from family members, were all ETR type 42235 but one (isolate 221) had a different subtype by IS*6110* RFLP. The other two patient isolates had indistinguishable IS*6110* RFLP patterns but differing MIRU codes (two different alleles). MIRU typing split ETR type 42235 into subtypes which correlated strongly with IS*6110* RFLP and epidemiological findings.

Investigation A

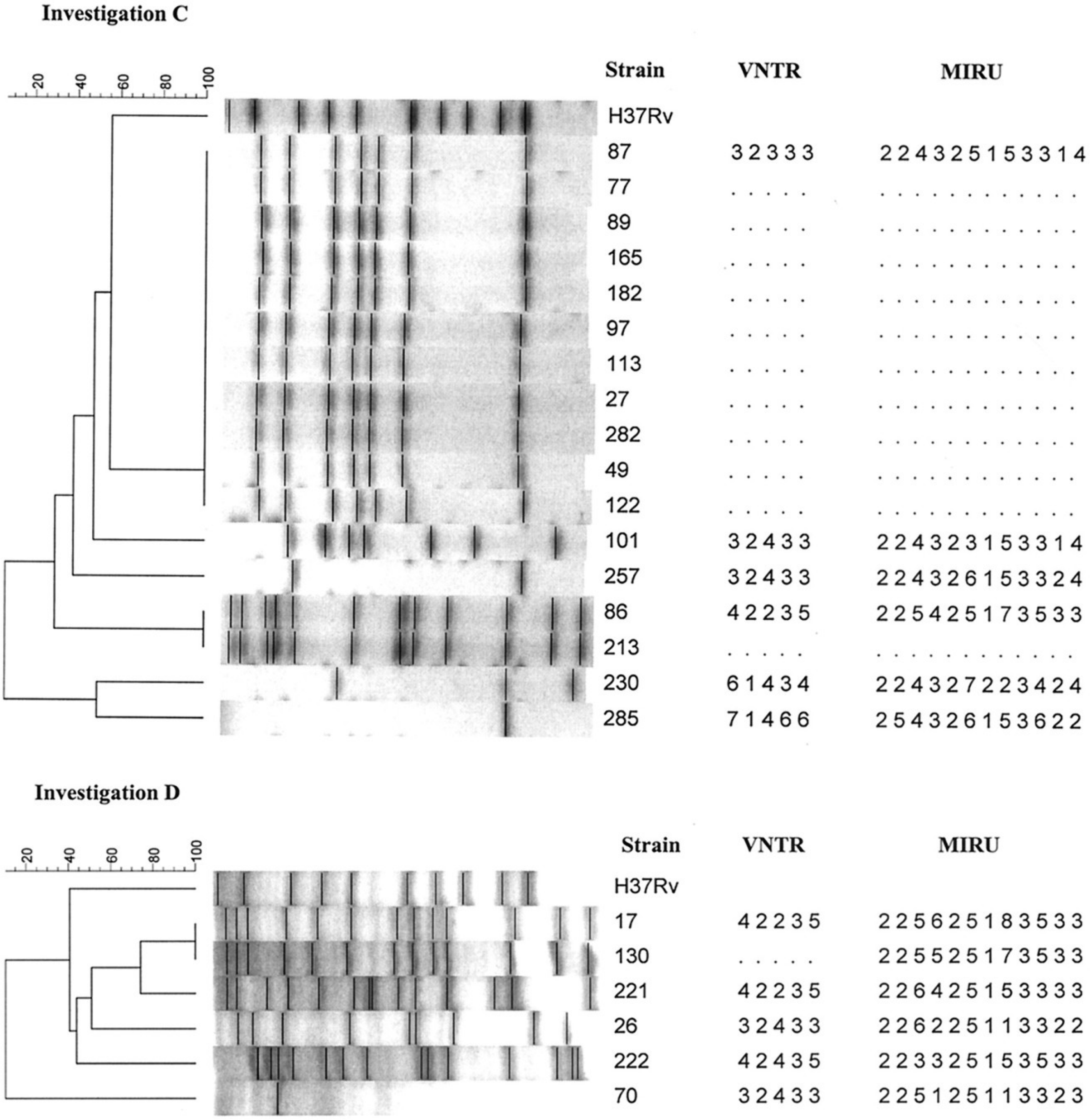


Strain	VNTR	MIRU
H37Rv	3 3 4 3 3	2 2 3 2 2 6 1 3 3 3 2 1
08	3 2 4 3 4	2 2 3 3 2 5 1 5 3 4 2 1
51	. . . . .	. . . .
90	4 2 2 3 5	2 2 5 4 2 5 1 5 3 5 3 4
56	4 2 2 3 5	2 2 6 3 2 5 1 5 3 5 3 3
201	4 2 2 3 5	2 2 5 4 2 5 1 7 3 5 3 3
40	. . . . .	. . . .
37	. . . . .	. . . .
80	. . . . .	. . . .
258	. . . . .	. . . .
11	. . . . .	. . . .
05	. . . . .	. . . .
35	. . . . .	. . . .
261	. . . . .	. . . .
09	4 2 2 3 5	2 2 5 4 2 5 1 7 3 5 3 3
92	4 2 2 3 5	2 2 5 4 2 5 1 7 3 5 3 3
163	4 2 2 3 5	2 2 5 5 2 5 1 5 3 5 3 3
67	4 2 2 3 5	2 2 5 4 2 5 1 5 3 5 3 3
68	. . . . .	. . . .
78	9 4 2 6 5	2 5 4 3 2 6 2 2 3 5 1 3
263	9 4 2 6 5	2 5 4 3 2 6 2 2 3 5 1 3

Investigation B



Strain	VNTR	MIRU
H37Rv		
121	3 2 3 3 3	2 2 5 3 2 5 1 5 3 3 2 3
239	4 2 2 3 4	2 2 5 4 2 5 1 7 3 4 3 3
134	3 2 4 3 3	2 2 4 3 2 5 1 5 3 3 2 4
03	. . . . .	. . . .
00	. . . . .	. . . .
158	. . . . .	. . . .
24	. . . . .	. . . .
75	. . . . .	. . . .
79	. . . . .	. . . .
32	. . . . .	. . . .



Investigation C

Strain	VNTR	MIRU
H37Rv		
87	3 2 3 3 3	2 2 4 3 2 5 1 5 3 3 1 4
77	. . . . . .	. . . .
89	. . . . . .	. . . .
165	. . . . . .	. . . .
182	. . . . . .	. . . .
97	. . . . . .	. . . .
113	. . . . . .	. . . .
27	. . . . . .	. . . .
282	. . . . . .	. . . .
49	. . . . . .	. . . .
122	. . . . . .	. . . .
101	3 2 4 3 3	2 2 4 3 2 3 1 5 3 3 1 4
257	3 2 4 3 3	2 2 4 3 2 6 1 5 3 3 2 4
86	4 2 2 3 5	2 2 5 4 2 5 1 7 3 5 3 3
213	. . . . . .	. . . .
230	6 1 4 3 4	2 2 4 3 2 7 2 2 3 4 2 4
285	7 1 4 6 6	2 5 4 3 2 6 1 5 3 6 2 2

Investigation D

Strain	VNTR	MIRU
H37Rv		
17	4 2 2 3 5	2 2 5 6 2 5 1 8 3 5 3 3
130	. . . . . .	2 2 5 5 2 5 1 7 3 5 3 3
221	4 2 2 3 5	2 2 6 4 2 5 1 5 3 3 3 3
26	3 2 4 3 3	2 2 6 2 2 5 1 1 3 3 2 2
222	4 2 4 3 5	2 2 3 3 2 5 1 5 3 5 3 3
70	3 2 4 3 3	2 2 5 1 2 5 1 1 3 3 2 3

Figure 3.1. IS*6110* RFLP similarity dendrograms

Dendrograms were created using the Dice co-efficient and displayed via UPGMA. Isolates were from four distinct geographical locations in the Midlands region of the United Kingdom (investigations A to D). VNTR and MIRU typing results are shown for each isolate. Dots represent alleles with the same numbers of repeat units, indicating isolates with indistinguishable types.

3.4.3. Collection 1: Epidemiology of clusters

Investigation A concerned a complex, large-scale outbreak centred on a single index case in a school. However, the high background level of tuberculosis in the community and in pupils in the school not directly related to the index case required the use of molecular typing in order to fully understand routes of transmission. It was originally thought that isolates 78 and 263 were the index cases who were part of the general community and not school pupils. However, application of five locus ETR typing excluded these two patients from the cluster and focussed public health efforts onto the school. Other pairs of cases (e.g., 67-68, 08-51) which were not caused by the outbreak strain but were subsequently linked epidemiologically with each other were identified. Two isolates, 09 and 92, showed single additional bands in the IS*6110* pattern compared to the outbreak strain but had identical MIRU codes. Epidemiologically, these isolates were indisputably part of the main outbreak, a case where the greater stability of MIRU enhanced the clarity of typing. A number of subsequent cases that may have represented failure of implemented control measures were quickly excluded from the outbreak cluster.

In investigation B, molecular typing was used to examine isolates from a community with a previously low incidence of tuberculosis where a sharp increase in the number of cases over the preceding 12 to 18 months had been noted. In investigation B, MIRU typing linked four family members (isolates 134, 03, 32, and 79), as was expected, but also linked these with other cases within the same city which were indistinguishable by typing and led to the discovery of an unsuspected social link. Two other isolates, 121 and 239, which might have been linked, were shown to be clearly unrelated. Typing was unable to distinguish between strains from two apparently distinct epidemiological clusters, prompting further investigation, which revealed unsuspected social links between the two clusters. The information was used to plan the extent of contact examination.

Investigation C was triggered by the recognition of an increase in the number of isoniazid-resistant strains over a five year period and the concern that transmission of a resistant clone was occurring. MIRU typing confirmed a common profile for 11/17 (65%) isolates which also had indistinguishable IS*6110* RFLP patterns. In this cluster of 11 patients, nine patients were UK-born and two patients originated from the Indian subcontinent (ISC). This was in contrast to the other six non-clustered isolates, all of which were from patients who originated from the ISC. Preliminary epidemiological investigation demonstrated geographical clustering of some cases with indistinguishable strains.

Investigation D concerned a small group of strains primarily from a family of three patients that were thought to be clustered, but IS*6110* RFLP typing showed them to be all unrelated with the exception of isolates 17 and 130, which were from a husband and wife. A close family contact (isolate 221) had a different RFLP and MIRU type, as did all the other contacts. MIRU typing showed differences at two loci, although the isolates were separated by three years.

3.4.4. Collection 2: The discriminatory power of 24 MIRU-VNTR loci compared to 15 locus ETR+MIRU typing.

The proportion of clustering and discriminatory power exhibited by ETR+MIRU typing, the recently published optimal set of 24 loci, and highly discriminatory set of 15 loci was assessed in seven clusters defined by ETR-MIRU typing (Table 3.3). Application of the 3rd generation of 24 MIRU-VNTR loci increased the number of clustered profiles from seven to eight and resulted in 19 strains differentiated as unique or orphan profiles. All 19 of these strains split further by 24 loci typing were considered to concord with the available epidemiological data. The 3rd generation loci reduced clustering from 92% to 59% and increased the HGDI from 0.793 (95% CI 0.730-0.856) to 0.890 (95% CI 0.835-0.944). Within this study, there were no differences in clustering or levels of differentiation obtained when the 24 loci (8 clusters, 47 clustered isolates, 59% clustering and HGDI = 0.890) or the subset of highly discriminatory 15 loci (8 clusters, 47 clustered isolates, 59% clustering and HGDI = 0.890) were applied.

Typing method	**ETR+MIRU (17 loci)**	**VNTR+ (24 loci)**	**HD15 (15 loci)**
No. of clusters	7	8	8
No. of clustered isolates	68	47	47
No. of unique isolates	0	19	19
Total isolates	66	66	66
Clustering (%)	92	59	59
HGDI (95% CI)	0.793 (0.730 - 0.856)	0.890 (0.835 - 0.944)	0.890 (0.835 - 0.944)

Table 3.3. Discriminatory power of three sets of MIRU-VNTR loci in Collection 2.

Collection 2 contained 66 isolates from seven clusters defined by ETR-MIRU typing.

3.4.5. Collection 2: Concordance between epidemiological and genotyping data.

The concordance between available epidemiological data and DNA fingerprinting data was evaluated (Table 3.4). The seven investigated clusters ranged in the degree of certainty of epidemiological links between patients in a cluster. In cluster 2, epidemiological links had been investigated by reviewing patient case notes but no common links were identified. Epidemiological links between patients were investigated with identification of common social links for most or all patients in clusters 5 and 6. In clusters 3, 4, and 7 some but not all of the patients within each cluster were linked by conventional epidemiological data. Cluster 1 was a single locus variant (SLV) of the ETR+MIRU profile present in cluster 5. The four patients in this cluster lived within the same county as Cluster 5 but in different towns. The hypothesis for investigating this cluster was to examine the potential further differentiation afforded by 24 loci in strains which only differ at one locus in the 15 loci.

Cluster 1 and 2 were separated by the 24 loci set into two smaller clusters with two and seven orphan profiles in cluster 1 and 2 respectively. Cluster 3 was split in half whereas cluster 4 had only one isolate split by the 24 loci with a cluster of 10 isolates remaining intact. The largest cluster (cluster 5) had 5/24 isolates split by the 24 loci with a central cluster of 19 isolates still indistinguishable. Cluster 6 still remained as a cluster of four isolates even after application of the 24 loci and cluster 7 was reduced from a cluster of six to four isolates.

In all seven clusters containing 66 isolates, there was an increase in the concordance between conventional epidemiological data from 38/66 (58%) isolates for 17 locus ETR+MIRU typing to 59/66 (89%) isolates for 24 locus MIRU-VNTR typing.

3.4.6. Collection 2: Conventional epidemiology of investigated clusters

The clusters investigated in Collection 2 were mainly from 2004. Due to the intensive nature of reviewing patient case notes, the epidemiological links between every isolate and associated patient were not investigated. Several of the clusters represented large, complex clusters where the social networks extended beyond the investigated patients. Definite epidemiological links were identified when a patient was named as a contact of another patient that was known to have been infected with the same MIRU-VNTR profile or when a common specific link was identified. Isolates from identified contacts were not always typed further. Review of case notes without positive identification of any possible links was accepted as proof of no epidemiological linkage between patients.

Cluster 2 was the second most prevalent ETR+MIRU profile in the Midlands in 2004. Within 2004, there were a total of 19 isolates with this profile in seven locations across the Midlands. Strains were selected such that at least one representative strain from each location was analysed. Epidemiological links were investigated but none were found in any of the nine patients. The 24 loci split this cluster into seven orphan profiles and a pair of isolates which increased the number of isolates with concordant epidemiological and genotyping data from none to seven (Table 3.4).

Cluster 3 was a cluster of eight patients identified by 17 locus ETR+MIRU typing where five patients were linked socially through common contact with a family and public houses were also a common factor. All eight isolates identified by ETR+MIRU typing in 2004 across the Midlands were selected for further analysis. The four patients who were linked with family 1 remained clustered after 24 loci typing. Two of the four other patients had possible connections through public bars but the two isolates were split into unique isolates by 24 loci typing. The other two patients had no epidemiological links at all and were also split into unique isolates.

Cluster 4 included 11 strains that are part of the most prevalent ETR+MIRU profile in the Midlands. In 2004, there was a total of 37 isolates across six cities in the West and East Midlands with this profile. At least one representative isolate from each location was chosen for further molecular and epidemiological analysis. Six patients had known contact with other patients infected with this strain. Only one isolate was split from this ETR+MIRU cluster by 24 loci. Four clustered isolates did not have any specific epidemiological links identified.

Cluster 5 was a long-standing outbreak first identified in the 1990's with multiple links between families, various public houses and homeless hostels. All 24 isolates genotyped between 2000 and 2007 were selected for further analysis. From the prospective phase of DNA fingerprinting presented later in chapter 5 of this thesis, 32433 2432515324 was the most prevalent strain (11/389, 3%) in the local area with the SLV only identified in one patient between 2004 and 2009 within the same town. The degree of relatedness between these two ETR+MIRU profiles was queried as the patients in cluster 1 were not epidemiologically linked to any patients in cluster 5. Two of the isolates in cluster 1 were still clustered by 24 loci with two orphan strains. All of the four patients in cluster 1 were differentiated further from all patients in cluster 5 by 24 loci. The further differentiation exhibited by the 24 loci was accepted as achieving concordance with conventional epidemiological data since the epidemiological hypothesis was that these strains are not related. The pair of clustered isolates (Cluster 1, 2 and 3) were actually resident in the same town. Application of the 24 loci revealed a central clone present in all of the mentioned overlapping links with 5/24 isolates separated from the main ETR+MIRU profile. The pair of patients (Cluster 5, 21 and 23) worked together and none of the orphan strains had any specific social links.

Cluster 6 was a cohort of four patients who all worked in two warehouses of the same company in 2004. Across the Midlands, there were only four isolates with the 42433 2331513321 profile. There was strong epidemiological linkage between patients which was confirmed as the four isolates were still indistinguishable after analysis by 24 loci. Cluster 7 contained six patients with various social links which included known contact with members of the same family that were culture positive or public bars or prison. This cluster was separated into a cluster of four isolates and two orphan profiles by the 24 loci. The patients with the two orphan profiles did not have any linking epidemiology to the family, public bars, or prison. Review of case notes hinted at possible links between patients 2, 3, 5, and 6 in this cluster but the linkage was probably the weakest of all the clusters. The presence or absence of epidemiological data in this cluster correlated with the 24 loci clustering. So, the further differentiation of the two patients without known social links and clustering of four patients with some form of epidemiological links was accepted as full concordance between conventional and molecular epidemiological data in this cluster.

Cluster	ETR+MIRU profile Cluster Description	Extra nine loci profiles	No. isolates where epidemiology concords with VNTR data (%)		Total
			ETR+MIRU	VNTR+	
1	**32433 224325153323 (2002-07)** Four culture positive patients resident within same county as Cluster 5 but in three different towns.	2 x 344423692 2 x orphan	0 (0)	4 (100)	4
2	**42235 225425173533 (2004)** 2nd most prevalent profile in 2004. Nine isolates from seven locations. Links investigated but none found.	2 x 442423384 7 x orphan	0 (0)	7 (78)	9
3	**42435 223325153533 (2004)** Members and contacts of family 1. Possible extended social links to public bars.	4 x 456443362 4 x orphan	4 (50)	8 (100)	8
4	**32333 224325153314 (2004)** Most prevalent profile. No specific social links. Nine isolates from six locations. Six patients had contact with individuals who were culture positive with an indistinguishable ETR+MIRU profile.	10 x 434443183 1 x orphan	6 (55)	7 (64)	11
5	**32433 224325153324 (2000-07)** Multiple well-defined overlapping links: family, hostel, bars.	19 x 443443153 2 x 442443151 3 x orphan	20 (83)	23 (96)	24
6	**42433 223315133321 (2004)** All four patients worked in the same warehouse.	4 x 236423252	4 (100)	4 (100)	4
7	**42235 226425163533 (2004)** Various social links including public bar, prison, and a family.	4 x 442423382 2 x orphan	4 (67)	6 (100)	6
		Total isolates	38 (58)	59 (89)	66

Table 3.4. Collection 2: Concordance between molecular and epidemiological data in clusters identified by ETR+MIRU typing.

Please see Appendix II for more detailed descriptions of the epidemiological associations.

3.5. DISCUSSION

This study has demonstrated that 17 locus ETR+MIRU typing provides comparable genotyping data to IS*6110* RFLP, the historical gold standard method for *M. tuberculosis* DNA fingerprinting, in a collection of clusters (Collection 1) where conventional epidemiological data was known.

In a follow-on evaluation of the 3rd generation set of 24 MIRU-VNTR loci applied to a collection of genotypically defined clusters with varying degree of concordance between conventional and molecular data (Collection 2), it was shown that the 24 loci set improved the degree of concordance between conventional epidemiological and molecular data. The optimal 24 loci set separates isolates with weak or non-existent epidemiological links but does not separate isolates with strong epidemiological links.

By comparing the results obtained from MIRU-VNTR and IS*6110* RFLP typing within this chapter to the criteria described in the Aims Section (Section 3.2) we can assess the utility of each DNA fingerprinting method in terms of typing ability, discriminatory power, reproducibility, and applicability.

The ability of IS*6110* RFLP to type all isolates in this chapter was not 100% as 17/53 (32%) strains had low copy numbers of IS*6110* and therefore cannot be accurately typed by RFLP. A typing result was obtained for all strains in all versions of MIRU-VNTR typing in both strain collections. Within collection 1, 12 MIRU loci and ETR+MIRU typing had equivalent HGDI values (0.881, 95% CI 0.830 - 0.932) to RFLP (0.890, (95% CI 0.845 - 0.936) when overlapping confidence intervals were taken into account and slightly lower clustering rates

(32% vs 40% respectively). The slight variation in diversity and clustering identified when using RFLP (higher HGDI and higher clustering) was caused by clustering of low copy number isolates by MIRU-VNTR typing (strains 78 and 263 in Investigation A) which were split by RFLP. Reproducibility within Collection 1 can be assessed by the serial detection of conserved clones within each cluster, 31/32 clustered isolates that were part of the major epidemiologically linked clusters were also clustered by ETR+MIRU typing. The only discrepancy was strain 221 in Investigation D. RFLP had a similar level of agreement but with the appearance of additional bands in strains 92 and 163 in the major cluster in Investigation A. However, clusters based on RFLP data are commonly analysed using a +/- one band difference (Savine *et al.*, 2002). Reproducibility could not be assessed within Collection 2 as some of these clusters have quite complex epidemiology which has not been fully elucidated from initial epidemiological investigations. MIRU-VNTR typing is a discriminatory, reproducible method that can type all strains of *M. tuberculosis*. The combination of these criteria makes MIRU-VNTR typing a highly applicable DNA fingerprinting method.

3.5.1. ETR+MIRU typing can be used as a rapid 1st-line alternative to IS*6110* RFLP

From the results of Collection 1, it was concluded that ETR+MIRU typing could be used as a rapid 1st-line alternative to IS*6110* RFLP typing. To enable full implementation as a 1st-line typing method, transfer from agarose gel electrophoresis to automated analysis is required to enable regional universal prospective typing of all *M. tuberculosis* strains. From Collection 2, it was concluded that the 17 loci should be extended to the optimal 24 loci set as the 1st-line method for DNA fingerprinting of *M. tuberculosis* strains in the Midlands. From 2003-2009, all isolates received at the HPA MRCM were typed using the ETR+MIRU loci with 24 loci analysis implemented in January 2010.

3.5.2. Collection 1: IS*6110* RFLP compared to 15 locus MIRU-VNTR data

In our study, ETR+MIRU typing accurately identified the clusters in investigations A, B, and C and confirmed the lack of a suspected cluster in the case of investigation D. Our experience of the MIRU-VNTR loci is in accord with the report of Supply and colleagues (Supply *et al.*, 2001), who applied MIRU typing to a blinded set of 90 strains from 38 countries, and demonstrated that it was 100% reproducible, sensitive, and specific for *M. tuberculosis*. Mazars and colleagues also applied MIRU typing to 12 clusters which contained 28 epidemiologically related isolates and found that all 28 were also clustered by 12 locus MIRU typing (Mazars *et al.*, 2001).

The ability to rapidly and conclusively exclude strains (and therefore cases) from large-scale outbreaks is very useful for public health teams. We found, as others have, that relying on absolute identity of IS*6110* RFLP patterns leads to erroneous exclusion of a significant proportion of isolates. In 49 patients with two isolates first cultured more than 90 days apart, 12/49 (25%) patients showed changes in their IS*6110* genotypes (Yeh *et al.*, 1998). MIRU typing in our hands resulted in the potential exclusion of only 2/56 isolates, a finding similar to a recent report (Savine *et al.*, 2002). In investigation B, the outbreak strain carried only two copies of IS*6110*, which can easily result in the false identification of clusters when IS*6110* RFLP is used as a single method (Gillespie *et al.*, 2000). However, distinct and unique MIRU codes were found.

In Collection 1, two of these isolates in one cluster (Investigation A) had IS*6110* RFLP patterns differing by one band; otherwise, the concordance was 100%. The pair of strains identified by IS*6110* RFLP but with differences in the MIRU-VNTR profile in Investigation D were isolated from a husband and wife three years apart, suggesting a higher rate of evolution at loci 16 and 26 compared to that of IS*6110* RFLP patterns. Both of these loci have been reported to exhibit a high degree of diversity in strains (Mazars *et al.*, 2001). The stability of MIRU has recently been investigated with pairs of isolates from patients separated by as many as six years (Savine *et al.*, 2002). Only in a single case were isolates found with identical IS*6110* patterns and a single change in a MIRU locus. Our finding for a single strain with a high copy number of IS*6110* is in contrast to the observation that the "molecular clock" for MIRU runs at a lower rate than that for IS*6110* RFLP (Supply *et al.*, 2001). In investigation A, IS*6110* RFLP failed to identify a pair of epidemiologically related strains identified by MIRU; these strains also had a rare VNTR type, confirming their close relationship.

This study demonstrated that ETR+MIRU typing provided equivalent data when compared to IS*6110* RFLP in the investigation of epidemiologically linked outbreaks of *M. tuberculosis*. The use of a two typing method strategy has been evaluated by others for population-screening studies (Wilson *et al.*, 1998). The greater technical reliability, ease of automation and data storage, and analysis of typing data make MIRU-VNTR typing, in my opinion, superior to IS*6110* RFLP. ETR+MIRU typing system could be automated and the digital results produced facilitate easier data storage and analysis than IS*6110* RFLP which makes MIRU-VNTR attractive for outbreak investigation, and as a method for compiling a comprehensive national database.

3.5.3. Successive refinements in the MIRU-VNTR loci used in typing schemes have improved the epidemiological relevance of MIRU-VNTR data.

The initial description of MIRU loci in 31 *M. tuberculosis* complex strains from around the world identified 12 loci that contained variable numbers of tandem repeats (VNTRs) in strains that originated from different geographical locations (Supply *et al.*, 2000). A second study compared IS*6110* RFLP and MIRU-VNTR data obtained from isolates involved in laboratory cross contamination, identified transmission events, relapses or different anatomical sites in Paris. This study showed that in 12 distinct clusters, 28/28 epidemiologically related isolates were also clustered by MIRU-VNTR typing which included strains that had few IS*6110* copies. This study in Paris indicated that MIRU-VNTR loci are stable enough to accurately track outbreaks (Mazars *et al.*, 2001). A third study of a blinded reference set of 90 strains from 38 countries then showed that high-throughput MIRU-VNTR typing using a DNA sequencer was 100% reproducible, sensitive, and specific for *M. tuberculosis* complex isolates, a performance that had not been achieved by any other typing method tested in the same conditions (Supply *et al.*, 2001).

Two studies from North America evaluated the potential for the implementation of MIRU-VNTR typing as a 1st line alternative method to IS*6110* RFLP. A study of 259 isolates from Wisconsin between 2000 and 2003 concluded that spoligotyping and MIRU-VNTR typing combined together provided adequate discrimination in most cases. If required, IS*6110* RFLP could be used as a secondary typing method (Cowan *et al.*, 2005). This study formed the basis for the US national strategy for genotyping *M. tuberculosis* strains from 2005 onwards. A study in Manitoba, Canada compared MIRU-VNTR to IS*6110* RFLP for 126 isolates identified in 2003 and found that both methods had similar levels of discrimination and the clustering of isolates using MIRU-VNTR data correlated with IS*6110* RFLP-derived clustering (Blackwood *et al.*, 2004).

In a comparison of nine recently described DNA typing methods that had been suggested as alternatives for IS*6110* RFLP analysis, MIRU-VNTR using the 12 MIRU loci was found to be the most reproducible method and second most discriminatory method (Kremer *et al.*, 2005a).

3.5.4. Two studies that demonstrated sub-optimal performance of 12 locus MIRU-VNTR typing

Two studies reported results that countered the equivalent performance of 12 locus MIRU-VNTR typing when compared to IS*6110* RFLP. The first study from Montreal showed that in high IS*6110* copy number isolates, the sensitivity of spoligotyping (83%) and MIRU-VNTR typing (52%) in identifying isolates clustered by IS*6110* RFLP and specificity in the identification of unique IS*6110* RFLP patterns was low (56% for MIRU-VNTR typing and 40% for spoligotyping). The authors warned that these techniques had unsuitable operating parameters for population-based molecular epidemiology studies (Scott *et al.*, 2005). A later study from Spain of 134 isolates analysed by IS*6110* RFLP, MIRU-VNTR using the 12 MIRU loci, and spoligotyping showed that the HGDI index was slightly higher for IS*6110* RFLP (0.989) than for MIRU-VNTR (0.978). IS*6110* RFLP clustered 42% of isolates whereas MIRU-VNTR clustered 58% of isolates. MIRU-VNTR data in IS*6110* RFLP-defined clusters showed full concordance in 7/17 (41%) of clusters. The addition of spoligotyping data to MIRU-VNTR data reduced the proportion of clustered isolates to 43% and increased the number of concordant clusters to 11/17 (65%) (Garcia de Viedma *et al.*, 2006).

Conclusions from these two studies required cautious interpretation as both accepted IS*6110* RFLP data as the gold standard without any comparison to conventional epidemiological data. However, no molecular genotyping method for *M. tuberculosis*, including IS*6110* RFLP is 100% sensitive and specific. Therefore, to confirm the validity of clustering, each new molecular genotyping method should be evaluated by comparison to data obtained from conventional epidemiological investigation (Koksalan, 2005).

3.5.5. Identification and evaluation of VNTR loci not part of the ETR or 12 MIRU-VNTR schemes for *M. tuberculosis*

The five ETR loci were identified using the preliminary H37Rv whole genome sequence (Frothingham and Meeker-O'Connell, 1998). The 12 MIRU loci were identified using the complete genome of H37Rv and preliminary genome sequences for *M. tuberculosis* CDC1551 and *M. bovis* AF2122/97 (Supply *et al.*, 2000) whereas additional genome searches utilised the subsequent availability of the complete genome sequences of CDC1551 (Fleischmann *et al.*, 2002), and *M. bovis* AF2122/97 (Garnier *et al.*, 2003) to identify more VNTR loci (Le Fleche *et al.*, 2002). Other groups used slightly different search strategies and parameters to identify additional VNTR loci (Smittipat and Palittapongarnpim, 2000).

From these evaluations of five and 12 locus MIRU-VNTR typing, various subsequent studies published evaluations of MIRU-VNTR loci that were not originally part of the five ETR or 12 MIRU and potentially offered even further strain differentiation (Kremer *et al.*, 2005b;Le Fleche *et al.*, 2002;Smittipat *et al.*, 2005).

3.5.6. The definition and evaluation of an internationally agreed set of 24 MIRU-VNTR loci

These various evaluations of multiple different sets of MIRU-VNTR loci led to an evaluation in a single study of the technical robustness, stability, and reproducibility of 29 putative loci which were applied to a collection of 824 isolates of *M. tuberculosis* that represented all of the main global lineages. Five loci were excluded for technical reasons which included: instability in serial isolates or isolates from epidemiologically linked patients; apparent multiple alleles within a single locus; uninterpretable long amplicons; or problematic amplification by PCR. The 24 remaining loci were found to increase the number of sub-types in the collection by 40% and reduced the clustering rate fourfold when compared to the original set of 12 MIRU loci. A subset of 15 loci with the highest evolutionary rates was identified that generated 96% of the total resolution shown with the full 24 locus set (Supply *et al.*, 2006). The predictive value for assessing transmission was shown to be equal to that of IS*6110* RFLP in a companion population-based study in Hamburg, Germany (Oelemann *et al.*, 2007). The authors of this study proposed the 15 and 24 locus systems as the new standard for routine epidemiological typing and phylogenetic analysis. Interestingly, all five of the original five ETR loci are included in the 24 loci and 4/5 are included in the high-definition 15 locus set (Frothingham and Meeker-O'Connell, 1998).

3.5.7. The internationally optimised set of 24 loci improves the correlation between conventional epidemiological and molecular data

Data obtained from seven clusters in the Midlands presented in this chapter showed that the 24 loci increased the level of discrimination (HGDI of 0.793 for ETR+MIRU and 0.890 for 24 loci) but not significantly and increased the concordance between molecular and conventional epidemiological data (58% to 89% in seven clusters containing 66 isolates) which correlated with recent subsequent evaluations of the "3rd generation" of 15 and 24 MIRU-VNTR loci by other groups. Other studies in Madrid, the Netherlands, Bulgaria, and Canada showed that 24 loci typing offered a higher discriminatory power than 12 MIRU-VNTR loci and better correlation with IS*6110* RFLP data (Alonso-Rodriguez *et al.*, 2008;Maes *et al.*, 2008;Valcheva *et al.*, 2008;Christianson *et al.*, 2010). However, it must be noted that genotyping data provided by the "3rd generation" 24 locus MIRU-VNTR set is not absolute as a 39-month universal population-based study undertaken in Madrid found that 24 loci typing and IS*6110* RFLP data was concordant in 85% of typed *M. tuberculosis* isolates. Epidemiological links were identified in 84% of cases clustered by both 24 loci typing and IS*6110* RFLP. Less than 40% of isolates with discordant typing data had identified epidemiological links (Alonso-Rodriguez *et al.*, 2009).

3.6. CONCLUSIONS

There are four criteria which can be used to assess the utility of a typing method: typing ability, discriminatory power, reproducibility, and applicability (Hawkey and Kerr, 2003). For IS*6110* RFLP, technical expertise is required to achieve good reproducibility and the lengthy culture, DNA preparation and analysis procedure limits applicability to newly positive cultures or rapidly expanding clusters. MIRU-VNTR typing can type all isolates, is highly discriminatory and reproducible and can be readily applied to newly positive cultures using simple, rapid extraction methods such as thermolysis by heating to 100°C. Therefore, MIRU-VNTR can be used as a 1st line method for genotyping of *M. tuberculosis*. Transfer of VNTR allele analysis from agarose gel electrophoresis to an automated analysis system would enable a regional assessment of the epidemiology and transmission of *M. tuberculosis* across the Midlands.

4. AUTOMATED HIGH-THROUGHPUT MYCOBACTERIAL INTERSPERSED REPETITIVE UNIT TYPING OF *M. TUBERCULOSIS*.

4.1. INTRODUCTION

4.1.1. Background

Data in chapter 3 demonstrated that MIRU-VNTR typing provided similar clustering data to IS*6110* RFLP typing in clusters with known epidemiological links. Transfer of MIRU-VNTR analysis from agarose gel electrophoresis to an automated analysis system would increase the sample through-put and enhance the utility of MIRU-VNTR typing by enabling application on a regional, prospective and universal basis. The evaluation of an option for an automated system is described in this chapter.

4.1.2. A novel variation of HPLC that has been used in SNP genotyping

dHPLC is a technique that uses reversed-phase HPLC to identify Single Nucleotide Polymorphisms (SNPs) (Oefner and Underhill, 1995). dHPLC offers an automated alternative to agarose-gel based SNP genotyping techniques such as single strand conformation polymorphism analysis or denaturing gradient gel electrophoresis (Fischer and Lerman, 1979;Orita *et al.*, 1989). dHPLC detects SNPs by using a solid phase that has a differential affinity for DNA dependent on whether the two strands are entirely complementary or if there is a mismatch when a SNP is present. dHPLC was utilized in the initial studies that deciphered human evolutionary history through the analysis of human Y chromosome haplotypes (Underhill *et al.*, 1997) and has been applied to a wide range of studies that have identified mutations involved in human genetic disorders including: BRCA1 and BRCA2 in breast and ovarian cancer (Gross *et al.*, 1999); tuberous sclerosis (Choy *et al.*, 1999); and the cystic fibrosis transmembrane conductance regulator gene (Liu *et al.*, 1998).

4.1.3. Mutation detection in bacteria

dHPLC was first used in bacteria to characterise mutations conferring quinolone resistance in *S. aureus* (Hannachi-M'Zali *et al.*, 2002), *Salmonella enterica* (Eaves *et al.*, 2002), and beta-lactam resistance in gram-negative organisms (Perez-Perez and Hanson, 2002). This method has also been used for a modified version of MLST in meningococci (Shlush *et al.*, 2002) and in the detection of mutations conferring antibiotic resistance in *M. tuberculosis* (Cooksey *et al.*, 2002). All bacterial studies prior to the data presented in this chapter had used the denaturing mode of dHPLC to detect mutations. The data in this chapter describes the validation of MIRU-VNTR typing using the non-denaturing mode (non-dHPLC) that can size PCR amplicons.

4.1.4. Principles of dHPLC analysis

dHPLC utilises a form of HPLC called reverse-phase ion-pair high performance liquid chromatography (RP-IP-HPLC) to perform analytical separations. A commercially available system specifically designed for dHPLC from Transgenomic called a WAVE® System was used for analysis (Figure 4.1). The central component of the WAVE® System is the separation cartridge which DNA binds to and elutes depending on the presence of a SNP or the length of the DNA fragment.

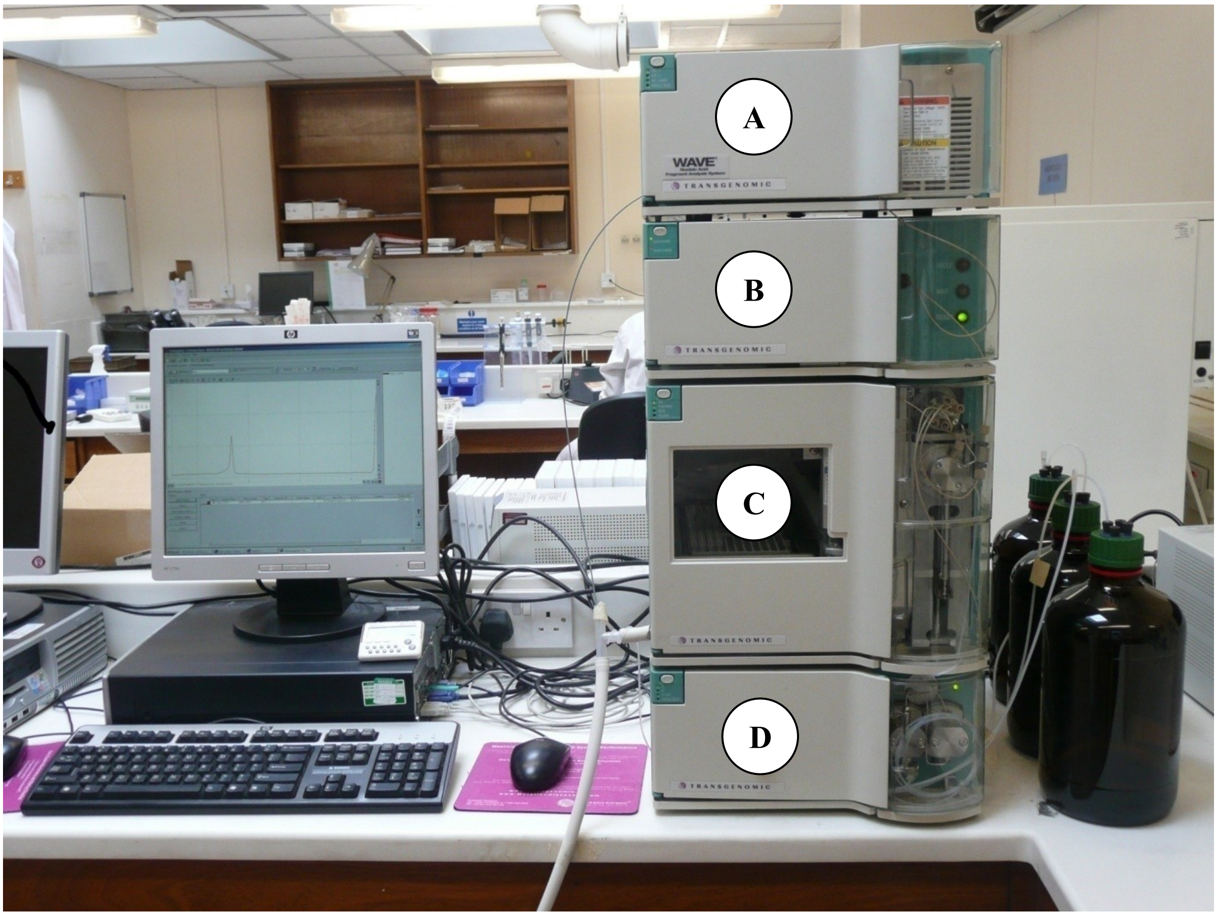


Figure 4.1. Transgenomic WAVE® dHPLC system.

The four sub-units are: UV detector (A); Oven containing analysis cartridge (B); 96-well three plate autosampler (C); and pump unit (D).

The WAVE® System DNASep separation cartridge is packed with a non-porous matrix consisting of polystyrene-divinylbenzene (PS-DVB) copolymer beads. The beads are alkylated with C-18 chains that form single C-C bonds. By nature, the beads within the cartridge are electrostatically neutral and hydrophobic, and do not readily react with nucleic acids. Triethylammonium acetate (TEAA) is an ion pairing reagent that acts as a 'bridging molecule' to aid in the adsorption of nucleic acids to the beads. The ion-pairing reagent is both hydrophobic and positively charged, and interacts with the negative charge of the anionic phosphate backbone of the nucleic acid, while the hydrophobic groups of the TEAA interact with the hydrophobic C-18 chains on the PS-DVB beads. This ion-pairing reagent serves as a bridge between the nucleic acid and the cartridge matrix as shown in Figure 4.2.

The cartridge matrix described above is referred to as the stationary phase. The mobile phase consists of the combination of buffers that elute the DNA off from the cartridge. A 0.1M solution of TEAA mixed with 25% acetonitrile is used to elute the DNA. As increasing concentrations of acetonitrile flow across the cartridge matrix, the hydrophobic interaction between the cartridge and the DNA/TEAA is broken which causes elution of the DNA.

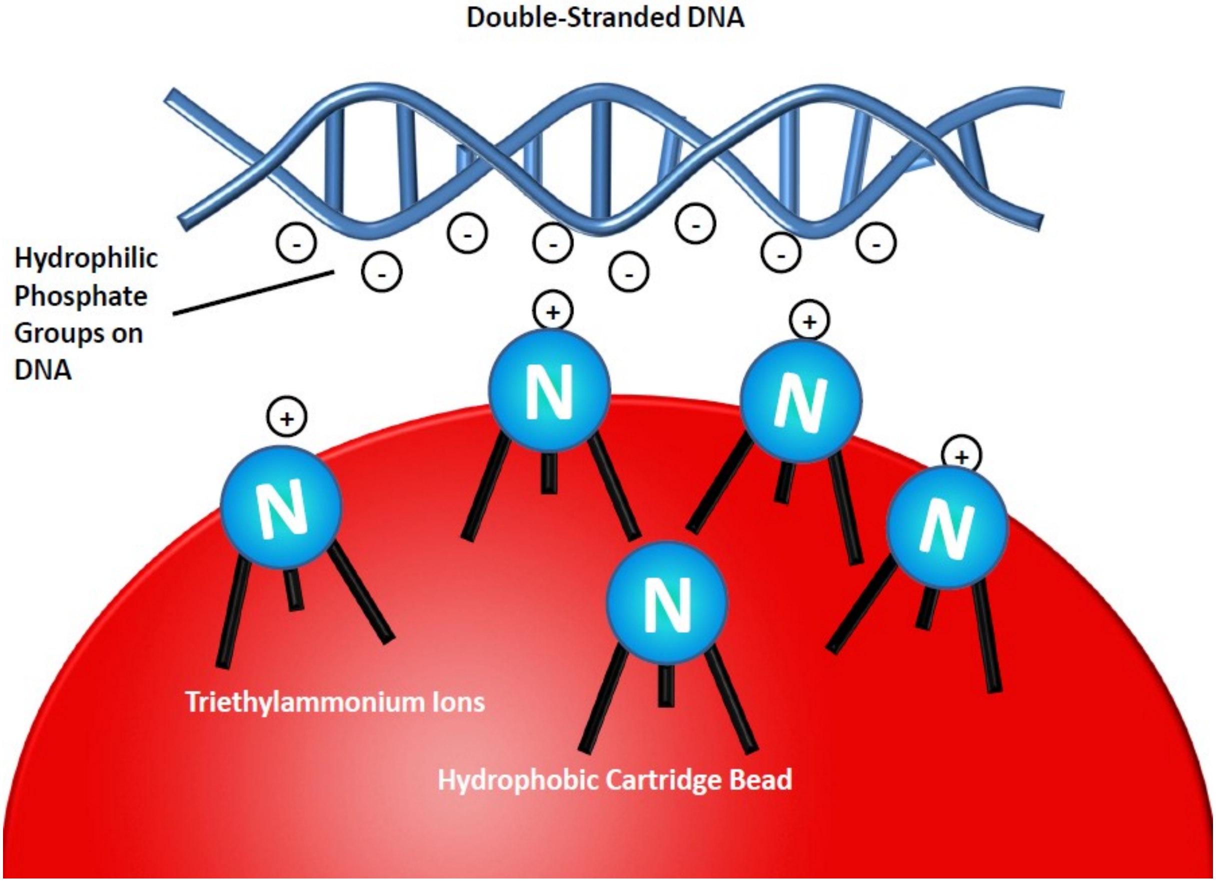

Double-Stranded DNA
Hydrophilic Phosphate Groups on DNA
-
+
N
Triethylammonium Ions
Hydrophobic Cartridge Bead

Figure 4.2. dHPLC Separation Cartridge Surface Chemistry.

In the separation cartridge, TEAA forms a positively charged triethylammonium ion (TEA+) that has both hydrophobic and hydrophilic ends. The DNASep cartridge contains beads that are hydrophobic. The positively charged portion of the TEA+ forms an association with the negatively charged phosphate backbone of the DNA creating a hydrophobic "fur" on the fragment. This entire hydrophobic entity then behaves as would a typical hydrophobic molecule and is attracted to the hydrophobic beads.

4.1.5. Denaturing conditions: Separation based on presence or absence of a SNP

The majority of studies have used dHPLC for SNP genotyping. PCR fragments containing the region of interest in the wild-type control containing no mutations and the sample of interest are amplified in separate reactions (Figure 4.3). After completion of the PCR, the amplicons can be quantified and then mixed together and denatured into single-strand forms by heating to 95°C and then annealed by cooling to 4°C. Once denatured, each strand can either re-anneal with its own entirely complementary strand of DNA (homoduplex), or it can anneal with the other type of DNA molecular present which will result in a DNA binding mismatch if there is a SNP present (heteroduplex). Each PCR fragment requires a specific analysis temperature and buffer proportions which are designed using WAVE® System software. The annealed amplicons are applied to the dHPLC cartridge and an increasing gradient of acetonitrile is applied which reduces the bridging affinity of TEAA and the amplicons are eluted from the cartridge. Heteroduplexes, containing mismatched SNPs, elute off before homoduplexes. The eluted amplicons then pass through a UV absorbance detector which measures absorbance over the time of each injection and the WAVE® System software constructs a chromatogram. If no SNP is present, there is one species of DNA (wild-type and sample which both have the same base pairing at the SNP site). These possess indistinguishable cartridge affinities so they elute at the same retention time which results in a single chromatogram peak. If a SNP is present, there are four species of DNA present (i.e. AC and GT heteroduplexes and AT and GC homoduplexes), which can potentially elute as four separate chromatogram peaks. The first two peaks contain the two heteroduplexes and the latter two peaks contain the two homoduplexes.

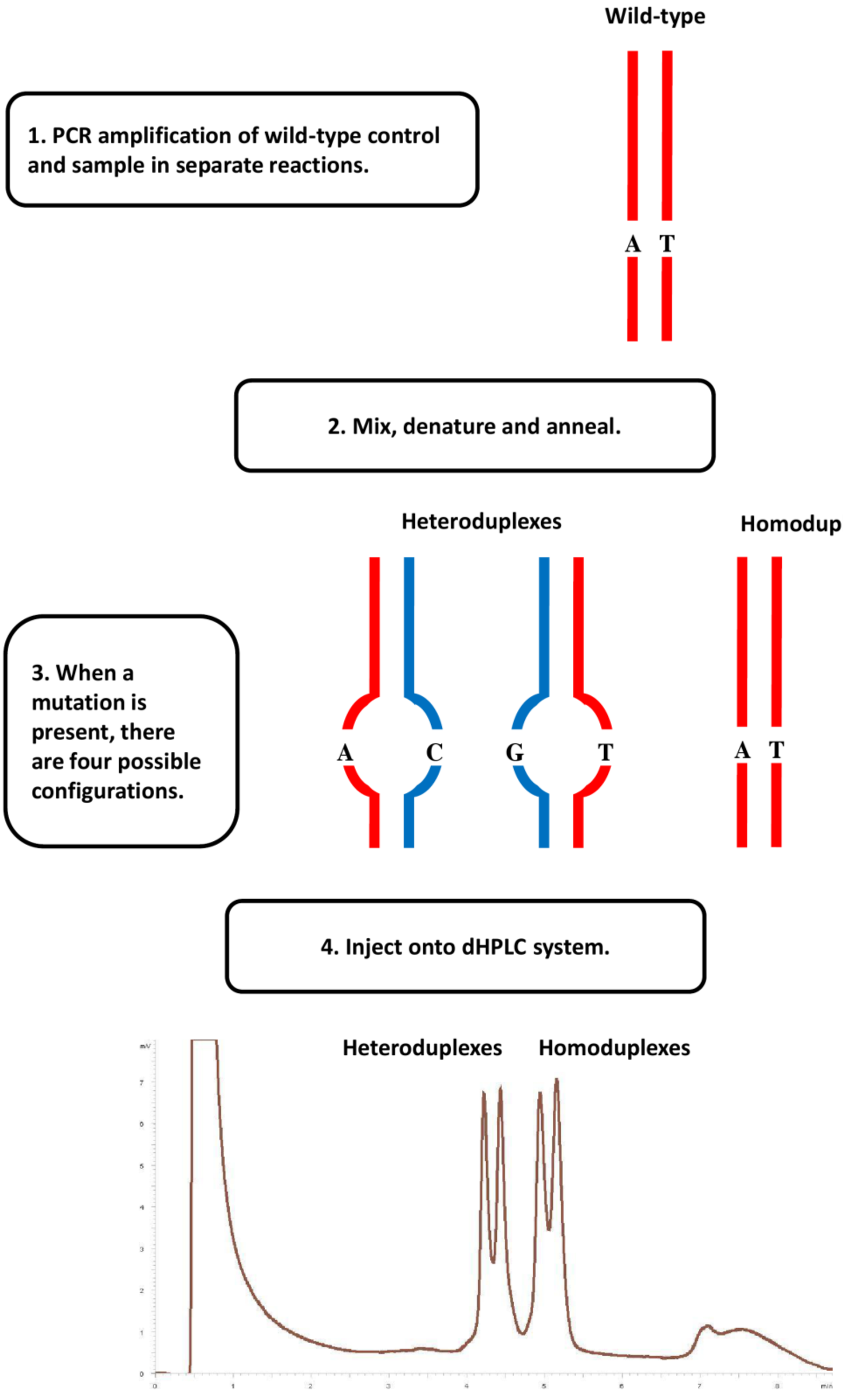

Wild-type
Mutant
1. PCR amplification of wild-type control and sample in separate reactions.
A T
G C
2. Mix, denature and anneal.
Heteroduplexes
Homoduplexes
3. When a mutation is present, there are four possible configurations.
A
C
G
T
A T
G C
4. Inject onto dHPLC system.
Heteroduplexes
Homoduplexes
mV
7
6
5
4
3
2
1
0
0
1
2
3
4
5
6
7
8
min

Figure 4.3. Mutation Detection by dHPLC.

Wild and test or mutant type alleles are amplified separately, mixed, heated and then cooled to form homoduplexes or heteroduplexes if a mutation is present. As the acetonitrile concentration increases the bridging capabilities of the TEA+ ions decrease and the DNA fragments are released from the cartridge. Heteroduplexes elute off of the cartridge first followed by the homoduplexes and pass through a UV detector which detects the absorbance over time and is recorded as a chromatogram. If no mutation is present, all of the homoduplex DNA fragments elute off of the cartridge at the same time producing a single peak on the chromatogram. If a mutation is present then two to four peaks will be visible.

4.1.6. Non-denaturing conditions: Separation based on size

dHPLC can be used in a non-denaturing form (non-dHPLC) by analysing double-stranded PCR amplicons at 50°C. This essentially changes the method of separation analysis from dHPLC analysis based on individual base composition at SNP sites to non-dHPLC analysis based on the relationship between amplicon length and the number of bonds between each amplicon and the cartridge matrix. Each phosphate group in the backbone of each DNA molecule binds to an available TEAA molecule which in turn binds to the cartridge matrix. Shorter amplicons have fewer phosphates to bind to the cartridge matrix and longer amplicons possess more phosphates that can bind to the cartridge matrix. For separation based on size, each amplicon of interest is generated by PCR and injected onto the WAVE® System. Within each run, a known molecular standard can also be analysed which can be used to generate a standard curve against which samples can be compared. Specific analysis conditions are also designed using the WAVE® System software which are not specific to each individually designed amplicon sequence but based on the size range of DNA fragments analysed. An increasing buffer gradient is used similar to SNP detection but as the concentration of acetonitrile across the cartridge rises the fragments elute, smallest to largest (Figure 4.4) (Devaney *et al.*, 2000).

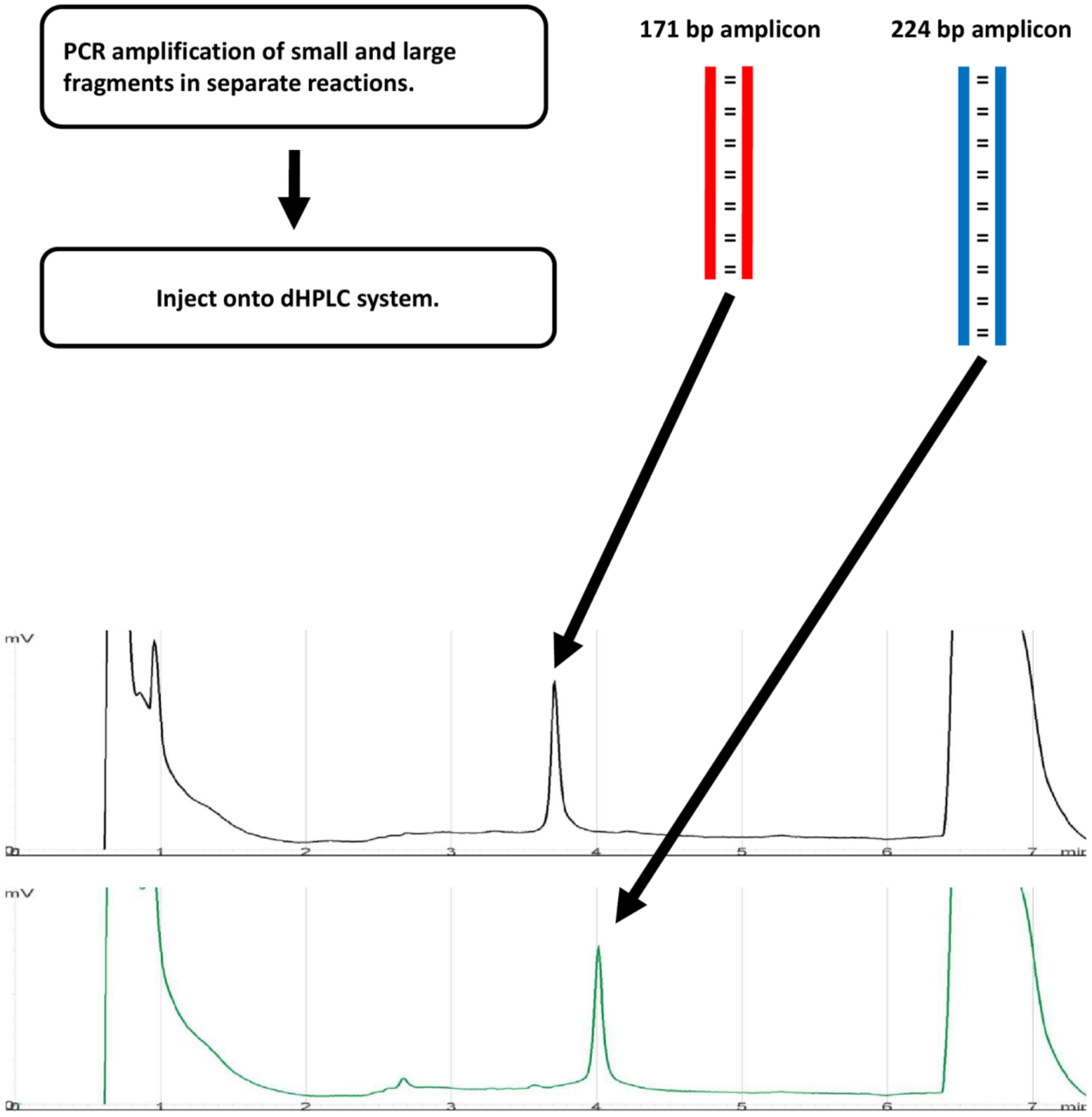

PCR amplification of small and large fragments in separate reactions.
Inject onto dHPLC system.
171 bp amplicon
224 bp amplicon
mV
0
1
2
3
4
5
6
7
min
mV
0
1
2
3
4
5
6
7
min

Figure 4.4. PCR amplicon sizing by non-denaturing HPLC.

PCR amplicons are amplified and then injected onto the sizing cartridge. Shorter fragments bind with less affinity and are eluted quicker than larger fragments. Two amplicons from MIRU-10 are shown with sizes of 171 bp and 224 bp which equate to two and three repeats respectively.

4.2. AIMS

MIRU-VNTR typing has been previously automated using capillary electrophoresis (Supply *et al.*, 2000). In this study we evaluated the utility of non-dHPLC in MIRU-VNTR typing because it has proved to be one of the most cost-effective methods for DNA mutation analysis in prokaryotic biology.

This study describes the initial validation of a WAVE® System for MIRU-VNTR typing of *M. tuberculosis* strains with use of 12 separate injections.

4.3. METHODS

Primary specimens were processed as described in Section 2.3, cultured using liquid media (Section 2.3.2), DNA was extracted (Section 2.3.3) and isolates were identified as *M. tuberculosis* using the HAIN GenoType MTBC assay (Section 2.4).

4.3.1. Strain collection

Seventy *M. tuberculosis* strains were selected from the MRCM culture collection. These strains had been previously characterized by MIRU-VNTR typing with the primers described by Philip Supply on agarose gels (Section 2.6.1). A dendrogram of the obtained results was constructed using the categorical coefficient algorithm and produced via UPGMA in BioNumerics. Twenty isolates were selected to represent the entire range of repeats observed at each locus from previous typing. The remaining 50 isolates were randomly selected from the clinical isolates typed at the Birmingham Regional Centre for Mycobacteriology. This collection represented the entire spectrum of MIRU locus alleles identified from clinical strain typing in the Midlands in 2003.

4.3.2. Allelic range of the Birmingham strain collection

The allele range of the Birmingham strain collection was compared to a global collection analysed in a previous study on MIRU-VNTR typing of 31 strains from seven countries (Supply *et al.*, 2000) in terms of the range of alleles identified at each of the 12 MIRU loci.

4.3.3. MIRU-VNTR typing using non-dHPLC

Oligonucleotides were designed for non-dHPLC analysis so that the length of flanking regions upstream and downstream of the actual VNTRs was as short as possible to enable analysis on a WAVE® system (Section 2.7.1). The PCR mastermix and thermal cycling conditions are described in Section 2.7.4.

4.3.4. MIRU-VNTR typing using non-dHPLC: Analysis gradient design

Each sample injected onto a WAVE® system undergoes an increasing gradient of acetonitrile which elutes bound DNA at a specific timepoint or "retention time". The exact increase in buffer percentage can be designed using Transgenomic WAVE® system software which enables the optimal separation of DNA molecules. For this study, a "DNA multiple fragments" gradient was chosen as the basis for development with parameters altered included the rate of analysis (minutes per 100 bp) and size range (bp) evaluated (Section 2.7.2).

4.3.5. MIRU-VNTR typing using non-dHPLC: Calculation of a standard curve

Data generated by a WAVE® system contains two data fields that were used for this study: retention time and peak height. Chromatogram peaks with a height ≥1 mV and within the range of the molecular standard were exported as a Comma Separate Variable file and imported for analysis into a laboratory-designed Microsoft Office Excel calculation file. The retention time of the samples was compared to the molecular standard and the size (bp) and number of repeats was calculated (Section 2.7.5).

4.3.6. Concordance between MIRU-VNTR data obtained from agarose gel electrophoresis and WAVE® System analysis.

MIRU-VNTR profiles obtained from the two methods were analysed in a blinded manner by a single operator and were then directly compared to calculate concordance.

4.3.7. Analysis of the performance of the WAVE® System

Within the Birmingham strain collection, the lowest and highest VNTR allele observed was identified along with the median allele for each MIRU locus. Replicate injections were carried out to ensure that each identified allele had at least 10 replicate data points for analysis. Statistical criteria analysed were the observed calculated size with a 95% confidence interval and the mean variance when compared to the expected size.

4.4. RESULTS

4.4.1. Allelic range of the Birmingham strain collection.

Table 4.1 is a comparison of the allelic range of the Birmingham strain collection with those obtained in a global strain collection (Supply *et al.*, 2000). For nine of the MIRU loci, isolates from a global collection displayed a greater range of alleles than in our collection. The highest allele observed in five loci in the global collection was only one repeat higher than the Birmingham strain collection with four loci containing two or more repeats in the global collection. For two MIRU loci (MIRU-02 and MIRU-20) the allele range was the same for the two strain collections. For one locus (MIRU 26) the collection from Birmingham Regional Centre for Mycobacteriology exhibited a greater range of alleles than the global collection.

Locus	**Allele Range**	
	West Midlands	**Global**
2	1-3	1-3
4	1-5	1-9
10	2-7	2-8
16	1-6	2-8
20	1-2	1-2
23	1-7	1-11
24	1-2	1-6
26	1-9	1-6
27	2-4	2-5
31	2-5	1-6
39	1-3	1-4
40	1-7	1-8

Table 4.1. Allelic range of typed strains from West Midlands compared to that of global strains in 2003.

4.4.2. MIRU-VNTR typing by agarose gel electrophoresis

In the Birmingham strain collection of 70 *M. tuberculosis* complex isolates, there were 56 distinct MIRU-VNTR profiles with a total of 26 isolates indistinguishable by MIRU-VNTR in 12 clusters. Forty-four isolates possessed a unique MIRU-VNTR profile compared to other strains in the collection. The largest cluster of four isolates (223325173533) represented a group of four epidemiologically linked patients from one sending location. Of the other 11 clusters each containing a pair of isolates, only two clusters contained isolates received from the same sending laboratory (Figure 4.5).

4.4.3. MIRU-VNTR typing using non-dHPLC

An example chromatogram of PCR fragments amplified from different strains is shown in Figure 4.6.

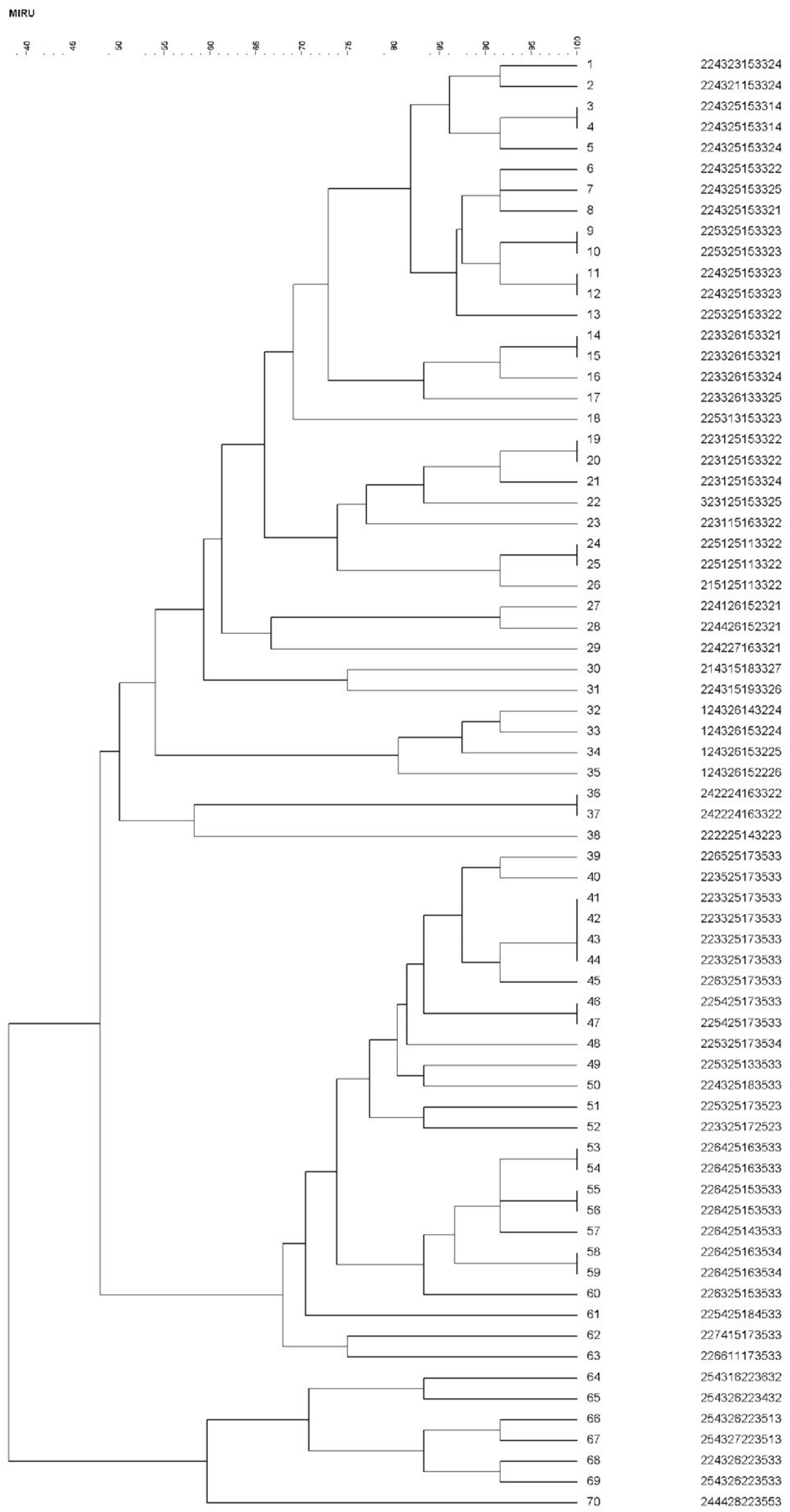


Figure 4.5. MIRU typing results obtained for each of the 70 *M. tuberculosis* strains analysed.

The dendrogram was constructed with use of the categorical coefficient algorithm and produced via the Unweighted Pair Group Method using Arithmetic Average.

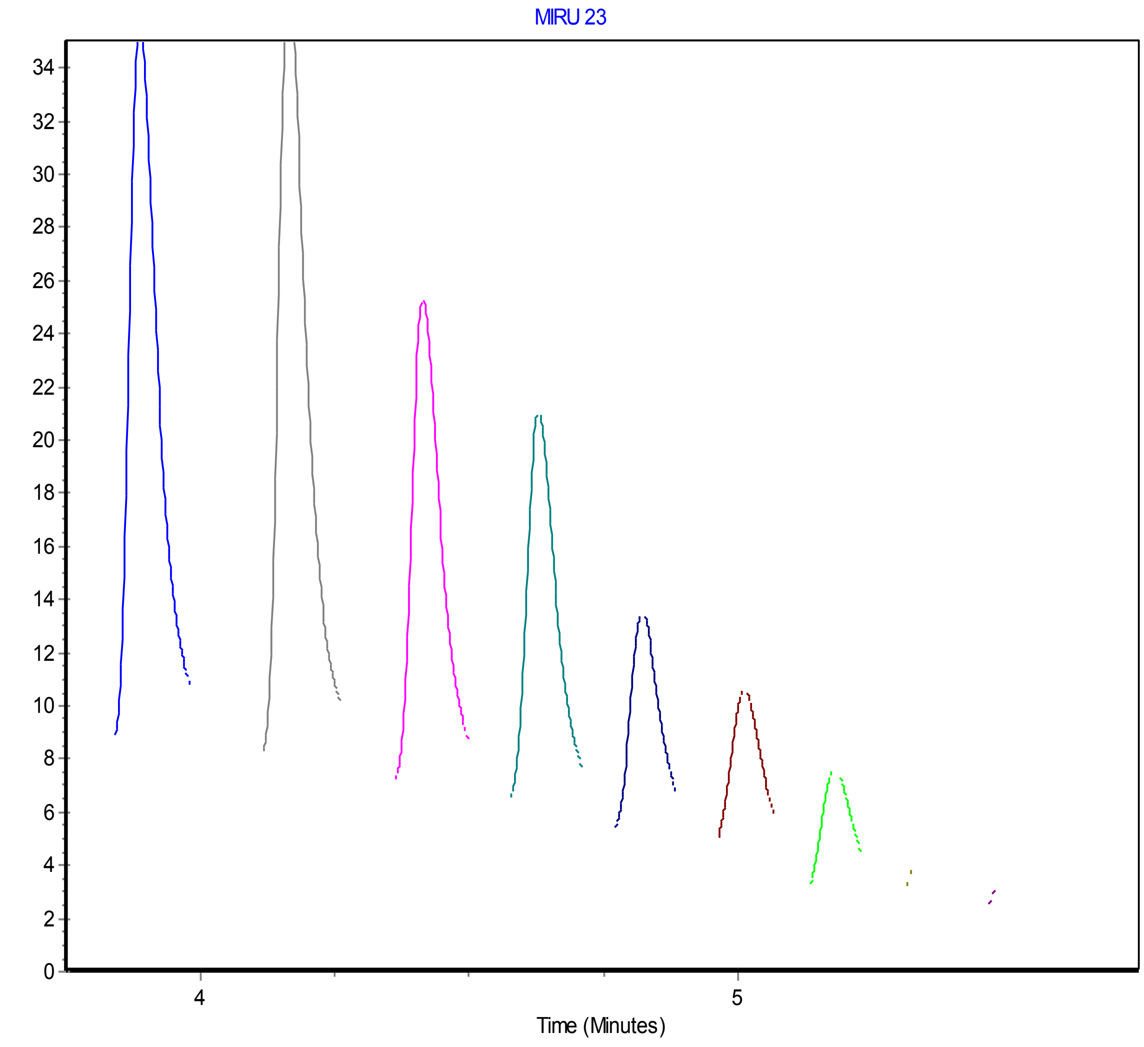


Figure 4.6 MIRU-23 alleles analysed by non-dHPLC.

The MIRU-23 locus in ten different *M. tuberculosis* strains was amplified by PCR and separated by non-dHPLC. Each strain had a different allele value which ranged from one (furthest left) to nine (furthest right) repeats.

4.4.4. MIRU-VNTR typing using non-dHPLC: Primer design

Primer flanking regions either upstream and downstream for the original primers used were an average 420 bp in length (range 44-565 bp) whereas for the oligonucleotides designed for sizing by non-dHPLC the average length of the primer flanking regions was 98 bp (range 39-213 bp). This meant that the highest allele and largest PCR amplicon in this study (7 repeats at MIRU-23) was reduced in size from 863 bp to 445 bp.

4.4.5. MIRU-VNTR typing using non-dHPLC: Analysis gradient design

From initial experiments, it was identified that for MIRU-VNTR typing using non-dHPLC to be a realistic proposition in terms of high-throughput genotyping, the gradient parameters would have to be set to provide the most rapid injection analysis gradient possible. Therefore, the evaluated rate of analysis was set to 0.5 minutes per 100 bp and the size range from 20-900 bp. This size range would cover all of the alleles in the Birmingham and global collection.

4.4.6. MIRU-VNTR typing using non-dHPLC: Calculation of a standard curve

From an initial evaluation, it was found that a MIRU allele ladder constructed from 10 PCR products (1-10 repeats) from MIRU-23 in 10 different strains provided more accurate VNTR allele assignation than a commercially available 50 bp standard. With use of this DNA standard, the effects of different sequence composition between the molecular standard and analysed amplicons were reduced.

4.4.7. Concordance between MIRU-VNTR Data obtained from agarose gel electrophoresis and WAVE® System analysis

Figure 4.5 displays the 70 strains analyzed by agarose gel electrophoresis and the WAVE® System as single amplifications. A concordance level of 100% between MIRU-VNTR profiles generated by agarose gel electrophoresis and by the non-dHPLC method was obtained when results from the two methods were compared.

4.4.8. Performance of MIRU-VNTR typing carried out using non-dHPLC.

Table 4.2 displays the statistical analysis of peaks obtained over multiple runs when the collection of 70 strains was typed using the WAVE® System. The minimum, median, and maximum numbers of repeats observed in clinical strains previously typed in the West Midlands were statistically analysed.

Locus	Value for product containing expected no. of repeats							
	Expected no. of repeats	No. of results	Expected (bp)	Observed (bp)			95% CI (bp)	Mean variance (bp)
				Min.	Mean	Max.		
Minimum								
2	1	10	168	155	160	164	158-162	-8
4	1	11	196	182	195	211	189-201	-1
10	2	14	195	183	190	198	187-193	-5
16	1	39	107	95	103	108	102-104	-4
20	1	12	156	144	150	156	148-152	-6
23	1	10	134	121	133	138	132-138	-1
24	1	59	179	166	173	178	172-174	-6
26	1	17	113	100	108	115	106-110	-5
27	2	10	210	193	201	208	198-204	-9
31	2	13	183	175	177	183	175-179	-6
39	1	17	187	178	182	187	181-183	-5
40	1	16	206	197	200	203	199-201	-6
Median								
2	2	67	221	209	216	222	215-217	-5
4	2	59	273	261	270	275	269-271	-3
10	4	29	301	287	293	301	292-294	-8
16	3	39	213	201	209	214	208-210	-4
20	2	63	233	221	228	235	227-229	-5
23	5	45	346	338	347	359	345-349	1
24	1	59	179	166	173	178	172-174	-6
26	5	30	317	305	316	323	315-317	-1
27	3	64	263	247	255	261	254-256	-8
31	3	37	236	225	231	235	230-232	-5
39	2	38	240	228	233	237	232-234	-7
40	3	31	314	301	308	314	307-309	-6
Maximum								
2	2	67	221	209	216	222	215-217	-5
4	5	10	504	481	490	503	485-495	-14
10	7	10	460	439	450	460	446-454	-10
16	6	14	372	365	368	374	367-369	-4
20	2	63	233	221	228	235	227-229	-5
23	7	18	452	441	451	465	448-454	-1
24	2	11	233	226	228	231	227-229	-5
26	9	14	521	512	517	522	515-519	-4
27	4	10	316	303	306	308	305-307	-10
31	5	28	342	332	340	347	338-342	-2
39	3	27	293	281	286	290	285-287	-7
40	7	10	530	505	515	524	512-518	-15

Table 4.2. Analysis of product sizes calculated from WAVE® system DNA fragment analysis chromatograms.

4.4.9. Performance of dHPLC in typing alleles with low numbers of repeats

For the products containing the lowest number of repeats observed at each locus (all 12 loci concatenated together: 112111112211), the 95% confidence interval (CI) ranged from 2-12 bp (0.04 to 0.16 repeats) with a mean interval of 5 bp (0.07 repeats). The variance of the mean calculated sizes from the expected product sizes ranged from 1 to 9 bp (0.01 to 0.16 repeats) less than the expected product size, with an average of -5 bp (-0.09 repeats) across the 12 MIRU loci.

4.4.10. Performance of dHPLC in typing alleles with median numbers of repeats

The concatenated median number of repeats for each locus was 224325153323. For these PCR products, the 95% CI varied from 2-4 bp (0.02 to 0.06 repeats) with an average interval of 2 bp (0.04 repeats). The mean of the calculated product sizes varied from the expected size by -1 to -8 bp (-0.01 to -0.13 repeats) with an average of -5 bp (-0.09 repeats) for all 12 MIRU loci.

4.4.11. Performance of dHPLC in typing alleles with high numbers of repeats

For PCR products analysed from isolates that possessed the maximum number of repeats observed at each locus in clinical strains (257627294537), the 95% CI ranged from 2-10 bp (0.02 to 0.14 repeats), with a mean interval of 4 bp (0.07 repeats). The variance of the mean calculated size from the expected product size ranged from -1 to -14 bp (-0.02 to -0.27 repeats) with an average variance from the expected size of -7 bp (-0.12 repeats).

4.4.12. Performance of non-dHPLC for all 840 PCR amplicons in this study.

Of 840 PCR fragments analyzed, 828 (99%) were sized accurately within a repeat integer range of ±0.25 repeats. For all 840 PCR fragments analyzed within this study, the WAVE® System exhibited a mean accuracy of +4.8 bp or +0.09 repeats. The 12 PCR products that were not within this range had an average variance of -0.29 repeats from the expected repeat number. Upon repeat analysis, all of the 12 PCR products that did not initially group within this interval were sized accurately to within ±0.25 repeats.

4.5. DISCUSSION

4.5.1. Allelic range of the Birmingham strain collection

The allele range of the Birmingham collection of 70 strains used in this study equates to approximately 80% of the total allelic range observed in a global collection of strains from seven countries analysed by MIRU-VNTR typing (Supply *et al.*, 2000). Therefore, the range of alleles tested is notable for a study based in one geographical location.

4.5.2. Non-dHPLC can be used for automated MIRU-VNTR typing

The results of this study showed that non-dHPLC using a WAVE® System produced accurate and reproducible sizing of MIRU-VNTR amplicons when utilized for single PCR MIRU-VNTR typing of *M. tuberculosis* strains. The highest number of repeats analyzed in this study was nine repeats (521 bp) for MIRU-VNTR locus 26. For nine repeats at this locus, the total range of calculated amplicon sizes was 512 to 522 bp, the 95% CI was 515 to 519 bp, and the mean variance of calculated product sizes from expected product size was -4 bp. The largest product size analyzed was 530 bp for MIRU 40 (seven repeats). For this locus the calculated amplicon size ranged from 505 to 524 bp, with a 95% CI of 512 to 518 bp, and a mean variance from calculated to expected amplicon size of -15 bp. The relatively narrow 95% CI values indicate that the majority of PCR products were grouped together. When sizing amplicons for tandem repeat typing, very high levels of accuracy of sizing are not necessary as the analysed MIRU-VNTR repeats are all longer than 50 bp. Therefore, it is necessary only to have sufficient accuracy to unambiguously determine the number of repeat sequences at each locus. Combining the 281 amplicons analyzed for the maximum number of repeats category, the mean variance for all samples compared to the expected amplicon size was -6 bp. For quality control purposes, certain ranges of allowable repeat numbers must be implemented. For single amplicon analysis, a "buffer zone" of 50% of the repeat size between predicted PCR fragment sizes was used. This meant that any PCR product calculated as within ±0.25 repeats when compared to a repeat integer was accepted. On a WAVE® System 828 of 840 (99%) of single PCR fragments analysed were within this interval. The 12 PCR products that were not within this range had an average variance of -0.29 repeats (i.e. only -0.04 repeats outside the limit) from the expected repeat number. This variation probably occurred as a result of systematic calculation because retention times were recorded to two decimal places. Repeat values were then calculated and rounded to two decimal places before rounding up or down to a repeat integer value.

Within this study, it was found that an allelic standard using 10 alleles in MIRU-23 provided more accurate allele calling than a 50 bp standard. This is most likely due to differences in GC content between *M. tuberculosis* amplicons and the 50 bp standard. For evaluation of any automated system, the effect of GC content should be taken into account when assessing performance. A different strategy to improve accuracy is the application of an offset to each calculated fragment size and allele value (Allix *et al.*, 2004).

4.5.3. Comparison of non-dHPLC to Capillary Electrophoresis

In a previous study (Supply *et al.*, 2001), sizing of MIRU-VNTR PCR amplicons using a DNA sequencer was found to be reproducible, with within-run and between-run average precisions of ±0.5 and ±0.6 bp, respectively. Mean errors for the sizing accuracy were 1.1 bp for fragments below 500 bp and 0.8 bp for fragments from 500 to up to 971 bp. The WAVE®

microbial analysis system exhibits a mean error of 4.8 bp for all PCR fragments analyzed within this study.

4.5.4. Other microfluidic technologies

At least two other equivalent microfluidic methods for automating MIRU-VNTR analysis in *M. tuberculosis* and one method for automated VNTR analysis in methicillin-resistant *S. aureus* have been subsequently evaluated and suggested as potential methods for high-throughput prospective clinical diagnostic purposes (Cooksey *et al.*, 2003;Mallard *et al.*, 2009;McMurray *et al.*, 2010).

4.5.5. Cost and logistics of MIRU-VNTR typing using the WAVE® System

To perform MIRU-VNTR typing with 12 single reactions, post-amplification analysis by non-dHPLC on a WAVE® System would cost $7 per isolate. After PCR amplification for non-dHPLC MIRU-VNTR typing, the next step is straightforward uncapping and loading of tubes into the autosampler unit of the WAVE® System. No intermediary manipulation of the PCR product such as addition of loading buffer, denaturation, and prerunning, which is required on a 96-well ABI 377 Capillary Electrophoresis System (Bio-Whittaker Molecular Applications, Rockland, Maine), is required (Supply *et al.*, 2000). Non-dHPLC involves single injection and single analysis of each MIRU-VNTR amplicon without multiplex analysis. We have estimated the annual capacity of a single WAVE® system to be >2,000 strains for MIRU-VNTR typing. Multiplex analysis using capillary electrophoresis can increase capacity and through-put but signal noise within and between capillaries caused by fluorescence “cross-talk” can be problematic when assigning alleles (Orita *et al.*, 1989;Thorne *et al.*, 2007b).

4.5.6. Primer design was optimised for non-dHPLC analysis

The total size range of amplicons amplified by our novel set of primers was 107 to 521 bp (1-9 repeats). This optimized range further reduced the analysis time to 7.4 minutes per injection, as the primer-binding flanking regions are relatively small (the largest is MIRU-40 at 152 bp).

This study is the first to report the successful use of non-dHPLC for screening for variations in the number of MIRU-VNTRs in mycobacterial DNA. The fragments of various retention times detected by agarose gel analysis could be rapidly and reproducibly identified in single PCR products by non-dHPLC, and produced specific DNA WAVE® patterns that correlated with MIRU-VNTR genotypes.

4.6. CONCLUSIONS

Non-dHPLC analysis was demonstrated to be a rapid, low-labour input method for the detection and analysis of PCR fragments containing various numbers of tandem repeats in *M. tuberculosis*. The development of automated MIRU-VNTR typing using non-dHPLC will enable the future implementation of prospective universal DNA fingerprinting of all *M. tuberculosis* strains within the Midlands which will greatly increase our knowledge and understanding of the epidemiology and transmission of *M. tuberculosis*. Non-dHPLC using a WAVE® System has been applied to genotype >8,000 *M. tuberculosis* strains between 2003 and 2010.

5. GLOBAL ORIGIN OF *M. TUBERCULOSIS* IN THE MIDLANDS, UK.

5.1. INTRODUCTION

Knowledge and understanding of the transmission dynamics of *M. tuberculosis* have been improved by the development of rapid molecular techniques that are being more extensively applied such as MIRU-VNTR typing (Evans *et al.*, 2004;Mazars *et al.*, 2001). On a global scale, the distribution of major *M. tuberculosis* phylogenetic lineages has been elucidated as a result of the application of various molecular typing tools with varying rates of evolution such as: the DR locus (Brudey *et al.*, 2006); SNPs (Sreevatsan *et al.*, 1997); MIRU-VNTRs (Mazars *et al.*, 2001); LSPs (Fleischmann *et al.*, 2002); and IS*6110*-RFLP (Kremer *et al.*, 2004).

It has been hypothesized that *M. tuberculosis* emerged in East Africa (Gutierrez *et al.*, 2005;Hershberg *et al.*, 2008;Wirth *et al.*, 2008), with dispersal and regional evolution in different continents that has generated distinct global populations of *M. tuberculosis* (Baker *et al.*, 2004). A study in San Francisco also found that these global lineages can exhibit specific associations between the global origin of a TB patient and the lineage of the associated strain (Figure 5.1) (Gagneux *et al.*, 2006). A similar replicate study in Montreal analyzed the robustness and reproducibility of the LSP methodology for global phylogenetic analysis in *M. tuberculosis* (Reed *et al.*, 2009). Comparison between the San Francisco and Montreal populations revealed that the initial associations detected between bacterial genotype and patient ethnic origins were conserved, even though the two patient populations differed in the distribution of their geographic origins.

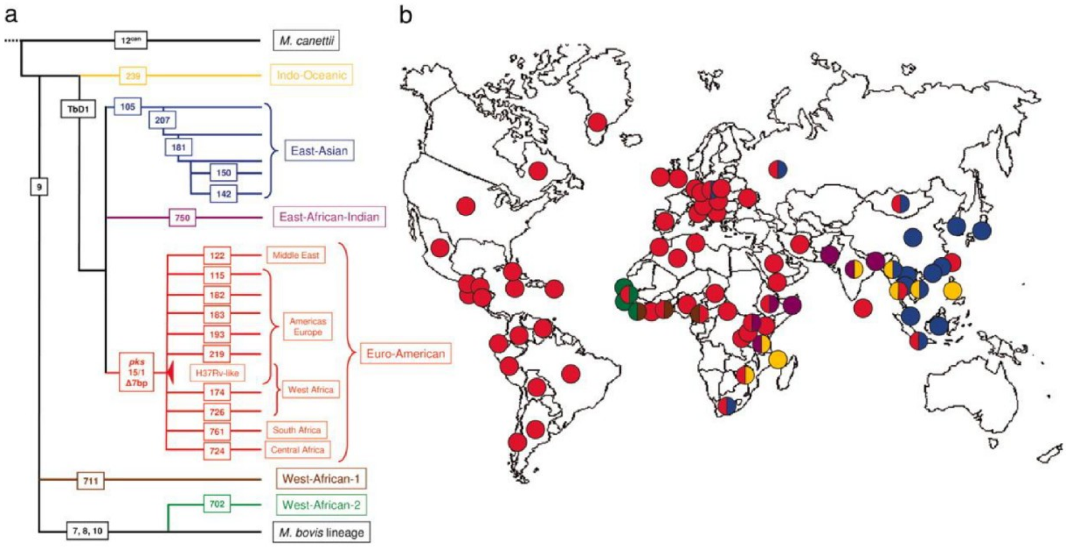


Figure 5.1 Global population structure and geographical distribution of *M. tuberculosis*.

LSPs define a global phylogeny for *M. tuberculosis*. The names of the lineage-defining LSPs or regions of difference are shown in rectangles. The geographic regions associated with specific lineages are indicated. (b) The six main lineages of *M. tuberculosis* are geographically structured. Each dot corresponds to one of 80 countries represented in the global strain collection analysed. The colours of the dots relate to the six main lineages defined and indicate the dominant lineage(s) in the respective countries (Gagneux et al., 2006).

The Midlands region of the UK contains a population of ~9.8 million people, and in 2009 had a rate of 15.8 TB cases per 100,000 population with 1,564 clinical cases. In 1,429 individuals where the place of birth was known, most individuals diagnosed with TB in the Midlands were born outside of the UK (966/1,429, 68%). Southern Asia was the most common region (609/1,429, 43%), and India was the most frequent country of birth (352/1,429, 25%) (Health Protection Agency, 2010). The HPA MRCM has typed all received *M. tuberculosis* isolates in the Midlands since January 2004 by MIRU-VNTR typing with the aim of prospectively detecting unsuspected transmission events and informing public health control efforts to reduce the incidence of *M. tuberculosis* (Evans *et al.*, 2007).

This chapter combines molecular epidemiological data with data obtained from a novel software programme (Origins) that assigns a CEL group based on the combination of given and family names. The Origins software program was developed by analyzing reference records from multiple countries (Australia, France, India, Ireland, Italy, Netherlands, Norway, Romania, Spain, Sweden, the USA, and the UK) to formulate algorithms based on the cultural, ethnic, and linguistic factors found to be associated with the given and family name. This has resulted in a reference database now containing 1,600,000 family names and 600,000 given names. From this database,>200 different CEL groups have been constructed which are applicable across the world.

In the UK, the non-UK population CEL groups were validated by comparison to the UK Consumer Dynamics Database (Experian, Nottingham) which essentially contains the UK electoral register and population groups used in the UK census. This comparison showed that the proportion of names in each main group generated by Origins was broadly similar to the proportions of the UK census population in each group. The proportion of individuals that Origins defined as originating from Southern Asia was broadly similar to that present in the UK census with 2-11% differences in proportions within subgroups between the two data sources. Population groups who originated from Eastern Asia were slightly underrepresented in Origins when compared to the UK census. Population groups from Africa had very similar representation in both datasets (Webber, 2007).

Origins software gives much greater accuracy and coverage of origin assignment than previous programs such as Nam Pehchan, Sangram and the Dictionary of American Family Names (Webber, 2007). The first use of the Origins software program in healthcare was the successful identification of the Polish population group sought clinical advice and treatment from the NHS in the UK (Leaman *et al.*, 2006).

5.2. AIMS AND HYPOTHESIS

The aim of this study was to combine mycobacterial fingerprinting data and patient origin as assigned by Origins software to relate the occurrence of major global *M. tuberculosis* lineages in populations that originate from around the world. Combining data obtained from universal typing and associated cultural and social links identified by Origins provides the potential for a deeper understanding of the causes for distribution of prevalent strains in specific population groups. We hypothesise that since patients originating from Southern Asia form the largest immigrant group in the Midlands that strains that are present in Southern Asia will also be present in the Midlands.

5.3. METHODS

5.3.1. Study population

Nonduplicate initial *M. tuberculosis* complex isolates (n = 5,731) were referred from the Midlands region (Figure 2.1) of the United Kingdom (population 9.5 million) to our laboratory between 1st January 2004 and 31st December 2009.

5.3.2. Mycobacterial culture, identification, and MIRU-VNTR genotyping

Primary specimens were processed as described in Section 2.3, cultured using liquid media (Section 2.3.2), DNA was extracted (Section (2.3.3) and isolates were identified as *M. tuberculosis* using the HAIN GenoType MTBC assay (Section 2.4). MIRU-VNTR typing of the five ETR (ETR-A to -E) and 10 MIRU loci (MIRU-02, -10, -16, -20, -23, -24, -26, -27, -39, and -40) was carried out using a Transgenomic WAVE® System (Section 2.7). MIRU-04 and MIRU-31 were removed from the 12 MIRU loci and analysed as ETR-D and ETR-E in the five ETR loci.

5.3.3. Assignation of CEL groups

Origins software (Experian, Nottingham, UK) was used to analyze given and family names and assign a CEL group to each patient (Section 2.9.2). The given and family name of each of the 5,731 patients were entered into Origins to obtain a CEL group, which was then assigned a continent based on the United Nations Standard Country and Area Codes Classification Scheme (United Nations Statistics Division., 2009). This classification places the British Isles in Northern Europe and India, Pakistan and Bangladesh in Southern Asia.

5.3.4. Available online databases of *M. tuberculosis* complex strains

MIRU-VNTR data was compared to MIRU-VNTR*plus* (See Methods Section 2.9.7) to assign *M. tuberculosis* strains to one of six lineages: East African-Indian, East Asian, Euro-American, Indo-Oceanic, West African-1, or West African-2.

There is also the fourth international global spoligotyping database (SpolDB4) that contains at least 39,925 strains from 122 countries classified into 62 clades or lineages (Brudey *et al.*, 2006). The spoligotype lineages used in MIRU-VNTRplus and SpolDB4 are mostly equivalent but there are some differences (Table 5.2). MIRU-VNTR*plus* was chosen for the ability to directly compare MIRU-VNTR data online.

Strains were compared to MIRU-VNTR*plus* using the categorical co-efficient which generated a match distance for each strain in the Midlands to the closest relative strain in MIRU-VNTR*plus*. A value of 0.000 equated to an absolute match at all of the 15 MIRU-VNTR loci and a value of 1.000 equated to a complete non-match at all 15 loci.

5.3.5. Global and regional geographical associations within the Midlands

For the analysis of the distribution of strain lineage and patient location of residence, three levels of geography were analysed within the Midlands: globally, regionally by counties (Figure 2.1), and within the county of the West Midlands (Figure 2.2A). Further examination of strain distribution was carried out by analysis of lineage distribution across the nine PCTs located in the county of the West Midlands (Figure 2.2B).

5.4. RESULTS

5.4.1. Prominent CEL Groups in the Midlands

To estimate the coverage of typing within the six year study period, the total proportion of *M. tuberculosis* isolates typed in 2006 and 2007 was compared to the national HPA MycobNet database that contains all drug sensitivity testing data in the UK and therefore all positive cultures. It was found that 1,686/1,765 (96%) isolates in MycobNet from the Birmingham laboratory had a MIRU-VNTR result. Of the 5,731 isolates typed between 2004 and 2009, a CEL group was identified in 5,625/5,731 (98%) of all patients (Table 5.1). The single most predominant CEL group overall was the Indian CEL group (1,202/5,731, 35%) with CEL groups who originated from Asia accounting for the highest number of patients from one continent with 1,587/5,731 (47%) of all patients. The European CEL group was the second most prevalent continental group (1,323/5,731, 39%), with the England CEL group the second most prevalent individual CEL group overall (1,190/5,731, 35%). Patients that originated from Africa (341/5,731, 10%), the Americas (47/5,731, 1%), and Oceania (5/5,731, <1%) accounted for 12% of all strains. The largest CEL group in Africa originated from Somalia (126/5,731, 4%). Within the Midlands patient population, there were a total of 134 individual CEL groups identified. The most diverse region was Southern Asia with 23 individual CEL groups which included: Pakistan (700/5,731, 12%); Bangladesh (193/5,731, 3%); and more specific geographical areas such as Kashmir (93/5,731, 2%) and 13 Indian States (172/5,731, 3%). Within the British Isles, patients that originated from each of the four nations were identified with: 1,190/5,731 (35%) from England; 130/5,731 (2%) from Ireland; 127/5,731 (2%) from Scotland; and 122/5,731 (2%) from Wales. The single largest non-specific CEL group was Western Asia (345/5,731, 6%). In the rest of Western Asia, there were two large specific CEL groups that originated from Lebanon (130/5,731, 2%) and Turkey (54/5,731, 1%).

Continental Origin of patient	Region	Most prevalent CEL Group	n (%)	No. of other CEL Groups In region	n (%)	Total Isolates (%)
Africa	Africa	Africa	18 (1)	1	3 (0)	21 (0)
	Eastern	Somalia	126 (4)	8	59 (3)	185 (3)
	Middle	Congo	12 (0)	1	1 (0)	13 (0)
	Northern	Morocco	69 (2)	2	17 (1)	86 (2)
	Southern	Black Southern Africa	74 (2)	7	49 (2)	123 (2)
	Western	Nigeria	42 (1)	5	61 (3)	103 (2)
		Region total	341 (10)	24	190 (8)	531 (9)
Americas	Caribbean	Black Caribbean	14 (0)	2	2 (0)	16 (0)
	Central	Central	2 (0)	0	0 (0)	2 (0)
	Northern	USA Black	11 (0)	3	11 (0)	22 (0)
	South	Brazil	20 (1)	2	4 (0)	24 (0)
		Region total	47 (1)	7	17 (1)	64 (1)
Asia	Central	Central	1 (0)	0	0 (0)	1 (0)
	Eastern	China Cantonese	24 (1)	5	36 (2)	60 (1)
	South-Eastern	Vietnam	15 (0)	6	21 (1)	36 (1)
	Southern	India	1,202 (35)	22	1,288 (55)	2,490 (43)
	Western	Western Asia	345 (10)	8	193 (8)	538 (9)
		Region total	1,587 (47)	41	1,538 (66)	3,125 (55)
Europe	Eastern	Poland	36 (1)	9	30 (1)	66 (1)
	British Isles	England	1,190 (35)	5	388 (17)	1,578 (28)
	Northern	Norway	10 (0)	6	24 (1)	34 (1)
	Southern	Italy	49 (1)	11	92 (4)	141 (2)
	Western	Germany	38 (1)	5	42 (2)	80 (1)
		Region total	1,323 (39)	36	576 (25)	1,899 (33)
Oceania		Region total	6 (0)	0	0 (0)	6 (0)
Not assigned	Not assigned	Not assigned	106 (3)	0	0 (0)	106 (2)
		Total	3,410 (100)	109	2,321 (100)	5,731 (100)

Table 5.1. Major Cultural, Ethnic, and Linguistic (CEL) Groups present in patients with typed *M. tuberculosis* isolates in the Midlands between 2004 and 2009.

5.4.2. Assignation of global strain lineage via MIRU-VNTR*plus*

MIRU-VNTR*plus* enables users to match against 186 reference strains using MIRU-VNTR, spoligotyping, LSP, or SNP data. Within the MIRU-VNTR*plus* database, there are 154 *M. tuberculosis* complex strains that are members of one of the six major LSP lineages and one of 17 spoligotype clades (Table 5.2). When the 5,731 strains from the Midlands were matched to MIRU-VNTR*plus*, 4,951/5,731 (86%) strains were matched to one of 19 specific spoligotypes including “non-human” clades and 5,636/5,731 (98%) strains were matched to one of the six LSP lineages. There were 780/5,731 (14%) and 95/5,731 (2%) strains with mixed lineages respectively. A strain was assigned as a mixed lineage if two or more MIRU-VNTR*plus* strains of different lineages matched with the closest distance to a strain from the Midlands. Therefore, to maintain the highest proportion of matched strains, the assignations to LSP lineages were used for further analysis.

Using the 15 MIRU-VNTR loci, 5,550/5,731 (97%) strains were matched to one of the six major human global lineages and 86/5,731 (2%) strains were matched to other recognised lineages with an average match distance of 0.16 (95 % CI: 0.15-0.16) for all 5,731 strains (Table 5.3). The West African-1 lineage showed the closest average match distance (0.05, 95% CI 0.01-0.08) with the Indo-Oceanic lineage exhibiting the highest average match distance (0.19, 95% CI 0.19-0.20) for a specific lineage. There were 3,475/5,731 (61%) typed strains from the Midlands that matched to strains in the MIRU-VNTR*plus* database with less than 0.17 match distance and 5,517/5,731 (96%) strains from the Midlands matched with less than 0.30 match distance. For this chapter, all unambiguous strain assignations were accepted as a global lineage assignation.

LSP Lineage	n (%)	MIRU-VNTR*plus* Spoligotype Lineage	n (%)	SpolDB4 Equivalent
East African Indian	10 (5)	Delhi/CAS	10 (5)	CAS
East Asian	10 (5)	Beijing	10 (5)	Beijing
Euro-American	91 (49)	Cameroon	10 (5)	LAM_CAM
		Ghana	10 (5)	T
		H37Rv	1 (1)	H37Rv
		Haarlem	13 (7)	Haarlem
		LAM	11 (6)	LAM
		NEW-1	3 (2)	Haarlem
		S	12 (6)	S
		TUR	4 (2)	LAM_TUR
		UgandaI	10 (5)	T
		UgandaII	10 (5)	T
		URAL	4 (2)	Haarlem
		X	3 (2)	X
Indo-Oceanic	12 (6)	EAI	12 (6)	EAI
M. bovis	11 (6)	Bovis	11 (6)	Bovis
M. canettii	2 (1)	Canetti	2 (1)	Canettii
M. caprae	11 (6)	Caprae	11 (6)	CAP
M. microti	6 (3)	llama	4 (2)	MICROTI
		vole	2 (1)	MICROTI
M. pinnipedii	2 (1)	Seal	2 (1)	PIN
West African-1	20 (11)	West African 1	20 (11)	AFRI
West African-2	11 (6)	West African 2	11 (6)	AFRI
	186 (100)		186 (100)	

Table 5.2. Distribution of lineages in the MIRU-VNTR*plus* database.

Equivalent LSP, MIRU-VNTR*plus* spoligotype clade and SpolDB4 spoligotype clades are shown. Global spoligotype lineages in MIRU-VNTR*plus* and SpolDB4 are equivalent but exact assignations differ for some lineages.

Strain lineage	n	Avg. Match Distance (95% CI)	No. strains with a MIRU-VNTR*plus* match distance of:		
			<0.17 (%)	<0.30 (%)	>0.30 (%)
East African Indian	1,888	0.15 (0.15-0.15)	1,287 (68)	1,868 (99)	20 (1)
East Asian	356	0.11 (0.10-0.11)	292 (82)	346 (97)	10 (3)
Euro-American	2,495	0.15 (0.14-0.15)	1,514 (61)	2,439 (98)	56 (2)
Indo-Oceanic	778	0.19 (0.19-0.20)	342 (44)	680 (87)	98 (13)
West African-1	16	0.05 (0.01-0.08)	15 (94)	16 (100)	0 (0)
West African-2	17	0.18 (0.13-0.22)	8 (47)	15 (88)	2 (12)
Mixed lineage	95	0.26 (0.25-0.27)	4 (4)	80 (84)	15 (16)
Other lineage	86	0.25 (0.23-0.27)	13 (15)	73 (85)	13 (15)
Total	5,731	0.16 (0.15-0.16)	3,475 (61)	5,517 (96)	214 (4)

Table 5.3. Similarity of 5,731 *M. tuberculosis* strains in the Midlands genotyped between 2004 and 2009 when compared to MIRU-VNTR*plus*.

5.4.3. Global distribution of *M. tuberculosis* Lineages in CEL groups resident in the Midlands

The Euro-American lineage was the single most prevalent lineage in our study with 2,495/5,731 (44%) strains and was present in each continental human population group (Table 5.4 and Figure 5.2). The Euro-American strain was the most prevalent lineage in patients that originated from Africa (296/531, 56%), the Americas (44/64, 69%), and Europe (1,342/1,899, 71%) and was the second most prevalent lineage in patients originating from Asia (752/3,125, 24%). The most prevalent *M. tuberculosis* lineage in patients that originated from Asia was the East African Indian lineage (1,499/3,125, 48%). Asia had the highest number of East Asian strains (218) with the two West African lineages most prevalent in Europe (16 and 17 isolates).

The East African Indian lineage was most closely associated with Asia (1,499/1,888, 79%), and in particular Southern Asia (1,258/1,888, 63%). Europe was the second closely associated continent with 270/1,888 (14%) and more specifically the British Isles CEL group (217/1,914, 11%).

Combining geographic data assigned by Origins and DNA fingerprinting data could influence public health efforts to control tuberculosis as this approach can identify strains in CEL groups where specific global *M. tuberculosis* lineages are not expected. The MIRU-VNTR profile 42235 2542517333 was the single East African Indian strain that infected the highest number of patients (n=152). This strain showed wide geographic distribution across the West and Midlands but was restricted in the distribution of patient origin as 142/152 (93%) patients originated from Asia whereas only 8/152 (5%) from Europe and one individual each from the Americas and Africa. The MIRU-VNTR profile 42435 2332515333 was assigned as a member of the East Asian lineage and was identified in 31 patients. Even though this strain originates from Eastern Asia, 19/31 (62%) of patients infected with this strain originated from the British Isles and only 12/31 (38%) patients in the Asian CEL group.

Continental origin of patient	**Region**	**No. isolates in each *M. tuberculosis* global lineage**								**Total Isolates (%)**
		East African Indian (%)	**East Asian (%)**	**Euro American (%)**	**Indo Oceanic (%)**	**West African-1 (%)**	**West African-2 (%)**	**Mixed Lineage (%)**	**Other Lineage (%)**	
Africa	Africa	1 (1)	1 (0)	17 (1)	2 (0)	0 (0)	0 (0)	0 (0)	0 (0)	21 (0)
	Eastern	45 (2)	10 (3)	74 (3)	42 (5)	0 (0)	2 (12)	11 (12)	1 (1)	185 (3)
	Middle	1 (0)	0 (0)	10 (0)	2 (0)	0 (0)	0 (0)	0 (0)	0 (0)	13 (0)
	Northern	36 (2)	2 (1)	26 (1)	17 (2)	0 (0)	2 (12)	3 (3)	0 (0)	86 (2)
	Southern	3 (0)	9 (3)	97 (4)	13 (2)	0 (0)	0 (0)	0 (0)	1 (1)	123 (2)
	Western	6 (0)	4 (1)	72 (3)	5 (1)	6 (38)	2 (12)	4 (4)	4 (5)	103 (2)
	Region Total*	92 (5)	26 (7)	296 (12)	81 (10)	6 (38)	6 (35)	18 (19)	6 (7)	531 (9)
Americas	Caribbean	3 (0)	2 (1)	11 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	16 (0)
	Central	0 (0)	0 (0)	2 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	2 (0)
	Northern	3 (0)	0 (0)	18 (1)	0 (0)	0 (0)	0 (0)	1 (1)	0 (0)	22 (0)
	South	4 (0)	2 (1)	13 (1)	5 (1)	0 (0)	0 (0)	0 (0)	0 (0)	24 (0)
	Region Total*	10 (1)	4 (1)	44 (2)	5 (1)	0 (0)	0 (0)	1 (1)	0 (0)	64 (1)
Asia	Central	0 (0)	0 (0)	1 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	1 (0)
	Eastern	1 (0)	36 (10)	15 (1)	7 (1)	0 (0)	0 (0)	1 (1)	0 (0)	60 (1)
	South-Eastern	5 (0)	12 (3)	8 (0)	11 (1)	0 (0)	0 (0)	0 (0)	0 (0)	36 (1)
	Southern	1,258 (67)	136 (38)	560 (22)	450 (58)	1 (6)	4 (24)	45 (47)	36 (42)	2,490 (43)
	Western	235 (12)	34 (10)	168 (7)	77 (10)	1 (6)	1 (6)	15 (16)	7 (8)	538 (9)
	Region Total*	1,499 (79)	218 (61)	752 (30)	545 (70)	2 (13)	5 (29)	61 (64)	43 (50)	3,125 (55)
Europe	Eastern	6 (0)	7 (2)	49 (2)	2 (0)	1 (6)	0 (0)	1 (1)	0 (0)	66 (1)
	British Isles	217 (11)	80 (22)	1,156 (46)	73 (9)	4 (25)	4 (24)	11 (12)	33 (38)	1,578 (28)
	Northern	6 (0)	2 (1)	13 (1)	11 (1)	0 (0)	1 (6)	1 (1)	0 (0)	34 (1)
	Southern	24 (1)	5 (1)	78 (3)	32 (4)	0 (0)	0 (0)	1 (1)	1 (1)	141 (2)
	Western	17 (1)	1 (0)	46 (2)	12 (2)	2 (13)	1 (6)	0 (0)	1 (1)	80 (1)
	Region Total*	270 (14)	95 (27)	1,342 (54)	130 (17)	7 (44)	6 (35)	14 (15)	35 (41)	1,899 (33)
Oceania	Region Total*	0 (0)	1 (0)	5 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	6 (0)
Not assigned	Not assigned	17 (1)	12 (3)	56 (2)	17 (2)	1 (6)	0 (0)	1 (1)	2 (2)	106 (2)
	Total	1,888 (100)	356 (100)	2,495 (100)	778 (100)	16 (100)	17 (100)	95 (100)	86 (100)	5,731 (100)

Table 5.4. Distribution of *M. tuberculosis* isolates in the Midlands between 2004 and 2009 according to lineage and continent of patient origin based on CEL group.

Percentages indicate the proportion of the total number of isolates in that particular lineage. *Region total, total number of isolates in that region and percentage of all isolates in that region that are members of that particular lineage.

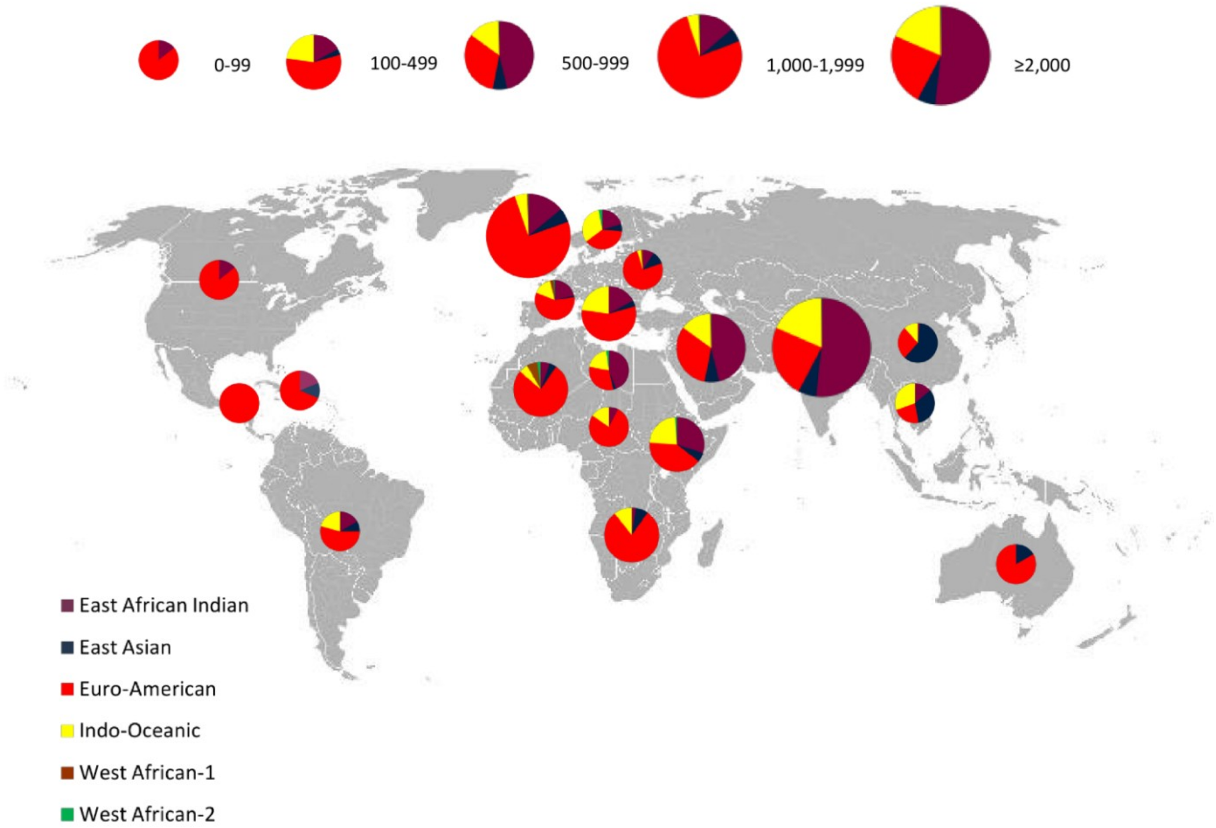


Figure 5.2. Global distribution of CEL groups and the six *M. tuberculosis* global LSP lineages present in the Midlands between 2004 and 2009.

A total of 5,550 strains were assigned to one of the six global lineages. Each circle is located in the continental region that each CEL group originates from. The proportions in each circle represent the proportion of each of the six *M. tuberculosis* lineages within that specific CEL group with the scale of the circles indicating the number of strains and patients within each region. Map adapted from http://commons.wikimedia.org/wiki/File:BlankMap-World-2009.PNG. The same colour scheme for lineages as used in Gagneux, 2006 was also used in this chapter to allow easy comparison.

5.4.4. Distribution of *M. tuberculosis* lineages in 11 counties across the Midlands

In eleven counties across the Midlands region, the county of the West Midlands had the highest number of typed isolates with 2,646/5,731 (46%) (Figure 5.3). In the East Midlands, Leicestershire had the highest number of typed isolates with 752/5,731 (13%). There were 3,193/5,731 (56%) typed isolates from the West Midlands, 1,672/5,731 (29%) isolates from the East Midlands, 143/5,731 (2%) patients resident outside of the Midlands, and 723/5,731 (13%) patients with no recorded place of residence. The number of isolates typed in each county ranged from 17 in Herefordshire to 2,646 in the West Midlands County. Three counties had less than 100 isolates typed (Lincolnshire, Herefordshire, and Shropshire) and six counties had between 100 and 499 isolates typed (Derbyshire, Northamptonshire, Nottinghamshire, Staffordshire, Warwickshire, and Worcestershire).

The Euro-American lineage was the most prevalent lineage in 10/11 (91%) counties with the highest proportion in Shropshire where it accounted for 40/61 (66%) isolates. Leicestershire was the only county where the most prevalent lineage was not the Euro-American lineage. In Leicestershire, the most prevalent lineage was the East African Indian lineage (272/752, 36%). Leicestershire was also the county with the highest proportion of the East Asian (65/752, 9%) and Indo-Oceanic (140/752, 19%) lineages. The two West African lineages did not account for more than 1% of typed strains in any of the eleven counties. The Euro-American and East African Indian lineages combined accounted for 4,383/5,731 (76%) of all strains typed in the Midlands.

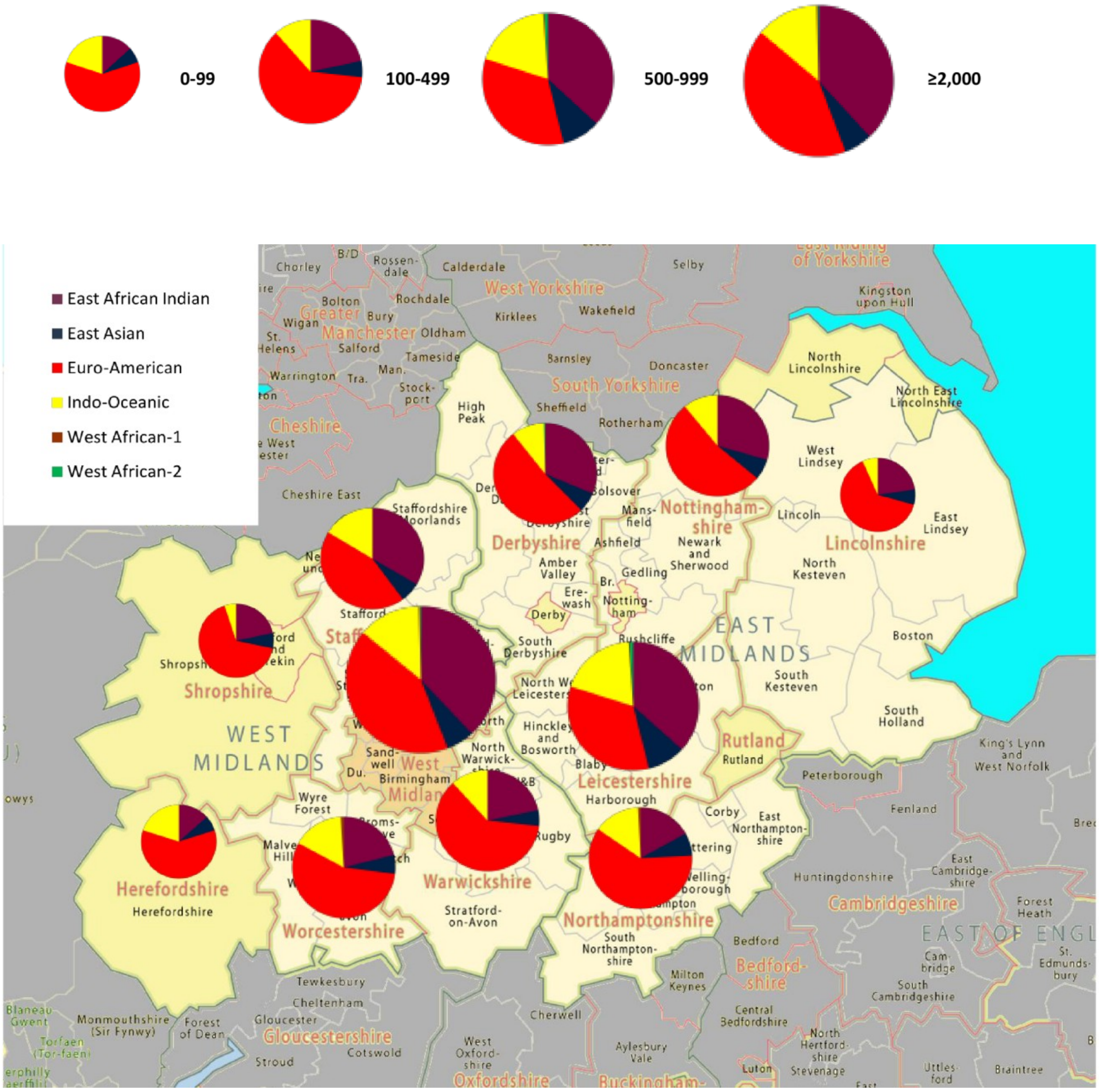


Figure 5.3. **Distribution of the six *M. tuberculosis* global LSP lineages in the Midlands between 2004 and 2009.**

A total of 4,782 strains had associated patient postcode of residence and were mapped. The scale of the circles indicates the number of strains within each county. Map was adapted from http://en.wikipedia.org/wiki/File:Map_of_the_administrative_geography_of_the_United_Kingdom.png.

5.4.5. Distribution of CEL groups in 11 counties across the Midlands

The Asian continental CEL group was the most prevalent CEL group across the Midlands with 3,125/5,731 (55%) patients and it was the most prevalent CEL group in 6/11 (55%0 counties with the highest proportion in Leicestershire where 531/752 (71%) patients originated from Asia (Figure 5.4). The Asian CEL group was also the most prevalent CEL group in Derbyshire (118/240, 49%), Staffordshire (114/238, 48%), and West Midlands county (1,610/2,646, 61%). The European continental CEL group was the most prevalent CEL group in Lincolnshire (54/75, 72%), Northamptonshire (119/252, 47%), Herefordshire (12/17, 71%), Shropshire (37/61, 61%), Warwickshire (69/124, 56%), and Worcestershire (74/107, 69%). In Nottinghamshire, there were almost equal proportions of the Asian (152/353, 43%) and European (151/353, 43%) CEL groups. The highest number of patients that originated from Africa (531/5,731, 9%) was in the West Midlands county (264/2,646, 10%) with the highest proportion in any county in Northamptonshire (35/252, 14%). The Asian and European CEL groups combined accounted for 5,024/5,731 (88%) of all strains typed in the Midlands.

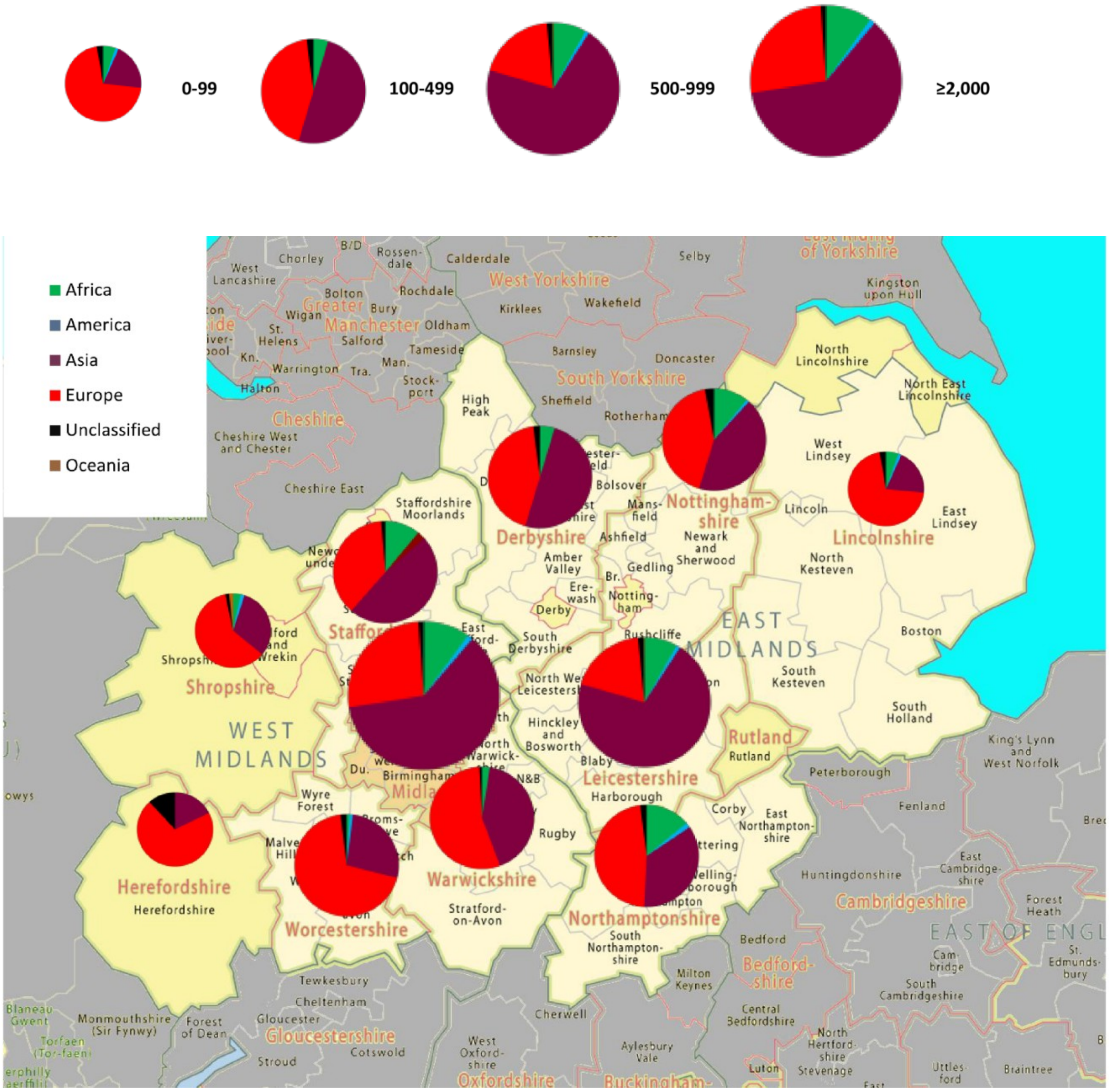


Figure 5.4. Distribution of CEL groups present in the Midlands (2004-2009).

A total of 4,782 strains had associated patient postcode of residence and were mapped. The scale of the circles indicates the number of strains within each county. Map adapted from http://en.wikipedia.org/wiki/File:Map_of_the_administrative_geography_of_the_United_Kingdom.png.

5.4.6. Distribution of *M. tuberculosis* lineages across nine PCTs within the West Midlands county

In nine Primary Care Trusts (PCTs) within the West Midlands county, the highest number of typed strains was in the Heart of Birmingham PCT with 811/2,646 (31%) typed strains (Figure 5.5). Two PCTs counties had less than 100 isolates typed (Dudley and Solihull) and six PCTs had between 100 and 499 isolates typed between 2004 and 2009 (Birmingham East and North, Coventry, Sandwell, South Birmingham, Walsall, and Wolverhampton City).

Within the nine PCTs in the West Midlands county, the Euro-American (1,075/2,646, 41%) and the East African Indian (972/2,646, 37%) lineages were the two most prevalent lineages accounting for 2,047/2,646 (77%) of all typed strains. The Euro-American lineage was the most prevalent lineage in 5/9 (56%) PCTs in the West Midlands county: Coventry (171/325, 53%); Dudley (43/97, 44%); Solihull (19/41, 46%); South Birmingham (110/217, 51%); and Wolverhampton City (167/283, 59%). The East African Indian lineage was the most prevalent lineage in 4/9 (44%) PCTs in the West Midlands county: Birmingham East and North (164/386, 42%); Heart of Birmingham (311/811, 38%), Sandwell (145/306, 47%); and Walsall (82/180, 46%). The highest number of Euro-American strains was in the Heart of Birmingham (283/811, 35%) with the highest proportion of Euro-American in a single PCT in Wolverhampton City where 167/283 (59%) were members of the Euro-American lineage whereas only 87/306 (28%) strains in Sandwell were Euro-American. The highest numbers of the East Asian (52/811, 6%) and Indo Oceanic lineages were both in Heart of Birmingham (127/811, 16%). The PCTs with the highest proportion of the East Asian and Indo Oceanic lineages were Dudley (10/97, 10%) and Solihull (7/41, 17%) respectively.

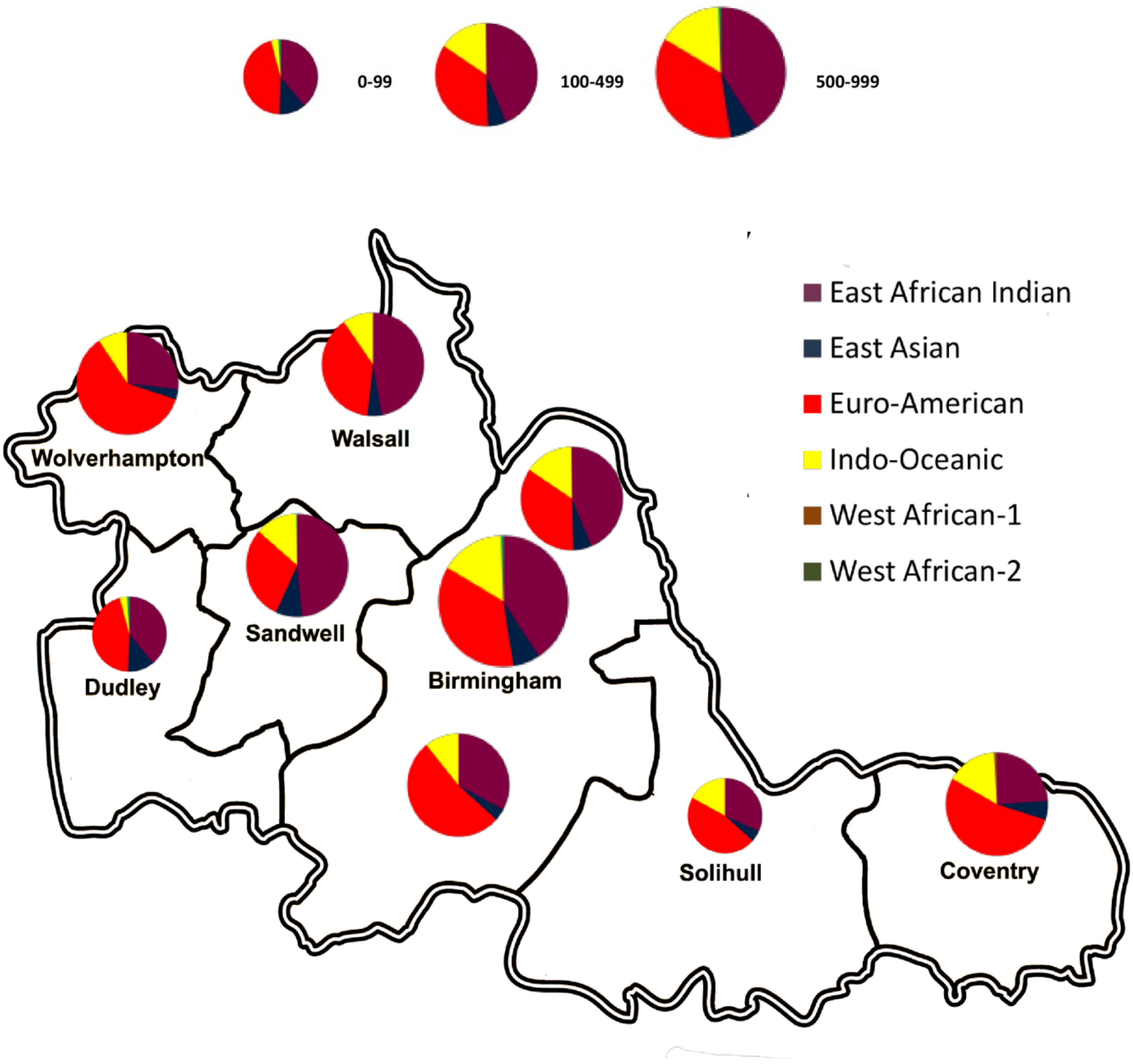


Figure 5.5. Distribution of the six *M. tuberculosis* global LSP lineages in the West Midlands county present between 2004 and 2009.

A total of 2,603 strains had associated patient postcode of residence and were mapped. The scale of the circles indicates the number of strains within each county. Map adapted from http://en.wikipedia.org/wiki/File:West_Midlands_County.png.

5.4.7. Distribution of CEL groups across nine PCTs within the West Midlands County

The Asian continental CEL group was the most prevalent CEL group in the West Midlands county with 1,610/2,646 (61%) patients and it was the most prevalent CEL group in all nine PCTs with the highest proportion in Sandwell where 210/306 (69%) patients originated from Asia (Figure 5.6). The lowest proportion in any PCT was in Dudley (45/97, 46%) and Solihull (19/41, 46%). Heart of Birmingham PCT had the highest numbers of patients originating from Africa (101/811, 12%), Asia (541/811, 67%), and Europe (156/811, 19%). Coventry had the highest proportion of patients originating from Africa (48/325, 15%) and Dudley and Solihull had the highest proportion of patients originating from Europe (43/97, 44% and 18/41, 44% respectively). The Asian and European CEL groups combined accounted for 2,322/2,646 (88%) of all strains typed in the West Midlands county.

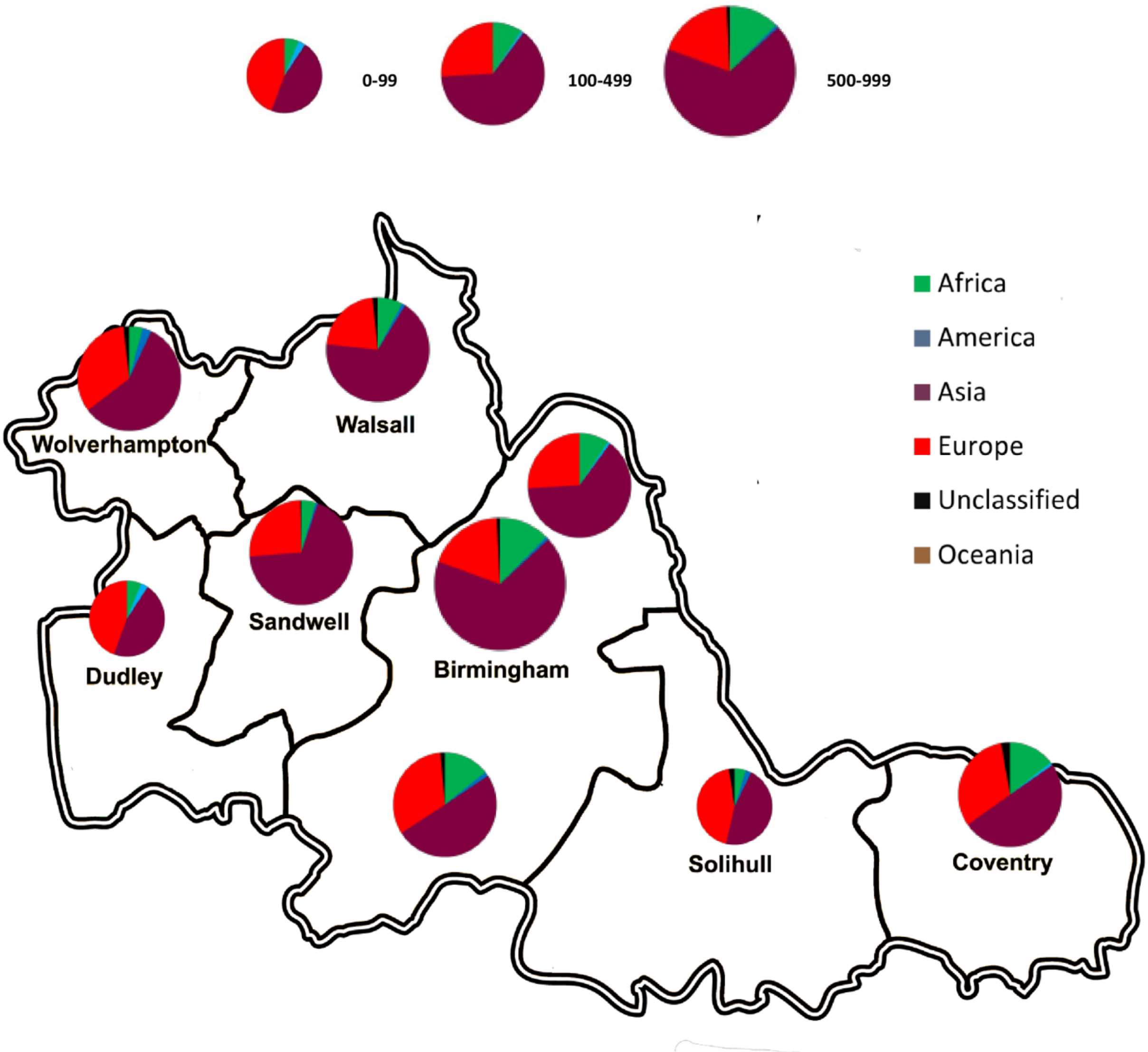


Figure 5.6. Distribution of continental CEL groups in the West Midlands County present between 2004 and 2009.

A total of 2,603 strains had associated patient postcode of residence and were mapped. The scale of the circles indicates the number of strains within each county. Map adapted from http://en.wikipedia.org/wiki/File:West_Midlands_County.png .

5.5. DISCUSSION

In this chapter, the analysis of 5,731 *M. tuberculosis* isolates typed in the United Kingdom demonstrated that the combination of molecular and population group data provided by novel software can provide information about the molecular epidemiology of *M. tuberculosis*.

The proportion of strains typed by MIRU-VNTR in this study was very high with 96% of all strains typed in 2006 and 2007 when compared to a national database. Most of the apparently untyped strains (4%) are probably data entries where an exact match between regional and national data was not obtained and so the data has remained unlinked.

The Midlands region of the UK encompasses a population of 9.8 million people with the single largest immigrant CEL group originating from India with 1,202/5,731 (21%) patients. However, the most predominant strain was the Euro-American lineage with 2,495/5,731 (44%) isolates. The data from this study support previous findings from studies in San Francisco and Montreal but also importantly extends the data set to Europe and includes a large number of strains from the ISC which were under represented in other studies (Gagneux *et al.*, 2006;Reed *et al.*, 2009). All of the six global lineages present in the Midlands exhibited previously identified associations with respective human population groups but crossover of all lineages between distinct human population groups was also identified which indicated that importation and transmission of strains between different human population groups may have occurred in the Midlands.

5.5.1. Distribution of global lineages in the Midlands: East African Indian

The association of lineages with distinct human population groups within the Midlands is similar when compared with previous studies (Gagneux *et al.*, 2006;Reed *et al.*, 2009), and also expands the global range of patients to the ISC, which were underrepresented in other studies. Only 17/875 (2%) strains (Gagneux *et al.*, 2006) and 91/730 (12%) strains (Reed *et al.*, 2009) in the San Francisco and Montreal studies originated from Southern Asia. In the Midlands, the East African Indian strain is tightly associated with patients who originate from Asia as 1,499/1,888 (79%) patients originating from Asia were infected with this lineage. However, East African Indian is not the sole lineage present in Asia as just under half of all patients that originated from Asia were infected with this lineage (1,512/3,125, 48%). The Euro-American (752/3,125, 24%) and Indo-Oceanic lineages (545/3,125, 17%) were also present in patients who originated from Asia. In a previous study based in San Francisco, there were only 11 isolates from Asia that were part of the East African Indian lineage (Gagneux *et al.*, 2006).

5.5.2. Distribution of global lineages in the Midlands: East Asian

In the Midlands, the East Asian lineage was not restricted only to the Eastern Asian CEL group as most patients infected with the East Asian lineage originated from Southern Asia (136/356, 38%), with only 36/356 (10%) patients from Eastern Asia. In two previous studies (Gagneux *et al.*, 2006;Reed *et al.*, 2009), most patients infected with the East Asian lineage did originate from Eastern Asia (190/277, 69% and 32/63, 51% respectively) and this lineage was also identified in patients who originated from Southeast Asia (73/277, 26% and 19/63, 30% respectively). However, in San Francisco and Montreal the East Asian lineage was rarely found in the autochthonous populations (7/277, 3% and 6/63, 10% respectively). In the Midlands, a higher proportion (80/356, 22%) of the autochthonous population was infected with the East Asian lineage.

The combination of DNA fingerprinting and origins data can provide deeper insights into the epidemiological backgrounds of specific MIRU-VNTR profiles. The two example MIRU-VNTR profiles shown (42235 2542517333 and 42435 2332515333) both originate from Asia but exhibit very different distribution in individuals within the Midlands. The 42235 strain is the most prevalent East African Indian strain and is highly restricted to patients who originate from Asia whereas the 42435 strain is a prevalent East Asian strain but has been identified predominantly in patients who originate from Europe. This indicates that transmission of the East Asian lineage may have occurred between two distinct CEL groups (Eastern Asia to British). However, it cannot be excluded that the original source may not have been resident in the Midlands and so transmission between CEL groups may have originally occurred outside the Midlands from a different CEL group altogether.

5.5.3. Distribution of global lineages in the Midlands: Euro-American

The Euro-American lineage was the only lineage to be identified in every one of the global regions represented in the Midlands and this lineage has been shown to be present throughout the world in studies on patients in San Francisco and Montreal (Gagneux *et al.*, 2006;Reed *et al.*, 2009).

5.5.4. Distribution of global lineages in the Midlands: Indo-Oceanic and West-African

Within the Midlands, the Indo-Oceanic lineage was identified predominantly in patients who originated from Southern Asia (450/2,490, 18%) whereas in the San Francisco and Montreal cohorts the Indo-Oceanic lineage was predominantly associated with South-Eastern Asia. This difference in associated regions may be caused by the paucity of patients who originate from South-Eastern Asia in the Midlands (36/5,731, <1%). In the previous studies in San Francisco and Montreal, West African strains were only identified in Africa whereas in the Midlands, 21/33 (64%) of the West African lineage were identified in patients that did not originate from Africa with 7/33 (21%) patients from Asia, 13/33 (39%) from Europe and 1/33 (3%) from Oceania.

5.5.5. Distribution of global lineages in the Midlands: Comparison of the Two Dominant Lineages

The two most prevalent lineages in the Midlands (East African Indian and Euro-American) have two contrasting patterns of distribution. The East African Indian lineage is highly restricted to patients originating from Asia as 79% (1,499/1,888) of patients infected with the East African Indian lineage originate from Asia. The Euro-American lineage exhibits a much more diverse distribution with Europe (1,342/2,495, 54%), Asia (752/2,495, 30%), and Africa (296/2,526, 12%) being the predominant continents associated with the Euro-American lineage. The dispersal of the Euro-American lineage throughout the world has been observed in two previous studies (Gagneux *et al.*, 2006;Reed *et al.*, 2009) but this is the first study to detail the restricted geographical association of the East African Indian lineage with Asia even in a dispersed population outside Asia.

5.5.6. Distribution of the East African Indian Lineage in Europe

The East African Indian LSP lineage or CAS spoligotype clade has been reported at low levels in mainland Europe in Hamburg. Germany (7.8%) (Oelemann *et al.*, 2007), 2007); 7% in Brussels, Belgium (Allix-Beguec *et al.*, 2008a); and in Spain (Alonso *et al.*, 2008); 9.4% in Greece (Rovina *et al.*, 2011). Within Southwest Ireland, only 4/171 (2%) patients were infected with a CAS strain (Ojo *et al.*, 2010). The highest reported proportion of CAS strains was also in England. In London, the most prevalent spoligotype was CAS (552/2,262, 24%) and patients who originated from the ISC were the largest patient population group (463/2,262, 20%) (Brown *et al.*, 2010).

5.5.7. Utility of Origins software in identification of CEL groups within countries

Origins software identified CEL groups within a country (e.g., Kashmir in Pakistan or various states in India) and divided the British Isles into four CEL groups. This enhanced differentiation could be useful in future population-based studies as migration patterns may be localized to specific areas within countries and common social networks could be identified. CEL groups can be assigned to any dataset in which the patient's name is known. Traditional epidemiologic identification of ethnic groups requires a questionnaire, but if patient names are not in a dataset, then CEL groups cannot be assigned (Webber, 2007).

The use of Origins software can enable a deeper understanding of epidemiological relationships between diverse CEL groups as a total of 134 individual CEL groups out of a potential total of >200 were identified whereas conventional epidemiological analysis such as the HPA Annual Report on the Epidemiology of Tuberculosis in the UK uses only nine ethnic groups (Bangladeshi, Black - other, Black-African, Black-Caribbean, Chinese, Indian, Mixed/other, Pakistani, and White) (Health Protection Agency, 2010). The completeness of ethnic group data in the HPA annual report is high as information was available on both ethnic group and place of birth for 8,064/9,040 (89%) TB cases in 2009 (Health Protection Agency, 2010). However, some studies do not have complete patient origin data as 39% (880/2,261) patients in a study based in London did not have any patient origin data (Brown *et al.*, 2010).

5.5.8. Identification of patient origin: Conventional versus Origins data

Comparing the data generated by the use of Origins software within the West and East Midlands to national data, the proportions for patients originating from Asia were very similar with 1,587/5,731 (47%) in the Midlands and 3,423/7,980 (43%) nationally in the UK. Proportions were also very similar for the few Black Caribbean patients diagnosed with TB each year with 14/5,731 (<1%) in the Midlands and 157/7,980 (2%) in the UK. The two most dissimilar groups were Africa/Black African with 341/5,731 (10%) in the Midlands versus 1,779/7,980 (22%) nationally and White/British with 1190/5,731 (35%) in the Midlands and 1648/7,980 (21%) nationally.

Origins was developed using electoral register data in the UK and in other countries around the world (Webber, 2007). It is known that Origins software has reduced sensitivity when applied to Black Caribbean populations as many individuals have apparently Anglo-Saxon names. Between the Midlands and UK data there are differences in the proportion of Black Caribbean patients. However, the total number of individuals originating from the Caribbean with TB in 2008 was only 2% of the total number of TB cases in the UK. CEL group data provided by Origins software cannot assign the country of birth of a patient whereas selfdeclared epidemiological data can analyse country of birth data. The distinction between country of birth is important as in all population groups in the UK, the incidence rates are higher in non-UK-born patients than in UK-born patients within the same ethnic groups (Health Protection Agency, 2010).

5.5.9. The impact of immigration on TB rates in the UK

In a study of the impact of migration within 21 European countries between 1996 and 2005, the UK was one of only three countries where an increase in rates between 1996 and 2005 was recorded (Gilbert *et al.*, 2009). Rates in the UK increased by 31% from 10.8 to 14.2 per 100,000 population. The two other countries that also had an increase in rates were Norway and Sweden. In Norway the incidence rate increased from 4.3 to 6.3 (31% increase) cases per 100,000 in nine years, and in Sweden the rate increased from 5.6 to 6.3 (13% increase) (Gilbert *et al.*, 2009). Portugal had the highest overall rate of TB in 2005 (34 cases per 100,000) but only 12% of cases occurred in the foreign-born population. Within the 21 countries analysed, Poland had the highest number of cases in 2005 with 9,269 (24.3 per 100,000) (European Centre for Disease Prevention and Control/WHO Regional Office for Europe., 2011). However, data from high incidence countries in Europe such as Russia were not included in the 21 countries studied.

The UK has the third largest foreign national population in Europe (3,035,000, 5% of total population) behind Germany (6,755,821, 9%) and France (3,263,200, 6%). In Europe, only the UK and Ireland have foreign national population groups that originate from countries with TB incidence rates of >500 cases per 100,000 with 168,000 individuals in the UK and 4,113 in Ireland (Gilbert *et al.*, 2009). Of the 21 countries included, the UK had the highest number and proportion of immigrants (3,575/5,994, 59.6%) with TB who originated from Asia in 2009 (European Centre for Disease Prevention and Control/WHO Regional Office for Europe., 2011;Gilbert *et al.*, 2009).

Between 1999 and 2003, the number of TB patients in England increased by 19% from 5,539 to 6,608. A study of this time period identified that TB in non-UK-born individuals who had presented less than five years after arrival in the UK was one of three groups significantly associated with the increase in TB across England. Patients co-infected with HIV and small increases in the number of cases living in deprived areas also contributed to the observed increase (Crofts *et al.*, 2008). Between 1999 and 2006, the proportion of patients with extra-pulmonary disease also significantly increased from 2,717 (48%) to 4,205 (53%). Analysis of this increase showed that it was significantly associated with an increase in the proportion of non-UK born cases (OR 2.7, 95% CI 2.6-2.8) (Kruijshaar and Abubakar, 2009).

5.5.10. Use of MIRU-VNTR*plus* and clade assignation

Within this chapter, patient origin was identified using Origins software. Mycobacterial lineages were assigned by using 15 locus MIRU-VNTR data with comparison to an online database (MIRU-VNTR*plus*). This database contains 186 reference *M. tuberculosis* complex strains with associated MIRU-VNTR, spoligotype, LSP, and SNP data for each strain. The 186 strains represent all of the major phylogenetic lineages known in the *M. tuberculosis* complex (Allix-Beguec *et al.*, 2008b).

The distribution of strains within MIRU-VNTR*plus* is heavily focused on the Euro-American lineage (91/186, 49%), "animal" strains (32/186, 17%) and the ancestral West African lineage

(31/186. 17%) whereas there are only 10 (5%) East African and 10 (5%) East Asian strains present. In the Midlands, an equivalent proportion of strains are members of the Euro-American lineage (2,526/5,731, 44%) but the East African Indian lineage is represented in a far greater proportion in the Midlands (1914/5731, 33%) than in MIRU-VNTR*plus*.

An internal evaluation of the MIRU-VNTR*plus* database by the database creators examined the effect of two different similarity distance cut-off values (0.17 and 0.30) on lineage assignation. These two values equate to 83% and 70% similarity respectively. Using MIRU-VNTR data only, a value of 0.17 corresponds to a maximum tolerance level of differences at four loci. A value of 0.17 matched 170/186 strains which resulted in a sensitivity value of 79% and specificity of 100%. The sensitivity of strain lineage assignation was increased to 95.1% when the distance cutoff was increased to 0.30 (up to seven variant MIRU-VNTR loci). However, specificity was reduced from 100% to 98.4% with two mismatches in the 186 strain database (Allix-Beguec *et al.*, 2008b).

Within the Midlands dataset, most strains (3,475/5,731, 61%) matched to a MIRU-VNTR*plus* strain with <0.17 match distance and almost all strains in the Midlands matched to a MIRU-VNTR*plus* strain with <0.30 match distance (5,517/5,731, 96%). Overall, all 5,731 strains in the Midlands matched to MIRU-VNTR*plus* exhibited an average match distance of 0.16 (95% CI: 0.15-0.16). This means that 61% of strains in the Midlands matched to MIRU-VNTR*plus* with differences in 4/15 analysed MIRU-VNTR loci and 96% strains matched with differences at seven loci.

5.5.11. The rationale for the use of LSP Lineages instead of spoligotype lineages

LSP lineages were used in this chapter as a greater proportion of strains from the Midlands were matched to a single LSP lineage in MIRU-VNTR*plus* (5,636/5,731, 98%) compared to single spoligotypes (4,951/5,731, 86%). There are also known discrepancies in MIRU-VNTR*plus* with the assignation of spoligotype clades that are members of the Euro-American lineage such as New-1, Clade X, and Haarlem (Allix-Beguec *et al.*, 2008b).

5.5.12. Improvement of matching to MIRU-VNTR*plus*

Between the human global spoligotype and LSP lineages, there are equivalent lineages present in each scheme as all but one of the LSP lineages corresponds to one spoligotype clade apart from the Euro-American lineage. The Euro-American lineage encompasses at least five separate spoligotype clades (Haarlem, LAM, S, T, X). This chapter used the 12 "classical" MIRU loci and three ETR loci. MIRU-VNTR*plus* allows a user to use various sets of loci including the 24 loci which make up the current internationally optimised set of MIRU-VNTR loci (Supply *et al.*, 2006). A study in Canada that examined the properties of 24 loci typing compared to the "classical" 12 MIRU loci typing when strain data was matched to MIRU-VNTR*plus.* This study showed that the number of strains that could not be assigned to a single spoligotype lineage were reduced from 282/650 (43%) using 12 loci to just one isolate using 24 loci. However, the increase in the number of loci increased the overall matching distance of all isolates as the number of strains that did not match any strains ("unknown lineage") in MIRU-VNTR*plus* increased from 23/650 (4%) using 12 loci to 429/650 (66%) using 24 loci (Christianson *et al.*, 2010). The MIRU-VNTR*plus* database does have a limited number of strains compared to other global databases such as SpolDB4 but MIRU-VNTR*plus* does provide a tool for direct assignation of lineages based on MIRU-

VNTR data that is obtained on a prospective basis in the Midlands. Spoligotyping data is not obtained prospectively in the Midlands. It is anticipated that MIRU-VNTR*plus* will eventually expand and include a wider collection of strains for comparison of *M. tuberculosis* strains from around the world.

5.5.13. Obtaining an accurate scenario for TB transmission in the Midlands

The data presented in this chapter is a large population based study with 5,731 patients from a population of almost 9.5 million people over a six year period from between 2004 and 2009. MIRU-VNTR data using the 15 loci does correlate well with IS*6110* RFLP data when epidemiological links are known (Hawkey *et al.*, 2003). However, knowledge of conventional epidemiological data across the 5,731 strains was not universal. When epidemiological data is not known, caution must be exercised when analysing MIRU-VNTR clusters as IS*6110* RFLP can still split 15 locus MIRU-VNTR clusters (Garcia de Viedma *et al.*, 2006). Therefore, an estimation of transmission rates and proportion of secondary cases generated by different strains or clades was not calculated. To improve the accuracy of MIRU-VNTR typing across the Midlands, prospective universal 24 loci typing of all strains was introduced in 2010. For 850 strains typed in 2010, an increase in the number of loci from 15 to 24 reduced the clustering rate across the Midlands from 45% to 23% (Evans *et al.*, 2011;Small *et al.*, 1994). To obtain an accurate estimation of transmission rates in the Midlands, strains typed between 2007 and 2009 will have the additional nine loci added to complete the 24 loci and combined with prospective typing data from 2010 to obtain an epidemiologically relevant four year time period (Vynnycky *et al.*, 2001).

Conventional epidemiological data is now analysed on a routine basis across the Midlands including the areas of highest incidence in Birmingham and Leicester. As part of a three year HPA national 24 locus typing project initiated in 2010 (Health Protection Agency, 2011), molecular and epidemiological data are being combined for the first time in the UK on a national level.

5.5.14. The use of repetitive sequences to identify phylogeographic associations

The use of MIRU-VNTR and spoligotype data for phylogenetic analysis has been examined in two recent studies that analysed the relative degree of convergent evolution or homoplasy when compared to SNP data. Homoplasy occurs when the same molecular epidemiological state such as strain type is identified in different lineages using a separate comparator method. A comparison of MIRU-VNTR, spoligotype, and SNP showed that SNP data exhibited the lowest rate of homoplasy in strains located in different parts of the global phylogeny of *M. tuberculosis* (Comas *et al.*, 2009). The authors recommended that MIRU-VNTR and spoligotype data can be used for initial exploratory screening of strains but that to overcome the phylogenetic limitations of these markers, complete strain assignment should be confirmed with SNP or LSP analysis. The SNP data analysed in this study was generated by an 89 gene target multilocus sequence analysis assay which generated 70 kbp of DNA sequence per strain. SNPs were detected by direct DNA sequencing of PCR products using the Sanger sequencing method (Hershberg *et al.*, 2008). However, the analysis of 89 potential SNPs per strain is beyond the realistic aims of most population-based phylogenetic studies of *M. tuberculosis* strains around the world especially in high-burden areas.

A subsequent larger study of 950 isolates in San Francisco compared spoligotypes obtained and the associated lineages based on LSP data (Kato-Maeda *et al.*, 2011). For 919 strains that had both spoligotype and LSP data, all isolates belonging to each of the recognised spoligotype families belonged to the same LSP/SNP derived lineage. No evidence of homoplasy in the main spoligotype lineages into different LSP/SNP lineages was identified. This study examined strains from the Euro-American (481/919, 53%), Indo-Oceanic (235/919, 26%), East Asian (198/919, 22%) but only five strains from the East African Indian LSP lineage and none from the two West African lineages.

5.5.15. Refinement of current phylogenies using minimal SNP sets

To overcome the need for massive gene sequencing to construct robust phylogenetic scenarios, minimal sets of informative SNPs involved in replication, repair and recombination (3R) in *M. tuberculosis* have been identified that provide accurate phylogenetic data (dos Vultos T. *et al.*, 2008). A minimal set of seven SNPs was used to further delineate the Haarlem, LAM, and T spoligotype clades which provides data that is almost as informative as massive sequencing projects (Abadia *et al.*, 2010) with identification of 30 SNPs in 192 clinical Beijing strains identified 26 robust sequence types within this clade (Mestre *et al.*, 2011).

5.5.16. Construction of robust global phylogenies using WGS

Strains analysed in this chapter will be selected for analysis by large-scale whole genome sequencing as part of the UK Clinical Research Collaboration Modernising Medical Microbiology Consortium which is a UK wide consortium funded by the Wellcome Trust and MRC. This will enable the construction of a robust phylogenetic scenario for the *M. tuberculosis* complex in the West and East Midlands. The three collaborative institutions in this project are the University of Oxford, HPA and the Wellcome Trust Sanger Institute. The aim of this project is to establish how revolutionary new technologies such as high-throughput next generation DNA sequencing can be optimally integrated into clinical microbiology. An Illumina system will be the primary next-generation platform that will be used for this project (Bentley, 2006;Balasubramanian and Bentley, 2001).

5.6. CONCLUSIONS

Many countries now routinely type *M. tuberculosis* isolates by using MIRU-VNTR typing. Our study of over 5,000 isolates shows that Origins data in combination with molecular typing data can be used to identify specific associations between lineages and apparent transmission into unexpected population groups. By using Origin software for identification of CEL groups, public health teams can identify and investigate possible cultural links for transmission of *M. tuberculosis*. We hypothesised that strains present in Southern Asia would also be present in the Midlands. This was proven true as the East African Indian clade was one of two predominant lineages in the Midlands.

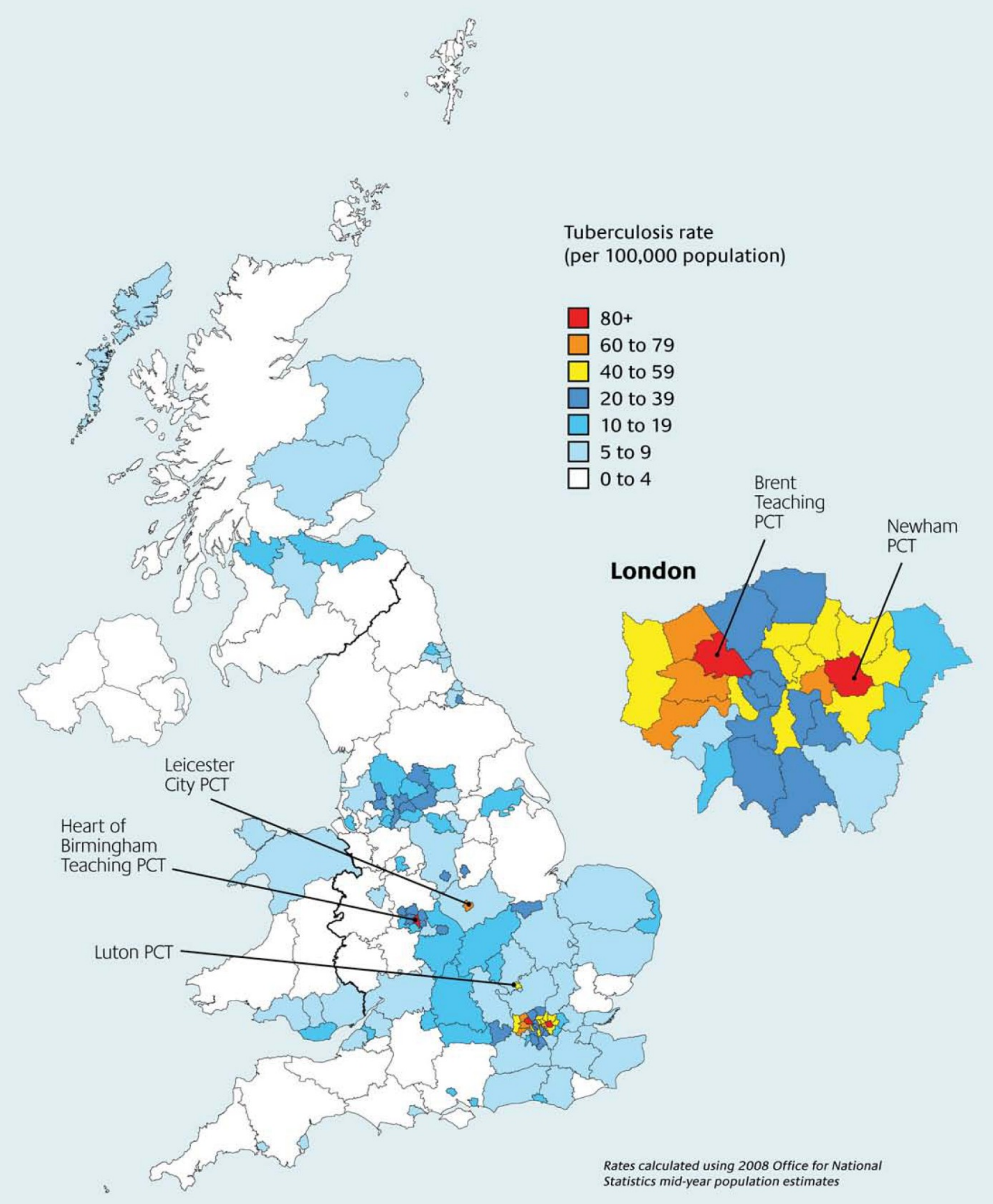
6. ANALYSIS OF THE CLONALITY OF *M. TUBERCULOSIS* STRAINS IN THE MIDLANDS AND IDENTIFICATION OF A GEOGRAPHICALLY-RESTRICTED BUT PREVALENT STRAIN.

6.1. INTRODUCTION

DNA fingerprinting of *M. tuberculosis* has a key role in TB control and cluster investigation as the molecular data obtained can be used to direct and focus public health control efforts (Evans *et al.*, 2007;Hawkey *et al.*, 2003). For example, DNA fingerprinting enhanced the investigation of a large outbreak in North London where many of the epidemiological links would not have been established by routine contact tracing or traditional epidemiological investigations alone (Ruddy *et al.*, 2004). Large-scale studies of *M. tuberculosis* strains have also enabled the accurate assessment of strain transmission. In one of the studies of TB transmission, a study in San Francisco identified that a third of new cases were the results of recent infection even with what was considered an efficient tuberculosis-control programme (Small *et al.*, 1994). In the Netherlands, demographic risk factors including male sex, urban residence, Dutch and Surinamese nationality, and long-term residence in the Netherlands were identified as significant risk factors for active transmission (van Soolingen *et al.*, 1999). More recently, a study in England revealed that recent transmission accounted for approximately 25% of cases and occurred over wide geographic areas, between ethnic groups and among the homeless (Love *et al.*, 2009).

6.1.1. TB in the Midlands and the UK

The geographical distribution of *M. tuberculosis* strains across health organisation boundaries (PCTs) will be described in this chapter (Figure 2.2). There is much variation in the incidence of TB across the Midlands (Figure 6.1), with rates highest in one urban area of Birmingham (>80 cases per 100,000) and lowest in rural areas such as Worcestershire (Figure 6.2) (Health Protection Agency, 2010).



Rates calculated using 2008 Office for National Statistics mid-year population estimates

Figure 6.1. Three-year average tuberculosis case rates in the UK by primary care organisation between 2007 and 2009.

The primary care organisations are primary care trusts (PCTs) in England, health and social services boards in Northern Ireland, NHS boards in Scotland, and local health boards in Wales (Health Protection Agency, 2010).

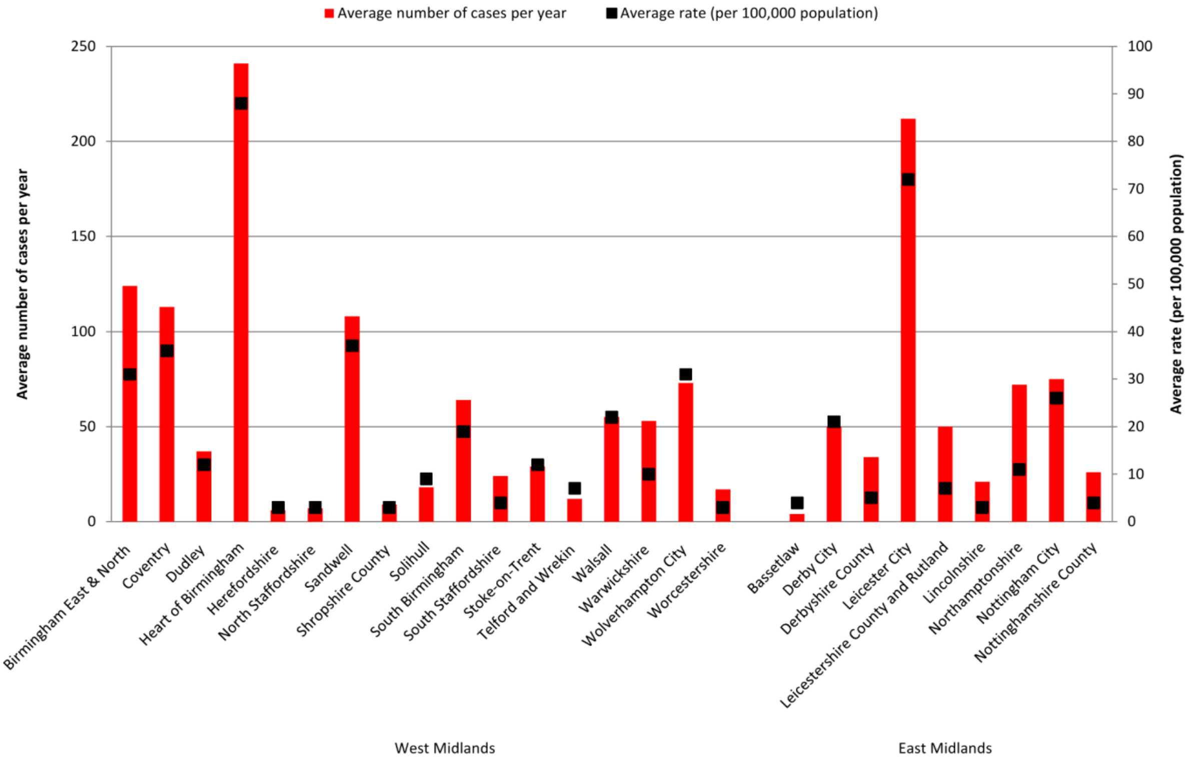


Figure 6.2. Three-year average tuberculosis case reports and rates by Primary Care Trusts in the West and East Midlands between 2007 and 2009.

Office for National Statistics mid-year figures from 2008 were used for population estimates (Health Protection Agency, 2010).

6.2. AIMS

We analysed all *M. tuberculosis* isolates in the Midlands region of the UK between 2004 and 2008 by universal prospective DNA fingerprinting with the aim of understanding the distribution and clonality of *M. tuberculosis* strains in the Midlands. The geographical distribution and epidemiological characteristics of patients infected with the most prevalent strain in the Midlands was analysed.

6.3. METHODS

6.3.1. Study population

Prospective universal DNA fingerprinting was undertaken between 2004 and 2008 with retrospective requested genotyping carried out on strains isolated before 2004. Retrospective observational epidemiological investigations were undertaken within one city and on a regional scale.

6.3.2. Mycobacterial culture, identification, and DNA fingerprinting

Primary specimens were processed as described in Section 2.3, cultured using liquid media (Section 2.3.2), DNA was extracted (Section 2.3.3) and isolates were identified as *M. tuberculosis* using the HAIN GenoType MTBC assay (Section 2.4). MIRU-VNTR typing of the five ETR (ETR-A to -E) and 10 MIRU loci (MIRU-02, -10, -16, -20, -23, -24, -26, -27, -39, and -40) was carried out using a Transgenomic WAVE® System (Section 2.7). MIRU-04 and MIRU-31 were removed from the 12 MIRU loci and analysed as ETR-D and ETR-E in the five ETR loci. IS*6110* RFLP (Section 2.5) and MIRU-VNTR typing was carried out on specific requested *M. tuberculosis strains* until 2004. A selection of strains were analysed by the nine additional loci (VNTR0424, 1955, 2163b, 2347, 2401, 3171, 3690, 4052, and 4156) that together with the 15 loci comprise the internationally optimised set of 24 loci (Section 2.7.4) (Supply *et al.*, 2006).

6.3.3. Case definition

Patients with the MIRU-VNTR profile of the most prevalent *M. tuberculosis* strain in the Midlands were included in further epidemiological investigations.

6.3.4. *M. tuberculosis* strain clustering analysis

MIRU-VNTR and IS*6110* RFLP data were entered into BioNumerics (Applied Maths, Saint Marten-Latem, Belgium) to identify clustered MIRU-VNTR profiles. MIRU-VNTR data was also queried against the HPA UK *M. tuberculosis* Strain Typing Database which contains >26,000 typed *M. tuberculosis* strains from six contributing centres in the United Kingdom (Health Protection Agency, 2011). Clustering percentages were calculated using the n-1 method (Section 2.9.3). The HGDI was used to assess the diversity and clonality of strains in the Midlands (2.9.4).

6.3.5. Assignation of global clade lineage

Spoligotyping was carried out to identify the global strain family that the most prevalent strain is part of. Spoligotyping was performed using the Luminex Multianalyte Profiling System as previously described (Cowan *et al.*, 2004) (Section 2.8). Spoligotype families were assigned by comparison to the international SpolDB4 database which contains 39,925 entries from 122 isolation countries (Section 2.9.6) (Brudey *et al.*, 2006).

6.3.6. West Midlands regional epidemiological analysis

Patients infected with the single most prevalent *M. tuberculosis* strain from 2004-2008 were compared to all patients with tuberculosis in the West Midlands between 2004 and 2008 in the HPA Enhanced Tuberculosis Surveillance System. The HPA Enhanced Tuberculosis

Surveillance System contains molecular, pathological, and treatment data on all notified cases of tuberculosis in England including culture-confirmed cases and clinically diagnosed cases. All patients with strain typing data were selected for comparison.

6.3.7. Patient location

M. tuberculosis strains referred from laboratories located in the West and East Midlands were included in this study. From the patient postcode, three levels of patient residence were analyzed: Health Protection Unit (HPU), PCT (Figure 2.2) and LA. Local HPA HPUs work alongside the National Health Service in England providing specialist support in communicable disease and infection control. There are three HPUs in the West Midlands: West Midlands East, West Midlands North, and West Midlands West. The major cities of Birmingham and Coventry are located in West Midlands East, Stoke-on-Trent in West Midlands North, and Wolverhampton in West Midlands West. More specific analysis of patient residence was undertaken by analyzing patient location within PCT or LA regions in the West Midlands. Primary Care Trusts provide primary and community services in England. LAs are administrative regions that are based on city or county boundaries (Figure 2.1).

6.3.8. Geographical distribution of the Mercian strain in the West Midlands region

Laboratory records of patients with the most prevalent strain were used to map patient residential location using postcode within the West Midlands with MapInfo software (Pitney Bowes Software, Watford, UK).

6.3.9. City specific epidemiological investigation

When it became apparent that there was a cohort of patients in Wolverhampton with an indistinguishable MIRU-VNTR profile, a retrospective review of patient case notes and interview of specialist tuberculosis nurses who were involved with the care of these patients was undertaken for culture-positive patients resident in Wolverhampton diagnosed with the same indistinguishable MIRU-VNTR profile between June 2003 and February 2006 to identify common factors and potential epidemiological links. These patients were compared to culture-positive cases diagnosed with other strains in 2004. Chest radiographs of all patients were also reviewed for the presence of cavitations.

6.3.10. Statistical analysis

Proportions calculated from epidemiological data obtained from the West Midlands regional and Wolverhampton city datasets were compared using Pearson's chi-squared test with Fisher's exact test was used where necessary. Univariate and multivariate logistic regression modelling was used to test the significance of odds ratios in Stata v10 (Stata Corp, College Station, TX, USA). The multivariable model was assembled by adding covariates individually in decreasing order of significance and the "goodness of fit" of each model was assessed using the likelihood ratio test. All cases with missing values for the variables examined were excluded from the multivariate model with 114 patients infected by the Mercian strain and 1,891 patients in the control group included. A univariate analysis of the epidemiological investigation of patients resident in Wolverhampton was undertaken using EpiData Analysis v2.2 (EpiData Association, Odense, Denmark). The extent of any association was expressed as an odds ratio (OR) with 95% confidence intervals (Section 2.9.9).

6.4. RESULTS

6.4.1. Summary of DNA fingerprinting data

Between 2001 and 2003, 290 *M. tuberculosis* strains identified between 1995 and 2002 were typed as part of requested retrospective epidemiological investigations (Figure 6.3). In 2003, a prospective service to genotype all referred *M. tuberculosis* strains in the Midlands was initiated, with 2004 the first full calendar year of typing.

Between 2004 and 2008, there were 4,830 isolates typed from 31 referring laboratories in the West and East Midlands. There were 171 duplicate isolates were removed so that 4,659 isolates representing the first typed isolate from each patient between 1st January 2004 and 31st December 2008 were selected. From 4,659 isolates typed, 2,791/4,659 (60%) originated from patients resident in the West Midlands and 1,464/4,659 (31%) from patients resident in the East Midlands. A summary flowchart of the molecular and epidemiological investigations undertaken in this study is shown in Figure 6.3.

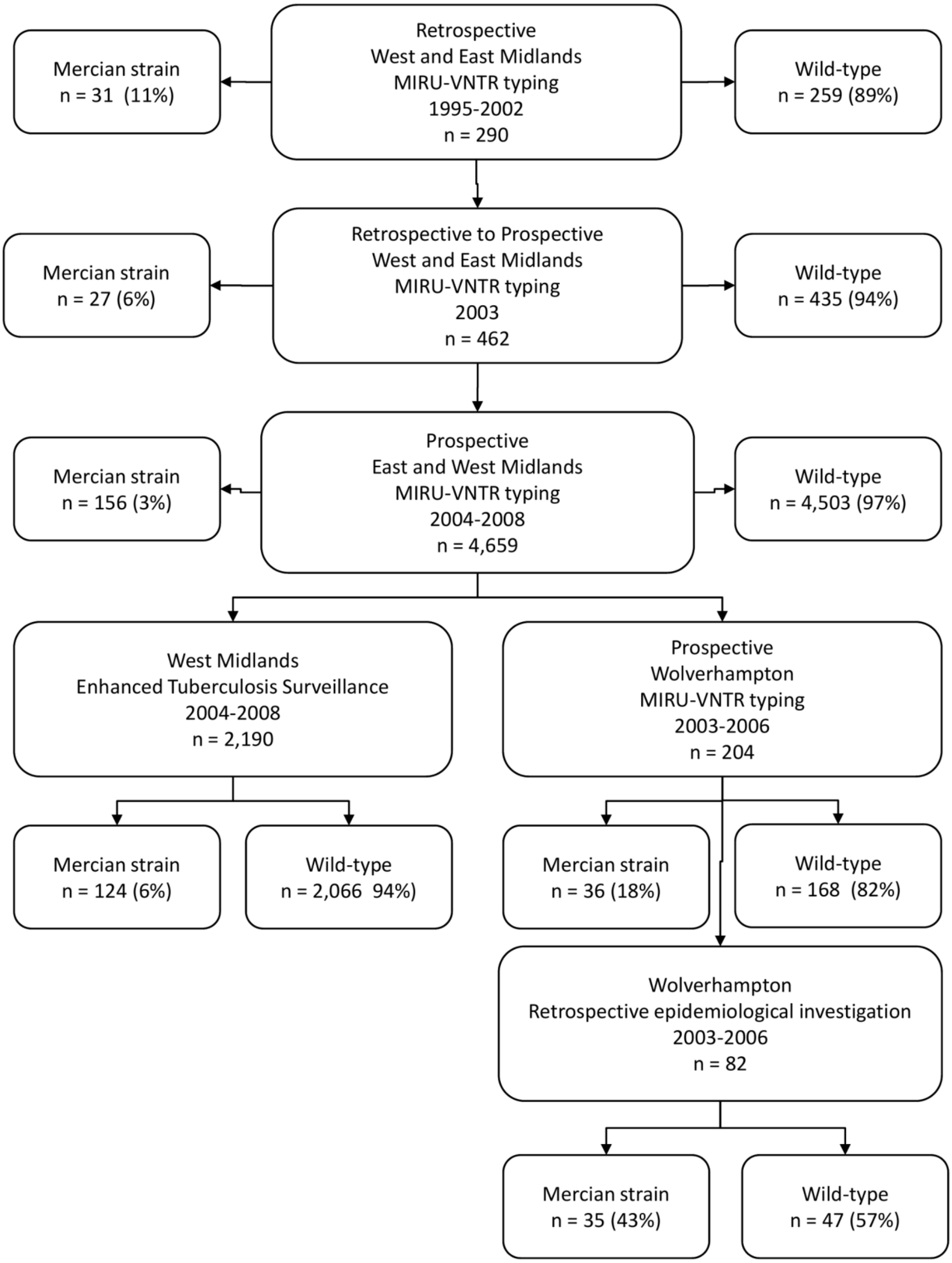


Figure 6.3. Flow diagram of DNA fingerprinting and epidemiological investigations.

6.4.2. MIRU-VNTR profile distribution across the Midlands

Across the Midlands, there were 489 molecular clusters that contained ≥2 isolates. Pairs of isolates were the most common cluster size with 216/489 (44%) clusters. Most clusters contained ≤5 isolates (372/489, 76%). There were three clusters that contained ≥85 isolates. The single largest cluster contained 156 isolates (Figure 6.4).

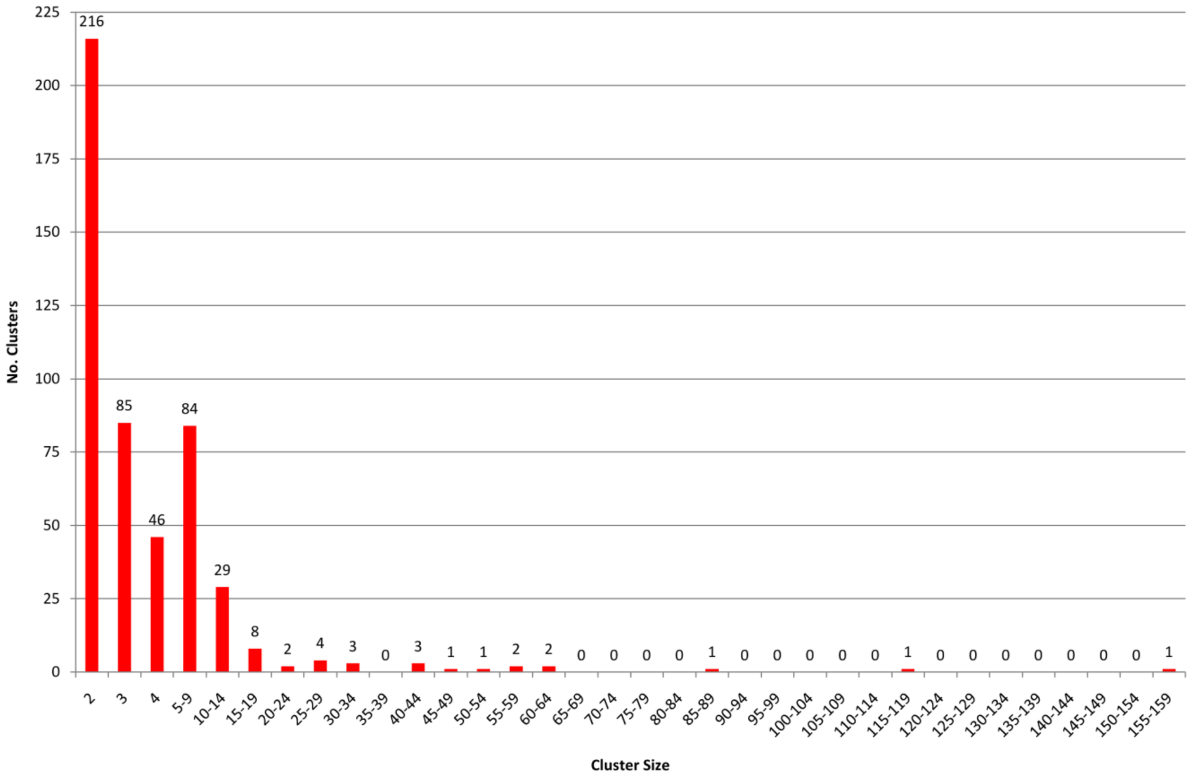


Figure 6.4. MIRU-VNTR profile distribution across the Midlands, 2004-2008.

6.4.3. Clonality and diversity of *M. tuberculosis* in regions across the Midlands

Heart of Birmingham and Leicester City PCTs had the highest number of MIRU-VNTR types (430 and 404), clusters (88 and 60), indistinguishable isolates (356 and 210) and total numbers of isolates (698 and 554) in the West and East Midlands respectively (Table 6.1). Two different areas had the highest proportion of strain clustering by MIRU-VNTR typing, Wolverhampton City (43%) and Northamptonshire (29%). Three areas had complete differentiation of all strains with 0% clustering and a HGDI value of 1.000 (Herefordshire, North Staffordshire, and Bassetlaw). Clustering rates and the HGDI were calculated to assess the clonality of strains within each PCT. Wolverhampton City PCT had the highest rate of clustering (43%) and a reduced HGDI value (0.948, 0.927 - 0.968), which indicated that in this location strains had a higher degree of clustering and clonality than other areas.

PCT	No. of MIRU-VNTR types	No. of MIRU-VNTR clusters	No. of indistinguishable isolates	No. of unique isolates	Total No. isolates	Clustering (%)	HGDI (95% CI)
West Midlands							
Birmingham East and North	232	35	131	197	328	29	0.994 (0.991 - 0.996)
Coventry	196	32	121	164	285	31	0.989 (0.983 - 0.994)
Dudley	60	15	46	45	91	34	0.979 (0.966 - 0.993)
Heart of Birmingham	430	88	356	342	698	38	0.994 (0.993 - 0.996)
Herefordshire	16	0	0	16	16	0	1.000 (0.974 - 1.000)
North Staffordshire	17	0	0	17	17	0	1.000 (0.977 - 1.000)
Sandwell	187	23	101	164	265	29	0.991 (0.987 - 0.995)
Shropshire County	25	2	4	23	27	7	0.994 (0.984 - 1.000)
Solihull	32	2	4	30	34	6	0.996 (0.990 - 1.000)
South Birmingham	143	16	56	127	183	22	0.993 (0.988 - 0.997)
South Staffordshire	62	8	23	54	77	19	0.991 (0.984 - 0.998)
Stoke On Trent	91	9	29	82	111	18	0.994 (0.989 - 0.998)
Telford and Wrekin	25	1	2	24	26	4	0.997 (0.989 - 1.000)
Walsall	123	21	62	102	164	25	0.994 (0.991 - 0.997)
Warwickshire	93	10	35	83	118	21	0.986 (0.975 - 0.997)
Wolverhampton City	147	22	133	125	258	43	0.948 (0.927 - 0.968)
Worcestershire	71	10	32	61	93	24	0.989 (0.982 - 0.997)
East Midlands							
Bassetlaw	3	0	0	3	3	0	1.000 (0.310 - 1.000)
Derby City	115	16	48	99	147	22	0.993 (0.989 - 0.998)
Derbyshire County	50	7	19	43	62	19	0.989 (0.980 - 0.999)
Leicester City	404	60	210	344	554	27	0.996 (0.995 - 0.998)
Leicestershire County and Rutland	91	11	30	80	110	17	0.994 (0.989 - 0.999)
Lincolnshire	41	5	15	37	52	19	0.980 (0.958 - 1.000)
Northamptonshire	154	23	87	131	218	29	0.988 (0.982 - 0.995)
Nottingham City	180	23	74	157	231	22	0.996 (0.994 - 0.998)
Nottinghamshire County	69	10	28	59	87	21	0.992 (0.987 - 0.998)
PCT not assigned	259	37	115	222	337	23	0.997 (0.996 - 0.998)
West Midlands	1,379	299	1,711	1,080	2,791	51	0.994 (0.993 - 0.995)
East Midlands	915	164	713	751	1,464	38	0.997 (0.997 - 0.998)
Midlands	2,195	489	2,953	1,706	4,659	53	0.996 (0.996 - 0.996)

Table 6.1. MIRU-VNTR profile distribution by PCT.

Profile distribution in shown by the number of distinct MIRU-VNTR profiles, the number of MIRU-VNTR clusters that contain ≥2 isolates, the number of clustered or indistinguishable isolates, and the number of unique isolates. The clonality or diversity of strains in each PCT was assessed by calculated the clustering rate using the n-1 method and the HGDI.

6.4.4. Distribution of the most prevalent MIRU-VNTR profile in the Midlands

The single most prevalent MIRU-VNTR profile was 32333 2432515314, which was identified in 156/4,659 (3%) isolates across the West and East Midlands and 155/2,791 (6%) isolates in the West Midlands. One isolate with this MIRU-VNTR profile was identified in the East Midlands (Table 6.2). The second and third most prevalent profiles were 42235 2542517333 (117/4,659, 3%) and 42235 2642515333 (88/4,659, 2%).

No.	MIRU-VNTR profile	MIRU-VNTR*plus* Lineage	Patient residence				Total
			West Midlands (%)	East Midlands (%)	Not assigned (%)	Non-Midlands (%)	
1	32333 2432515314	Euro-American	149 (96)	1 (1)	6 (4)	0 (0)	156
2	42235 2542517333	East African Indian	82 (70)	28 (24)	4 (3)	3 (3)	117
3	42235 2642515333	East African Indian	48 (55)	33 (38)	7 (8)	0 (0)	88
4	42235 2642516333	East African Indian	55 (86)	7 (11)	2 (3)	0 (0)	64
5	21433 2412615221	Euro-American	34 (55)	21 (34)	4 (6)	3 (5)	62
6	32433 2312515322	Euro-American	49 (84)	2 (3)	7 (12)	0 (0)	58
7	42235 2642517333	East African Indian	37 (64)	18 (31)	3 (5)	0 (0)	58
8	42234 2642517323	East African Indian	41 (76)	9 (17)	4 (7)	0 (0)	54
9	32433 2312515324	Euro-American	36 (77)	7 (15)	4 (9)	0 (0)	47
10	-2235 2542517333	East African Indian	34 (79)	5 (12)	4 (9)	0 (0)	43
11	32433 2512511322	Euro-American	24 (57)	17 (40)	1 (2)	0 (0)	42
12	32333 2532515323	Euro-American	17 (41)	21 (51)	3 (7)	0 (0)	41
13	32433 2432515323	Euro-American	13 (41)	16 (50)	3 (9)	0 (0)	32
14	42234 2742511334	East African Indian	12 (39)	16 (52)	2 (6)	1 (3)	31
15	42435 2332517333	East Asian	11 (37)	14 (47)	3 (10)	2 (7)	30
16	32433 2432515324	Euro-American	18 (67)	6 (22)	3 (11)	0 (0)	27
17	32333 2432515324	Euro-American	11 (42)	11 (42)	4 (15)	0 (0)	26
18	42435 2332515333	East Asian	22 (85)	2 (8)	2 (8)	0 (0)	26
19	32433 2332615321	Euro-American	18 (72)	5 (20)	1 (4)	1 (4)	25
20	32333 2532315323	Euro-American	14 (58)	10 (42)	0 (0)	0 (0)	24
21	42433 2331513321	Euro-American	1 (5)	19 (95)	0 (0)	0 (0)	20
22	22433 2342515322	Euro-American	7 (37)	10 (53)	2 (11)	0 (0)	19
23	32333 1532515323	Euro-American	16 (84)	1 (5)	1 (5)	1 (5)	19
24	42235 2642513333	East African Indian	9 (47)	9 (47)	1 (5)	0 (0)	19
25	61464 2432622334	Indo-Oceanic	14 (78)	4 (22)	0 (0)	0 (0)	18
26	32333 2531315323	Euro-American	13 (81)	1 (6)	1 (6)	1 (6)	16
27	42235 2442517333	East African Indian	7 (44)	8 (50)	1 (6)	0 (0)	16
28	42235 2542516333	East African Indian	10 (67)	3 (20)	2 (13)	0 (0)	15
29	22434 243251a322	Euro-American	13 (87)	0 (0)	2 (13)	0 (0)	15
30	21433 2312615221	Euro-American	5 (36)	6 (43)	2 (14)	1 (7)	14
	459 other profiles		1,008 (58)	550 (32)	147 (8)	26 (2)	1,731
	Unique profiles		963 (56)	604 (35)	112 (7)	27 (2)	1,706
	Total		2,791 (60)	1,464 (31)	338 (7)	66 (1)	4,659

Table 6.2. Most prevalent MIRU-VNTR profiles in the Midlands (2004-2008).

The 30 most prevalent MIRU-VNTR profiles are shown. The MIRU-VNTR*plus* lineage was assigned as described in Chapter 5. The total number and proportion of each lineage in the West and East Midlands is also shown.

6.4.5. Distribution of the most prevalent MIRU-VNTR profile across the Midlands

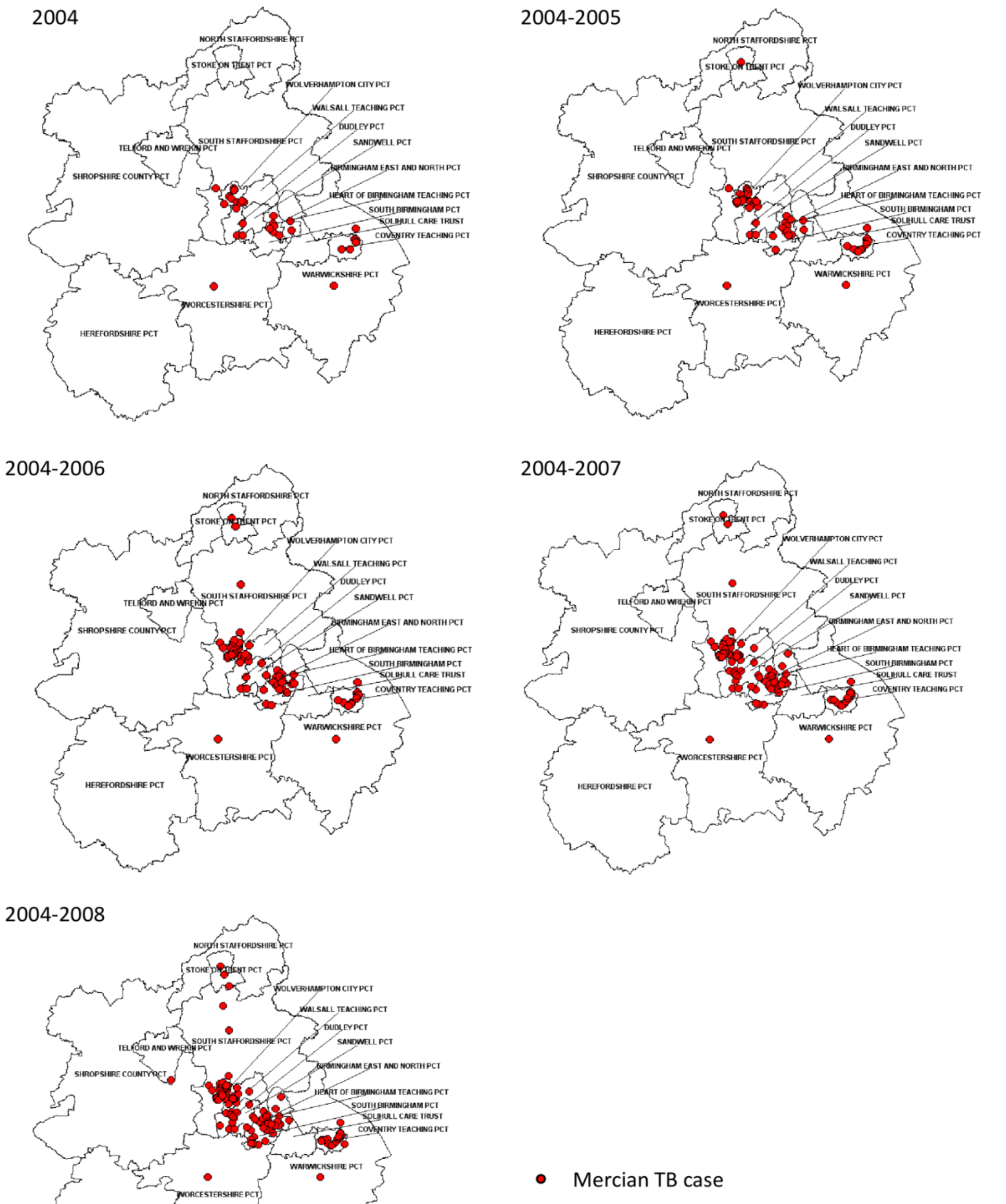
Analysis of the geographical distribution of the most prevalent MIRU-VNTR profile (32333 2432515314) showed that this strain was identified in 14/17 (83%) PCTs in the West Midlands. Wolverhampton had the highest proportion of this strain with 55/258 (21%) of all strains in Wolverhampton identified as 32333 2432515314 (Table 6.3).

Overall, 121/156 (78%) isolates of the Mercian strain were identified in patients resident in three cities in the West Midlands (Birmingham, Coventry and Wolverhampton) which are located within a 25 km radius of each other. There were 55/156 (35%) isolates that originated from Wolverhampton, 40/156 (26%) isolates from Birmingham, and 26/156 (17%) isolates from Coventry. Figure 6.5 shows the geographical mapping of 32333 2432515314 in the West Midlands between 2004 and 2008.

PCT	32333 2432515314 (%)	488 Other profiles (%)	Unique strains (%)	Total
West Midlands				
Birmingham East and North	12 (4)	189 (58)	127 (39)	328
Coventry	26 (9)	159 (56)	100 (35)	285
Dudley	10 (11)	61 (67)	20 (22)	91
Heart of Birmingham	16 (2)	448 (64)	234 (34)	698
Herefordshire	0 (0)	8 (50)	8 (50)	16
North Staffordshire	0 (0)	13 (76)	4 (24)	17
Sandwell	4 (2)	171 (65)	90 (34)	265
Shropshire County	0 (0)	16 (59)	11 (41)	27
Solihull	1 (3)	20 (59)	13 (38)	34
South Birmingham	12 (7)	100 (55)	71 (39)	183
South Staffordshire	4 (5)	48 (62)	25 (32)	77
Stoke On Trent	3 (3)	64 (58)	44 (40)	111
Telford and Wrekin	1 (4)	14 (54)	11 (42)	26
Walsall	2 (1)	109 (66)	53 (32)	164
Warwickshire	2 (2)	65 (55)	51 (43)	118
Wolverhampton City	55 (21)	140 (54)	63 (24)	258
Worcestershire	1 (1)	54 (58)	38 (41)	93
East Midlands				
Bassetlaw	0 (0)	3 (100)	0 (0)	3
Derby City	0 (0)	101 (69)	46 (31)	147
Derbyshire County	0 (0)	38 (61)	24 (39)	62
Leicester City	0 (0)	299 (54)	255 (46)	554
Leicestershire County and Rutland	0 (0)	67 (61)	43 (39)	110
Lincolnshire	0 (0)	30 (58)	22 (42)	52
Northamptonshire	0 (0)	132 (61)	86 (39)	218
Nottingham City	1 (0)	132 (57)	98 (42)	231
Nottinghamshire County	0 (0)	57 (66)	30 (34)	87
Not assigned	6 (2)	220 (65)	111 (33)	337
West Midlands	149 (5)	1,679 (60)	963 (35)	2,791
East Midlands	1 (0)	859 (59)	604 (41)	1,464
Non-Midlands	0 (0)	39 (58)	28 (42)	67
Total	156 (3)	2,797 (60)	1,706 (37)	4,659

Table 6.3. **Distribution of the Mercian strain in PCTs across the Midlands.**

The number and proportion of the Mercian strain, all other clustered strains and all unique strains in each PCT is shown.



Boundaries indicate West Midlands Primary Care Trusts. Crown copyright. All rights reserved. Health Protection Agency, 100016969 (2009).

Figure 6.5. Geographical mapping of patients infected with the Mercian strain in the West Midlands between 2004 and 2008.

The postcode map was produced under license by ©Crown Copyright and database right 2010. Ordnance Survey License 100016969/100022432.

6.4.6. Temporal distribution of the most prevalent MIRU-VNTR profile

The Mercian strain has been identified in each year of prospective typing between 2004 and 2008 with an average of 31 isolates (range 27-37) each year (Figure 6.6). The average annual proportion of all strains that are the Mercian strain across the Midlands was 3.35% (range 2.93%-4.13%).

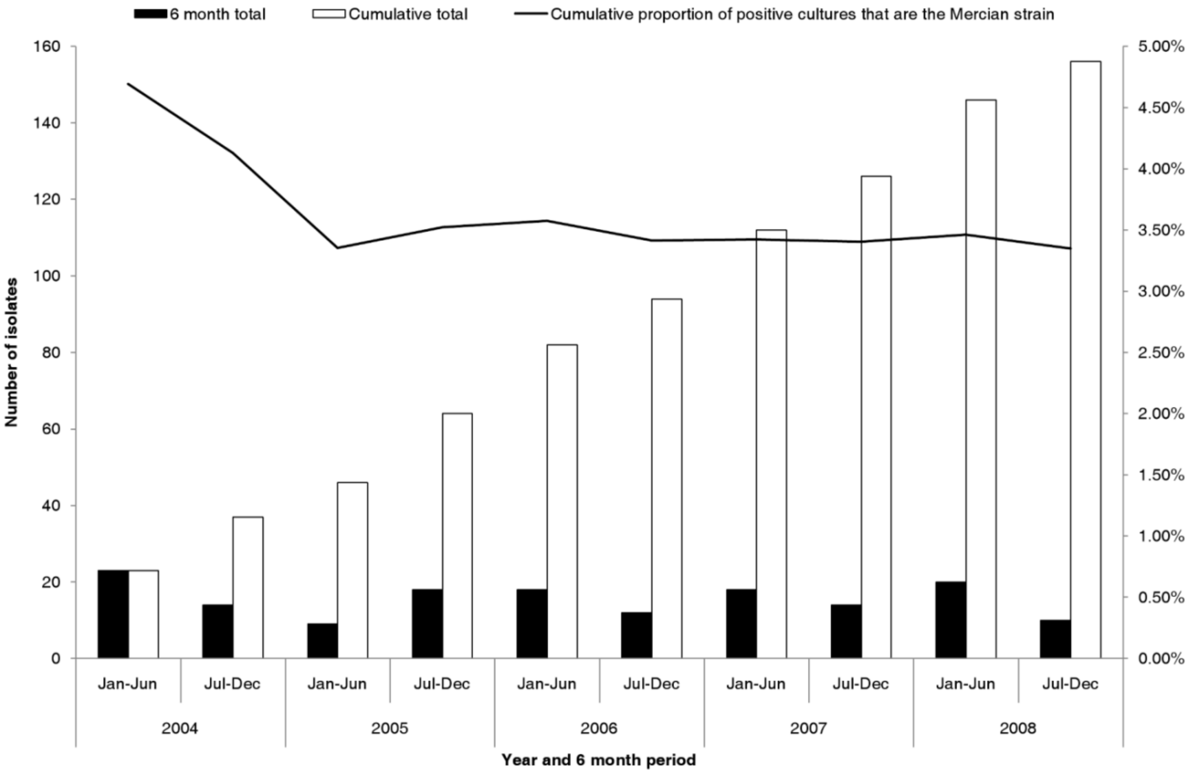


Figure 6.6. Incidence and cumulative total of the number of patients infected with the Mercian strain in the Midlands between 2004 and 2008.

The total for each six month period is shown in black with the white bars showing the cumulative total. The proportion of all strains in the Midlands that are the Mercian strain is shown on the right-hand y axis.

6.4.7. Distribution of the most prevalent MIRU-VNTR profile in other regions in the UK

The HPA UK *M. tuberculosis* Strain Typing Database was interrogated to analyze the national distribution of the Mercian strain. This database found a total of 176 isolates identified as 32333 2432515314 across the UK from 2004-2008. Only 6/176 (3%) of these isolates were identified in patients resident outside of the Midlands. The six strains were identified by regional laboratories in Cardiff (n=1), London (n=3), Newcastle (n=1) and Edinburgh (n=1) with 14 duplicate isolates present in the national database from the Midlands. Since this MIRU-VNTR profile appeared to be geographically restricted to the West Midlands in the UK, we have named the profile the “Mercian strain”, after the Anglo-Saxon kingdom of Mercia (Zaluckyj , 2001).

6.4.8. Case definition

Inclusion of a patient into the Mercian strain profile was based on 15 locus MIRU-VNTR typing, with a confirmed case defined as a patient with microbiologically confirmed tuberculosis and an isolate that had the 32333 2432515314 MIRU-VNTR profile.

6.4.9. Confirmation of genetic homogeneity of the Mercian strain across the Midlands by IS*6110* RFLP

Cluster investigations that involved retrospective typing of stored *M. tuberculosis* isolates originally identified between 1995 and 2003 (before the introduction of universal typing) resulted in the identification of an additional 51 isolates from six different locations as members of the 32333 2432515314 Mercian strain MIRU-VNTR profile. Upon further investigation by IS*6110* RFLP, 50/51 (98%) of these strains were still indistinguishable (

Figure 6.7) with a 7-band RFLP pattern. One isolate from Wolverhampton in 2003 possessed eight copies of IS*6110* but was still considered as part of the same strain and the MIRU-VNTR profile did not vary in this single isolate. A total of 58 isolates were retrospectively identified as the Mercian strain by MIRU-VNTR typing between 1995 and 2003 (Figure 6.3). When requested retrospective typing was undertaken prior to 2003, 31/290 (12%) of all strains genotyped by MIRU-VNTR were identified as the Mercian strain.

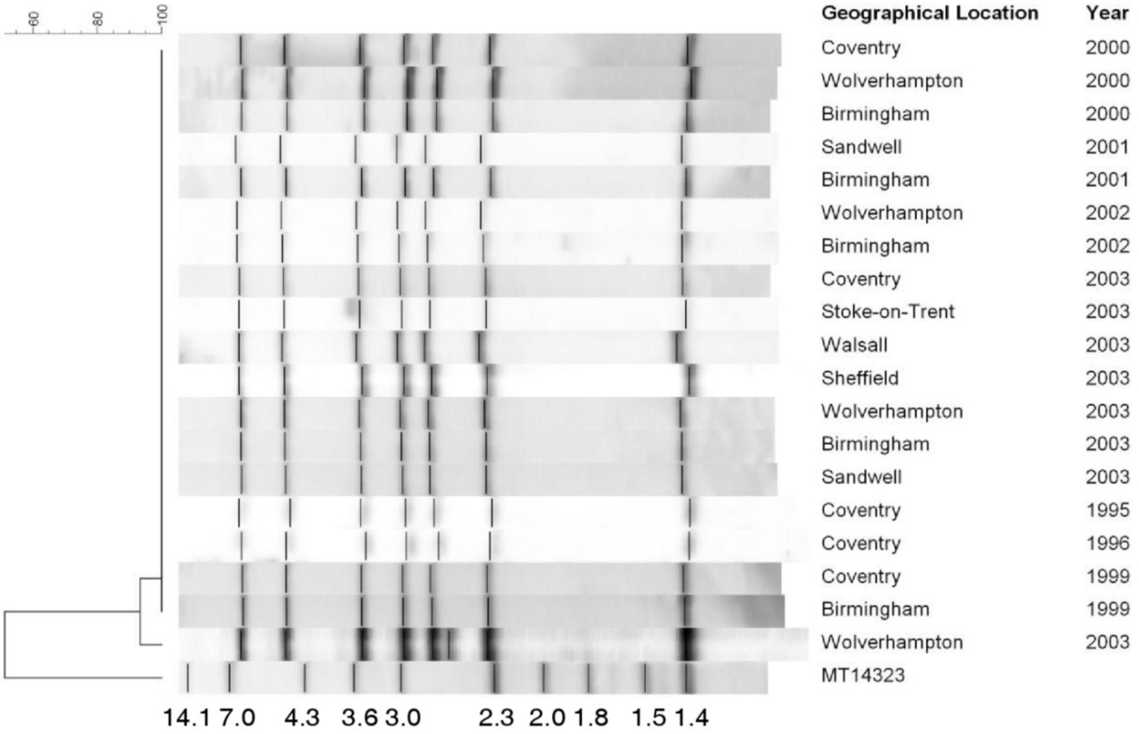


Figure 6.7. RFLP analysis of the Mercian strain across the Midlands.

Each horizontal lane is an example individual *M. tuberculosis* strain from each location and year. The molecular sizes of the digested DNA fragments of MT14323 are shown in kilobases. The 8th IS*6110* fragment was identified in an isolate from Wolverhampton in 2003.

6.4.10. West Midlands regional epidemiological analysis

Since the Mercian strain was restricted to the West Midlands and in particular Wolverhampton, regional and city-wide epidemiological analyses were carried out to understand the basic epidemiological associations and more in-depth social links within Wolverhampton. A total of 124/156 (79%) tuberculosis patients with the Mercian strain were successfully matched to notification data in the HPA Enhanced Tuberculosis Surveillance system. There were 2,190 tuberculosis patients with other strain types notified in the West Midlands between 2004 and 2008. A univariate analysis identified that patients who were resident in the West Midlands West HPU areas and specifically resident in Wolverhampton, UK-born, black Caribbean or white ethnic group had a significantly higher risk of disease caused by the Mercian strain (Table 6.4 and Table 6.5). Significant negative associations were identified with age not greater than 65 years old, the Black African ethnic group or extrapulmonary disease. No significant associations with drug resistance were identified ($p>0.05$). The significant variables were then included in a multivariate logistic regression which identified that being UK-born (OR 9.03, 95% CI 4.56-17.87), Black Caribbean ethnic group (OR 5.68, 95% CI 2.96-10.91), >65 years old (OR 0.25, 95% CI 0.09-0.67), and resident in Wolverhampton (OR 9.29, 95% CI 5.69-15.19), were significantly associated with the Mercian strain. A significant negative association of the Mercian strain with patient age >65 years old was identified. Therefore, age <65 years old is positively associated with the Mercian strain.

Variable	No. patients		Unadjusted			Adjusted		
	Mercian (n=124)	WT (n=2,066)	Odds Ratio	95% CI	p	Odds Ratio	95% CI	p
Gender								
Male	77	1,118	1.38	0.95-2.01	0.09	1.03	0.65-1.62	0.91
Female	47	944	1.00	1.00	Reference			
Age group								
0-14	4	45	1.27	0.45-3.59	0.69	0.35	0.07-1.63	0.18
15-44	92	1310	1.00	1.00	Reference			
45-64	23	379	0.86	0.54-1.38	0.543	0.77	0.42-1.41	0.40
>65	5	330	0.22	0.09-0.53	<0.01*	0.25	0.09-0.67	<0.01*
HPU location in West Midlands								
East	47	1,126	1.00	1.00	Reference			
North	10	206	1.16	0.58-2.34	0.67			
West	67	727	2.21	1.50-3.24	<0.01*			
Local Authority								
Wolverhampton	51	169	7.83	5.30-11.58	<0.01*	9.29	5.69-15.19	<0.01*
32 other Local Authorities	73	1,895	1.00	1.00	Reference			
Place of birth								
UK-born	100	546	18.03	10.22-31.81	<0.01*	9.03	4.56-17.87	<0.01*
Non-UK-born	14	1,378	1.00	1.00	Reference			
Ethnic group								
Black Caribbean	32	70	14.84	8.55-25.77	<0.01*	5.68	2.96-10.91	<0.01*
Black African	1	362	0.09	0.01-0.68	0.02	0.19	0.02-1.51	0.12
Indian Subcontinent	33	1,102	1.00	1.00	Reference			
Other	6	105	1.57	0.60-4.10	0.36	1.11	0.37-3.37	0.85
White	49	366	4.47	2.79-7.15	<0.01*	1.75	0.95-3.22	0.07

Table 6.4. Statistical analysis of sociodemographic data for patients infected with the Mercian strain.

The wild-type (WT) control group consisted of tuberculosis patients infected with *M. tuberculosis* strains other than the Mercian strain (n = 2,066) in the West Midlands between 2004 and 2008. *P-values were considered as statistically significant if <0.05. Significant unadjusted values were included in the multivariate model.

Variable	No. patients		Unadjusted			Adjusted		
	Mercian (n=124)	WT (n=2,190)	Odds Ratio	95% CI	P	Odds Ratio	95% CI	p
Site of disease								
Pulmonary sputum smear positive	59	655	1.00	1.00	Reference			
Pulmonary sputum smear other	47	711	0.73	0.49-1.09	0.13	1.06	0.64-1.74	0.83
Extra pulmonary	18	685	0.29	0.17-0.50	<0.01*	0.62	0.33-1.18	0.15
Clinical history of TB								
Previous diagnosis of TB	10	85	0.60	0.30-1.20	0.15			
No previous diagnosis of TB	82	1,164	1.00	1.00	Reference			
Treatment								
Patient admitted as in-patient	25	329	1.04	0.64-1.65	0.89			
Patient admitted as out-patient	58	737	1.00	1.00	Reference			
Treatment outcome at 12 months								
Treatment completed	71	1,157	1.00	1.00	Reference			
Died	3	108	0.45	0.14-1.46	0.19			
Lost to follow up	4	72	0.91	0.32-2.55	0.85			
Still on treatment	2	32	1.02	0.24-4.34	0.98			
Treatment stopped	1	6	2.72	0.32-22.87	0.36			
Transferred out	0	18	-	-	-			
Not completed unknown	2	21	1.55	0.36-6.75	0.56			
Unknown	2	13	2.51	0.56-11.32	0.23			
Treatment outcome at 12 months								
Successful	73	1,189	0.89	0.47-1.66	0.71			
Not successful	12	220	1.00	1.00	Reference			
Drug Sensitivity Testing								
Resistance to any 1st line drug	5	93	0.89	0.35-2.22	0.79			
No resistance to any 1st line drug	119	1,959	1.00	1.00	Reference			
MDR	1	15	1.10	0.14-8.43	0.92			
Not MDR	123	2,037	1.00	1.00	Reference			

Table 6.5. Statistical analysis of clinical and bacteriological data for patients infected with the Mercian strain across the West Midlands.

6.4.11. Epidemiological investigation in the city with the highest proportion of the Mercian strain

The Mercian strain in Wolverhampton was significantly associated with white UK-born patients who presented with cavitations on chest X-ray and produced smear positive specimens (Table 6.6 and Table 6.7). Patients with the Mercian strain continued to experience weight loss at eight weeks after being started on anti-tubercular chemotherapy ($p<0.05$). However, there was no significant difference between treatment completion rates after 12 months.

Examination of the epidemiological factors revealed that cases with the Mercian strain were more likely to have a previous history of TB (9/35, 26%), and would have had significant previous contact with a case of TB (24/35, 69%), and in particular patients with the Mercian strain (13/35, 37%). Significant social factors detected were evidence of excess alcohol intake (OR 8.78, 95%CI 2.42-41.01, $p<0.01$) and cannabis use (OR 10.02, 95%CI 1.96-100.33, $p<0.01$).

Variable	Mercian (n=35)	WT (n=47)	Unadjusted			Adjusted		
			Odds Ratio	95% CI	p	Odds Ratio	95% CI	p
Patient gender								
Female	18	20	1.42	0.54-3.77	0.43			
Male	17	27	1.00	1.00	Reference			
Age group								
0-14	1	1	1.15	0.01-92.99	1.00			
15-44	27	31	1.00	1.00	Reference			
45-64	4	10	0.46	0.10-1.85	0.22			
>65	3	5	0.69	0.10-3.95	0.72			
Ethnic group								
Indian Subcontinent	11	23	1.00	1.00	Reference			
Black African	0	6	-	-	-			
Black Caribbean	9	6	3.06	0.75-13.46	0.07			
Other	1	5	0.44	0.01-4.71	0.65			
White	14	7	4.06	1.15-15.77	0.01*			
Country of birth								
UK-born	32	12	29.42	7.32-177.18	<0.01*	9.68	2.00-46.78	<0.01*
Non-UK-born	3	35	1.00	1.00	Reference			
Epidemiological History								
Previous contact with TB case	24	11	6.94	2.42-21.50	<0.01*	3.40	0.92-12.59	0.07
No previous contact with TB case	11	36	1.00	1.00	Reference			
Previous contact with Mercian strain	13	0	15.65	4.76-51.48	<0.01*			
No previous contact with Mercian strain	13	46	1.00	1.00	Reference			
Previous history of TB	9	3	4.97	1.11-31.13	0.01*			
No previous history of TB	26	44	1.00	1.00	Reference			
Clinical co-factors								
Malignancy	1	0	10.41	0.20-547.59	0.43*			
No evidence of malignancy	34	47	1.00	1.00	Reference			
Diabetes	2	7	0.35	0.03-2.01	0.29			
No evidence of diabetes	33	40	1.00	1.00	Reference			
Social factors								
Excess alcohol use	16	4	8.78	2.42-41.01	<0.01*	6.26**	1.45-27.02	0.01
No evidence of excess alcohol use	19	43	1.00	1.00	Reference			
Cigarette smoking	6	3	2.99	0.58-19.96	0.12			
Non-smoker	29	44	1.00	1.00	Reference			
Cannabis use	11	2	10.02	1.96-100.33	<0.01*			
No evidence of cannabis use	24	45	1.00	1.00	Reference			
Employed	14	24	1.00	1.00	Reference			
Unemployed	21	23	1.56	0.59-4.18	0.32			

Table 6.6. Epidemiological associations of patients infected with the Mercian strain in Wolverhampton.

P-values were considered as statistically significant if ≤0.05. **OR for excess alcohol use and cannabis use was combined.

Variable	Mercian (n=35)	WT (n=47)	Unadjusted			Adjusted		
			Odds Ratio	95% CI	p	Odds Ratio	95% CI	p
Clinical presentation								
Cavitary disease	22	12	4.83	1.74-14.24	<0.01*	1.57	0.40-6.17	0.52
Non-cavitary disease	13	35	1.00	1.00	Reference			
Pulmonary disease	31	34	2.93	0.79-13.64	0.07			
Non-pulmonary disease	4	13	1.00	1.00	Reference			
Sputum specimen	25	24	2.37	0.87-6.84	0.06			
No sputum specimen	10	23	1.00	1.00	Reference			
Positive microscopy	22	17	2.94	1.10-8.19	0.02*			
Negative microscopy	13	30	1.00	1.00	Reference			
Treatment								
Therapy adherence	28	44	1.00	1.00	Reference			
Therapy non-adherence	7	3	3.61	0.75-23.42	0.06			
No weight loss after initiation of treatment	27	47	1.00	1.00	Reference			
Weight loss after initiation of treatment	8	0	12.99	3.00-56.28	<0.01*			
Completed treatment within 12 months	26	36	0.64	0.17-2.29	0.43			
Did not complete treatment within 12 months	8	7	1.00	1.00	Reference			
Drug Sensitivity Testing								
Resistance to any 1st line drug	0	1	-	-	-			
No resistance to any 1st line drug	35	46	1.00	1.00	Reference			
MDR	0	0	-	-	-			
Not MDR	35	47	1.00	1.00	Reference			

Table 6.7. Clinical and bacteriological associations of patients infected with the Mercian strain in Wolverhampton.

P-values were considered as statistically significant if ≤0.05.

6.4.12. Additional genotyping by spoligotyping and MIRU-VNTR using the internationally optimised set of 24 loci

A selection of 201 strains identified as the Mercian strain between 1996 and 2008 were analysed by spoligotyping and nine additional MIRU-VNTR loci to complete the recently described optimal set of 24 loci. This was to identify which global lineage the Mercian strain is a member of and to confirm or refute presence of a clone by the current internationally recommended set of MIRU-VNTR loci. When compared to the international spoligotyping database (SpoldDB4), the predominant spoligotype clade for the Mercian strain (189/201, 94%) was identified as ST490 (octal code 767776777760771). This shared typed is a member of clade X1. There are 13 strains with this spoligotype in the SpolDB4 database, of which nine originate from our initial studies and two originate from New York with one strain each from London and Washington (date accessed 5th December 2010). Of the 201 strains analysed by the complete 24 loci set, 170 (85%) were still indistinguishable (extra nine loci profile 434443183). The other 31 isolates contained two single locus variants (23 and five isolates), two unique isolates and a double locus variant. The first variant profile was identified in a strain originating from 2000 with subsequent variants identified in 2002, 2005, and 2006 (Gibson *et al.*, 2010).

6.5. DISCUSSION

This chapter describes the identification of the most prevalent *M. tuberculosis* strain in the West Midlands, which we have termed the Mercian strain. Concordant MIRU-VNTR and RFLP data from six different geographical locations across the West Midlands indicated that this strain has continued to expand in three major cities. Regional, national, and global databases provided evidence that this strain was restricted to the West Midlands region in England. Regional epidemiological data showed that this strain primarily infected UK-born, Black Caribbean patients less than 65 years old.

6.5.1. Significant epidemiological associations have been identified but the reasons for regional transmission are still to be determined

The regional and Wolverhampton epidemiological investigations presented in this report identified significant associations for the Mercian strain. However, they do not provide a full explanation as to why the Mercian strain is more prevalent when compared to other strains in the West Midlands. Drug and alcohol use were identified as significant social factors in Wolverhampton. These two risk factors have been identified as significant associations in previously reported tuberculosis outbreaks particularly in low-incidence countries (Jackson *et al.*, 2009;Fok *et al.*, 2008;Oeltmann *et al.*, 2006;Perlman *et al.*, 1997). The cumulative number of cases and continuing presence of the Mercian strain does not follow a typical point-source outbreak pattern. The significant association with younger age suggests that cases caused by the Mercian strain have arisen as a result of recent transmission and not re-activation in older patients. A possible transmission scenario is that after the initial emergence of the Mercian strain there have been several independent clusters of transmission each with their own common social link. This has resulted in a large, complex transmission network that is still expanding and is yet to be fully elucidated.

6.5.2. The Mercian strain has been present in the Midlands since at least 1995

Retrospective epidemiological studies have identified the earliest isolate of the Mercian strain from 1995 in an archive strain collection. This isolate was part of a cluster of 11 isoniazid resistant strains identified between 1995 and 2000 which was reported previously before the full regional extent of the Mercian strain was known (Chapter 3) (Hawkey *et al.*, 2003). We have typed very few archived *M. tuberculosis* strains from 1995 so the full extent of drug sensitive and drug resistant Mercian strains 15 years ago has not been assessed. The cluster of isoniazid resistant Mercian strains was present in one specific location. Between 1995 and 2000, there was no other investigated instance of increased isoniazid resistance in the rest of the West Midlands.

It is likely that the Mercian strain first emerged in the West Midlands well before 2004 and perhaps even before 1995. The Mercian strain has been present since prospective DNA fingerprinting commenced in 2004 with a median of 30 isolates per year (range 27-37), and has represented a consistent proportion of all strains (Figure 6.1),

6.5.3. Wolverhampton is a geographical focus for the Mercian strain

The distribution of *M. tuberculosis* strains in Wolverhampton was significantly altered by the Mercian strain. Compared to all other regions, Wolverhampton had a reduced level of strain diversity and an increased rate of clustering. The calculation of clustering rates using the “classical” 15 locus MIRU-VNTR should be analysed with caution as has been discussed previously in this thesis (Chapter 5, Section 5.5.13). The Wolverhampton epidemiological investigation applied a detailed questionnaire that was only used in this location. Patients with the Mercian strain in Birmingham and Coventry might differ in their use of drugs and alcohol. The results from the Wolverhampton and region-wide analysis do not concord exactly as different ethnic population groups were identified as at highest risk: the White population in Wolverhampton but the Black Caribbean group across the West Midlands.

6.5.4. The origin of the Mercian strain

There are three possible scenarios for the emergence of the Mercian strain in the West Midlands. The first scenario is that it is an endemic strain that has been present in the West Midlands for decades perhaps even centuries. The other two scenarios are based on the epidemiological association with Black Caribbean patients. The second scenario is that the strain has been recently imported from the Caribbean to the West Midlands by human migration in the 20th century. A further extension of the second scenario is that the Mercian strain spread from the UK to the Caribbean centuries ago and then back to the UK again more recently. The clonality of the Mercian strain even after 24 loci and IS*6110* RFLP analysis suggests that there was a single original common source for this strain.

The Mercian strain is significantly associated with UK-born Black Caribbean patients. The fact that the Mercian strain is associated with UK-born patients may provide some evidence for the timeline of emergence of this strain. The Mercian strain is not associated with Black Caribbean patients not born in the UK which could represent either recent transmission into UK-born patients or transmission between Caribbean-born and UK-born descendents in the early part of the 20th Century.

Across the UK, there are other areas that have a high proportion of people who originate from the Caribbean. This population groups accounts for more than ten per cent of the total population in the London boroughs of Lewisham, Lambeth, Brent and Hackney (Office for National Statistics, 2003). Yet, only three *M. tuberculosis* isolates typed in London had the same indistinguishable MIRU-VNTR profile as the Mercian strain. In England in 2009, only 160/8,064 (2%) of all TB cases were patients identified as Black Caribbean. The relative paucity of the Mercian strain in other areas of the UK indicates that the Mercian has expanded by successful recent transmission in the Midlands and not by importation from a common country outside of the UK.

6.5.5. Identification of Clade X and location in the global phylogeny of *M. tuberculosis*

The Mercian strain is a member of the Clade X global spoligotype family. A characteristic spoligotype signature for Clade X was first identified by data mining of the third international spoligotyping database (SpolDB3) that contained 13,008 strains from >90 countries (Sebban *et al.*, 2002). Clade X was identified as a family of strains that do not have spacer 18 or spacers 33-36. Three subtypes have been identified: X1 (spacer 18 and spacers 33-36 absent); X2 (as X1 plus spacers 39-42 deleted as well); and X3 (as X1 with spacers 4-12 deleted) (Filliol *et al.*, 2002). IS*6110* typing has shown that clade X strains are usually considered as strains containing low-copy numbers of *IS6110* (Warren *et al.*, 2004;Soini *et al.*, 2000).

The global distribution of Clade X has been associated with English-speaking countries and countries with a past history of colonization by Anglo-Saxon populations (Sebban *et al.*, 2002;Soini *et al.*, 2000). Clade X is most prevalent in North America (14% of all strains)

(Filliol *et al.*, 2002), South Africa (19%) (Stavrum *et al.*, 2009a), Southwest Ireland (26.2%) (Ojo *et al.*, 2010), and London (9% of UK-born patients) (Brown *et al.*, 2010).

The distribution of spoligotypes on Caribbean islands have been intensively studied and it was revealed that clade X is present on both French- and English-speaking islands including Guadeloupe (9.3%), Martinique (3.4%) and continental French Guiana (7.6%) (Brudey *et al.*, 2006;Duchene *et al.*, 2004). Trinidad & Tobago is the only English-speaking Caribbean island investigated so far. A novel spoligotype was identified that is a member of Clade X and accounted for 56% of all strains (Baboolal *et al.*, 2009;Millet *et al.*, 2009).

Clade X has been rarely detected in Southern (Italy) and Eastern Europe (Poland) (Lari *et al.*, 2007;Jagielski *et al.*, 2010); Saudi Arabia and Turkey in Western Asia (Al-Hajoj *et al.*, 2007;Otlu *et al.*, 2009); India in Southern Asia (Gutierrez *et al.*, 2006;Kulkarni *et al.*, 2005;Singh *et al.*, 2004;Singh *et al.*, 2007); China in Eastern Asia (Dong *et al.*, 2010); Honduras and Paraguay in Central and South America (Candia *et al.*, 2007;Rosales *et al.*, 2010); Gambia in Western Africa (de Jong *et al.*, 2009); and Malawi(Glynn *et al.*, 2010), Mozambique (Viegas *et al.*, 2010), and Zambia in Eastern Africa (Mulenga *et al.*, 2010) with proportions ranging from 0.9-6.0%.

Clade X strains have been analysed by various phylogenetic techniques to locate their position in the global evolution of *M. tuberculosis*. Essentially, clade X is a modern clade of *M. tuberculosis* that is closely related to other modern strains such as Haarlem, LAM, and T but is a distinct spoligotype. Clade X is a member of: the Euro-American lineage based on LSPs (Gagneux *et al.*, 2006); PGG 2/3 (Sreevatsan *et al.*, 1997); Lineage 4 (Gagneux and Small, 2007); Lineage II (Baker *et al.*, 2004); SNP clusters III-VII (Gutacker *et al.*, 2006); and SNP group 3b-6b (Filliol *et al.*, 2006). The link between Anglo-Saxon history and the Caribbean fits in with a recent hypothesis that ancestral *M. tuberculosis* complex strains spread by land and modern strains spread by sea travel (Hershberg *et al.*, 2008).

6.5.6. Utility of universal prospective DNA fingerprinting

Detection of the Mercian strain was only possible with the commencement of universal prospective MIRU-VNTR typing of all *M. tuberculosis* isolates in the West and East Midlands. Only with universal prospective DNA fingerprinting was the full extent of the Mercian strain in the West Midlands fully characterized. It would only have been possible to detect the Mercian strain by obtaining the genotype as it is not a drug-resistant strain and there are no other phenotypic properties that could have easily differentiated it from other *M. tuberculosis* complex strains, The patient population that the Mercian strain has been identified in is different to the UK-wide situation for TB as the majority of patients diagnosed each year in the UK are not born in the UK and originate from the ISC (Health Protection Agency, 2010).

The 156 individual patients detected between 2004 and 2008, make the Mercian strain one of the largest known community-based clusters in the world. Previous major prevalent strains have been identified in New York (Alland *et al.*, 1994;Moss *et al.*, 1997), Rotterdam (de Vries *et al.*, 2007), North London (Ruddy *et al.*, 2004), and Rio de Janeiro (Lazzarini *et al.*, 2007).

6.5.7. Comparison to other reported prevalent strains in the UK

The most prevalent strain detected in the UK was an isoniazid resistant strain in North London that was previously reported in 70 patients (Ruddy *et al.*, 2004), with a current total of over 300 cases caused by this strain (Ibrahim Abubakar, Consultant Epidemiologist & TB Section Head, Respiratory Diseases Department - Tuberculosis Section, HPA, personal communication). The 24 locus MIRU-VNTR profile of this strain is 42433 2431515321 226423-52. Isoniazid resistance acted as a very useful marker for detection of the strain. It was noted by the outbreak investigators that without the drug resistance marker only prospective typing of all isolates would have detected this large, complex outbreak. The North London strain was predominantly found in young White or Black Caribbean UK-born adults with drug misuse a common epidemiological factor (Ruddy *et al.*, 2004). It is possible that patients in this population group take longer to present clinically as TB may not be suspected when initial symptoms develop or they might not seek medical help soon after onset. Both factors aid strain transmission and disease progression.

6.5.8. Future investigations

Both epidemiological investigations presented in this report were retrospective and did not involve direct patient interviews. The Mercian strain continues to be identified in the West Midlands. Enhanced epidemiological knowledge could be obtained by prospectively investigating social links as each new patient is diagnosed. Investigation of potential factors which may cause a delay in diagnosis should be investigated as well. The data presented identified the infected patient population and also important common social factors. The exact interaction of patient population and social factors should be investigated further to identify and fully understand any confounding factors.

6.5.9. Potential impact of mycobacterial genomic variation on strain transmission

A strain was identified in 93/314 (30%) patients in Rio de Janeiro that uniquely lacked a major region of genomic DNA (>26.3 kb). This region contained 10 genes including two potentially immunogenic PPE (proline-proline-glutamic acid) genes (Gibson *et al.*, 2008;Lazzarini *et al.*, 2007). This strain (RDRio) was associated with a higher frequency of cavitary pulmonary disease (Lazzarini *et al.*, 2008). The major deletion identified in the RDRio strain has been hypothesized as having a major impact on the virulence properties of the RDRio strain. As the genomic content of the Mercian strain has not been characterized, further work should determine whether such a deletion or other similar major genomic variation has altered the virulence of this strain leading to multiple transmission events in and between three cities in the West Midlands.

6.6. CONCLUSIONS

This chapter describes the identification of the most prevalent *M. tuberculosis* strain in the West Midlands region of the UK, with 156 isolates in a five year period between 2004 and 2008. The Mercian strain has been significantly associated with UK-born patients, appears to be geographically restricted to the West Midlands region in the UK with evidence of ongoing transmission and specific risk factors for acquisition that require further investigation.

7. WHOLE GENOME ANALYSIS SHOWS A COMMON EVOLUTIONARY ORIGIN FOR THE DOMINANT STRAINS OF *M. TUBERCULOSIS* IN A UK SOUTH ASIAN COMMUNITY.

7.1. INTRODUCTION

7.1.1. TB in the ISC and UK

One third of the world's population is infected by *M. tuberculosis*. The bacillus causes 1.6 million deaths each year and more than eight million new cases annually. India has the highest burden of TB of any single country in the world with two million new cases in 2009 with 420,000 new cases in Pakistan and 360,000 in Bangladesh. These three countries in the ISC account for 30% of the global burden of all new TB cases with India alone accounting for 21% (World Health Organisation, 2010b). Within the UK in 2009, the majority of TB patients with a known country of birth were not born in the UK (5,793/8,234, 70%). The most frequent country of birth in non-UK-born patients was India (1,615/5,793, 28%). Pakistan (982/5,793, 17%) and Bangladesh (247/5,793, 4%) were the second and fourth most frequent respectively. Collectively, patients born in these three countries accounted for 49% (2,844/5,793) of all TB cases in non-UK-born patients (Health Protection Agency, 2010).

7.1.2. The Central Asian Strain

This strain was initially described as a spoligotype in a UK study in 1997 (Goyal *et al.*, 1997) and the association with patients that originate from the ISC was first identified in a cohort of patients from South Asia (SA) who were resident in Leeds and Bradford where this clade was identified as the most prevalent strain. In all patients in Leeds and Bradford, 23% were infected with this clade and all patients infected with this clade originated from the ISC (Gascoyne-Binzi *et al.*, 2002).

The identification and definition of a distinctive spoligotype present in Southern Asia was concomitantly and independently recognised by two groups in 2002 (Bhanu *et al.*, 2002;

Filliol *et al.*, 2002). A study in Delhi that analysed 83 isolates by IS*6110* RFLP and spoligotyping identified that 60/ 83 (75%) isolates were at least 61% similar on IS*6110* RFLP typing with two common IS*6110*-containing *Pvu*II RFLP fragments present at 10.1 and 12.1 kb, and a distinctive set of spacer sequence patterns. These strains were named the Delhi type (Bhanu *et al.*, 2002). This strain was defined as a distinctive spoligotype clade by the absence of spacers 4-7 and 23-34 and was named CAS (Filliol *et al.*, 2002). Within the CAS clade, there are several global variants including CAS1-Delhi, CAS-Kili identified in Tanzania that has additional spacers absent between positions 8-22 (Kibiki *et al.*, 2007;McHugh *et al.*, 2005); CAS1-DAR in Dar es Salaam (Eldholm *et al.*, 2006); and a separate variant in Sudan (Sharaf-Eldin *et al.*, 2002). The initial spoligotype identified as the Delhi type is specifically spoligotype Shared Type (ST) 26. CAS encompasses ST26 and other shared spoligotypes, including further variants such as CAS2 (additional absent spacers 4-10 and 23-34) (Bhanu *et al.*, 2002;Filliol *et al.*, 2002). Subsequent analysis of *M. tuberculosis* strains by genomic frameworks based on SNP and LSPs have identified CAS as a major phylogenetic clade and have assigned it to various global lineages including the East-African-Indian lineage (Table 7.1) (Gagneux and Small, 2007).

By MIRU-VNTR typing, the CAS spoligotype has characteristic alleles at the five ETR loci (A-E) of 42234 or 42235 with some strains possessing a non-amplifiable ETR-A locus (Brown *et al.*, 2010;Brudey *et al.*, 2004;Cheah *et al.*, 2010;Freidlin *et al.*, 2009;Gutierrez *et al.*, 2006;Gutierrez *et al.*, 2006).

Scheme	Experimental Data	Lineage	Reference
TbD1	TbD1 deleted	Modern	Brosch et al. 2002
LSP	RD750	East African-Indian	Gagneux et al. 2006
SNP	*katG*463 CTG (Leu) *gyrA*95 ACC (Thr)	PGG 1	Sreevatsan et al. 1997
SNP	*rpoB* 2646 G *ahpC*-46 A	Lineage III	Baker et al. 2004
SNP	See reference	Cluster II.A	Gutacker et al. 2006
SNP	See reference	Cluster group 3a	Filliol et al. 2006
Spoligotyping	Spacers absent between 4-7 and 23-34.	CAS	Brudey et al. 2006
ETRs	-2234, -2235, 42234, 42235	South Asian strain	Gascoyne-Binzi et al. 2002 Cheah et al. 2010
MIRU-VNTR	ETR-C: 2 MIRU-23: 5	Lineage III/CAS	Gibson et al. 2005
MIRU-VNTR	See reference	CAS	Brown et al. 2010
IS*6110* RFLP	Two common bands at 10.1 and 12.1 kb.		Bhanu et al. 2002
	One of most variable IS*6110* groups.		Gutacker et al. 2006

Table 7.1. Molecular markers for the predominant *M. tuberculosis* clade present in Southern Asia.

Table adapted from Gagneux et al. 2007.

7.1.3. Global geographic distribution of CAS

The CAS spoligotype clade is predominantly associated with South Asia and to a lesser extent, the Middle-East. The fourth version of the international spoligotype database (SpolDB4) identified that 21.2% of all strains in South-Asia were CAS (Brudey *et al.*, 2006). One of the initial studies of strain distribution in Delhi identified that 75% of all strains were members of the CAS or Delhi-type clade (Bhanu *et al.*, 2002).

Intensive successive studies of *M. tuberculosis* strains in India have repeatedly identified CAS as the most prevalent clade in many cities and states with particular focus on New Delhi (Singh *et al.*, 2004;Stavrum *et al.*, 2009b) and Mumbai (Kulkarni *et al.*, 2005;Almeida *et al.*, 2005). There is a geographical gradient of strain distribution between North and South India. CAS is more predominant in the North of India with the ancestral EAI strain more common in the South (Gutierrez *et al.*, 2006;Singh *et al.*, 2007). There are instances of other clades present in India as there is a prevalent MDR Beijing strain in Mumbai (Almeida *et al.*, 2005). A third distinct clade, Manu, was identified as the most prevalent clade in Western India (Chatterjee *et al.*, 2010). In the rest of the ISC, CAS has also been identified as the most prevalent clade in Karachi, Pakistan (Hasan *et al.*, 2006;Tanveer *et al.*, 2008) but not in Bangladesh where Beijing and EAI predominate (Banu *et al.*, 2004;Rahim *et al.*, 2007). CAS has been frequently identified in immigrants originating from South Asia in other regions of the world (Sola *et al.*, 2001). CAS makes up approximately half of the *M. tuberculosis* isolates circulating in Leicester (Cheah *et al.*, 2010). In 2001, there was a large outbreak among people of South Asian origin in a school in Leicester with 70 cases of active tuberculosis and 254 cases of latent infection in a total population of 1,208 pupils (Rajakumar *et al.*, 2004). In a recent study of 2,261 isolates in London, CAS was the single most prevalent clade overall (552/2,261, 24%) and in patients originating from the ISC (203/463, 44%) (Brown *et al.*, 2010).

7.1.4. The post-genomic era in *M. tuberculosis*

From the determination of the complete genome sequence of *M. tuberculosis* H37Rv in 1998, a new era of comparative genomics was heralded. *M. tuberculosis* H37Rv contains approximately 4,000 ORFs (Cole *et al.*, 1998). However, the molecular basis of pathogenicity, virulence and transmissibility in *M. tuberculosis* is still not well understood as the function of only one-third these 4,000 predicted ORFs can be can be accurately predicted. Of the other ORFs, the function of another third can be predicted with some confidence; with little or nothing known about the function of the remaining third. Further determination of significant deleted regions in the genome of *M. tuberculosis* by DNA microarrays and whole-genome sequencing will help to reveal the basis for transmission and pathogenicity especially with regards to the interaction between host and pathogen (Barry, III *et al.*, 2000;Cole *et al.*, 1998).

7.1.5. PPE genes

One of the major gene families identified in *M. tuberculosis* are the PE and PPE gene glycine-rich families which make up 10% of the coding capacity of the genome. There are multiple copies of repetitive sequences (PGRS and MPTRs) in each family. The PE family contains Pro-Glu (PE) motifs and the PPE family contains Pro-Pro-Glu (PPE) motifs, both usually found near the N terminus (Cole *et al.*, 1998). It is hypothesised that both PE and PPE genes are involved in antigenic variation. One member (Rv1759) is a fibronectin-binding protein that elicits a variable antibody response (Abou-Zeid *et al.*, 1991) and it these genes are likely associated with the mycobacterial cell wall (Sampson *et al.*, 2001).

7.1.6. DNA microarrays: the technique

For DNA microarray analysis, oligonucleotides are designed to amplify each of the predicted ORFs in the reference sequenced genome. Each amplicon is then spotted onto a glass slide. Extracted DNA from the strain of interest and reference strain is extracted and fluorescently labelled with two different dyes. The labelled DNA from both strains is then hybridised to the microarray and scanned by laser excitation. The data obtained for each dye is then combined to identify regions that are present or absent in the strain of interest (Figure 7.1). “Deleted” regions are confirmed by PCR and DNA sequencing to determine the extent of the deleted region (Pym and Brosch, 2000). There are various types of DNA microarrays including: printed; in situ-synthesized oligonucleotide; high-density bead; electronic; and suspension bead arrays (Miller and Tang, 2009). The study described in this chapter used a printed microarray produced by The Bacterial Microarray Group at St George's Hospital (BμG@S), University of London.

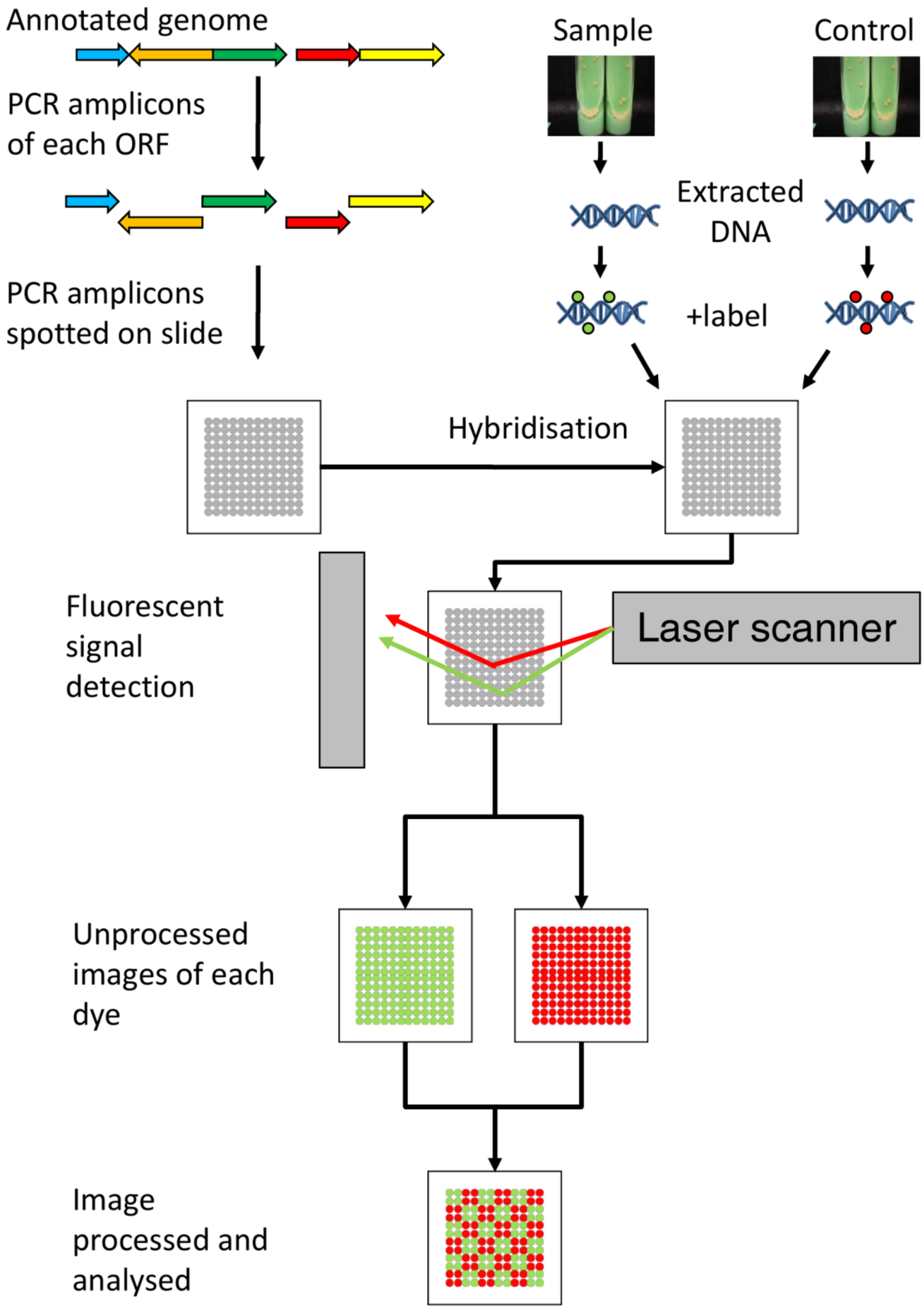

Annotated genome
PCR amplicons of each ORF
PCR amplicons spotted on slide
Sample
Control
Extracted DNA
+label
Hybridisation
Fluorescent signal detection
Laser scanner
Unprocessed images of each dye
Image processed and analysed

Figure 7.1. Specimen preparation and printed microarray analysis.

PCR amplicons are generated for each identified ORF in a genome and spotted onto a glass slide. DNA from the strain of interest and a suitable control strain are extracted, labelled and hybridised to the spotted microarray in a competitive hybridization. The presence or absence of ORF regions is then identified by a laser scanner, images combined and a list of deleted regions identified. Figure adapted from Miller and Tang, 2009.

7.1.7. Delineation of evolution in the *M. tuberculosis* complex using DNA microarray technology

Comparison of the *M. tuberculosis* H37Rv and *M. bovis* BCG vaccine strain genomes identified 14 sequences present in *M. tuberculosis* H37Rv but absent in *M. bovis* BCG. These were called regions of difference (RD1-14) and demonstrated the genomic changes involved in the production of the BCG vaccine strain (Gordon *et al.*, 1999). Similarly, six regions were identified, that were absent from the *M. tuberculosis* H37Rv genome relative to other members of the *M. tuberculosis* complex: H37Rv relative deletions (RvD1-5) and *M. tuberculosis* specific deletion 1 (TbD1) which divided global clades into “ancestral” and “modern” lineages. These deletions provided evidence that *M. africanum*, *M. microti*, and *M. bovis* diverged from the progenitor of *M. tuberculosis* before a deletion in TbD1 occurred which altered the previous hypothesis that *M. tuberculosis* had evolved from *M. bovis*. (Gordon *et al.*, 1999;Brosch *et al.*, 2002). A global collection of BCG vaccine strains were also analysed by DNA microarrays which demonstrated the evolution of BCG strains since the original passaged strain (Behr *et al.*, 1999). This study of BCG strains in comparison to H37Rv identified 14 regions (RD1-14) absent in BCG Pasteur but present in H37Rv with two additional deletions (RD15 and 16) in specific BCG substrains (Behr *et al.*, 1999). DNA-DNA microarrays have also been used to map IS*6110* insertion sites to specific regions or genes and the implications for strain virulence examined (Kivi *et al.*, 2002**.** Evolutionary studies have been undertaken to characterize genomic deletions and determine the probable evolution in a large number of different clinical isolates of *M. tuberculosis.* This found that deletions in clinical strains are not randomly distributed and tend to be aggregated in related lineages but that there were also deletions found in unrelated strains in regions apparently vulnerable to disruption (Tsolaki *et al.*, 2004).

A comparative study of 100 *M. tuberculosis* strains revealed that a patient's region of origin is predictive of the strain that the patient is infected with even when transmission has occurred in a geographic location outside the region of the patient's origin. Two lineages that originated from South-East Asia diverged centuries ago and have remained distinct lineages which suggests that the associations between host and pathogen are highly stable (Hirsh *et al.*, 2004).

DNA-DNA microarrays have been used to assess the varying spectrum of virulence in a virulent strain (H37Rv) and a member of the *M. tuberculosis* complex that is less virulent in humans (*M. microti).* Heterogeneous deletions were identified which make it difficult to assign virulence properties to any specific pattern of deletion. Perhaps the most significant factors being that ESAT-6 antigens and PE/PPE proteins are frequently deleted in *M. microti* (Garcia-Pelayo *et al.*, 2004;Frota *et al.*, 2004) .

Some of the genomic deletions of strains of *M. tuberculosis* studied were identified as useful markers for defining the Beijing/W family of strains (Tsolaki *et al.*, 2005) and six distinct global lineages of *M. tuberculosis* (Gagneux *et al.*, 2006). Genomic deletions generally behave as being unique event polymorphisms that once generated are not regained. Therefore, genomic deletions can be used to construct robust, unambiguous phylogenies.

7.1.8. Gene expression profiling microarrays

In expression microarrays, two populations of cDNA derived from RNA extracted from two *M. tuberculosis* strains exposed to two contrasting environments are labelled with two different fluorochromes. The two cDNA populations are then compared by hybridisation. The values for each of the two populations are then compared to calculate the relative degree of expression or repression for each ORF (Barry, III *et al.*, 2000).

Whole-genome microarrays were first used in *M. tuberculosis* for the definition of gene expression responses to isoniazid and ethambutol (Wilson *et al.*, 1999). Subsequent expression microarray studies have analysed the response of *M. tuberculosis* to various environmental conditions and other antibiotics (Betts *et al.*, 2003;Rengarajan *et al.*, 2005;Voskuil *et al.*, 2003); and survival within the host (Sassetti *et al.*, 2003). Differences in human gene expression between infection and active disease have also been identified. Genes expressed by monocytes involved in antimicrobial defence, inflammation, chemotaxis, and intracellular trafficking were significantly over-expressed with a minimal set of genes that included lactoferrin, CD64, and a GTPase were sufficient to distinguish between tuberculosis patients, *M. tuberculosis*-infected health donors and non-infected healthy donors. (Jacobsen *et al.*, 2007).

The first large-scale description of the human transcriptional signature of TB infection discovered a 393-gene transcript expression signature. This was dominated by IFN-induced genes that are specific to individuals with active TB disease when compared to other bacterial or inflammatory diseases and correlated with radiographic findings. This signature may also be able to identify latently infected individuals who will develop active TB, as 10-25% of latently infected patients had similar transcriptional profiles to those with active disease. The transcriptional signature decreased after two months of treatment of active disease and reverted to being indistinguishable from health controls after completion of treatment. Therefore, the transcriptional signature could be used to monitor patient response to infection and treatment. The largest collection of expressed transcripts in active tuberculosis were those induced by type I IFNs or IFN-y which were produced by neutrophils and monocytes (Berry *et al.*, 2010).

7.1.9. Microbial diagnostic microarrays

Microarrays have been developed to identify a broad range of pathogenic bacteria including *M. tuberculosis* using *gyrB* and speciation across the genus (Kostic *et al.*, 2007;Park *et al.*, 2005). Microarrays have been developed to identify mutations that confer drug resistance to rifampicin, isoniazid, and pyrazinamide (Wade *et al.*, 2004;Park *et al.*, 2006). Five genomic deletions were identified by DNA-DNA microarray analysis in the specific strain of CAS that caused a large outbreak in Leicester in 2001. These five deletions were restricted to the outbreak strain. All strains within Leicester were analysed for the presence of these five deletions on a local level to identify outbreak strains using a Genome Level-Informed PCR (Rajakumar *et al.*, 2004). A similar approach was applied to the investigation of a TB outbreak in Seattle (Freeman *et al.*, 2005). A high-throughput membrane-based hybridisation method (deligotyping) that uses a similar principle to spoligotyping has been developed to analyse LSPs (Goguet de la Salmoniere YO *et al.*, 2004). Comparative genomic hybridisation has also been used to examine a prevalent strain family in Quebec. Two example clinical isolates were analysed by microarray and identified deletions were confirmed in 302 clinical strains which identified a 11.4 kb "DS6$^{\text{Quebec}}$" deleted region in 143 strains. (Nguyen *et al.*, 2004).

7.2. HYPOTHESIS & AIMS

In comparison to the Beijing strain (Parwati *et al.*, 2010), little is known about the selective advantages that CAS possesses which has enabled it to become the most predominant clade in the ISC. Also, global strains exhibit a stable association with the global continental origin of the host (Gagneux *et al.*, 2006). Recent studies in San Francisco suggested that patients tend to become infected with strains of *M. tuberculosis* that have similar origin to those associated with their region of birth (Hirsh *et al.*, 2004).

The study of genetic variability within natural populations of pathogens can provide insight into their evolution and pathogenesis with comparative genomics a powerful tool providing important data that can be used to control transmission, and complements the more extensive studies of variation using other sequences such as IS*6110* (van Soolingen *et al.*, 1994;Shamputa *et al.*, 2004).

Therefore, it was hypothesized that microbial determinants are likely to play an important role in the specificity of strains with certain patient populations. To further understand the host-pathogen interaction between host and strains originating from the ISC a whole genome analysis of strains originating from the ISC was carried out. Analysis of the whole genome content of this strain will provide further insights into the evolution and biological properties of this strain.

The aim of this chapter was to obtain a further understanding of the diversity, origin, and biological properties of the CAS strain from patients originating from the ISC in the UK, by analysing a selection of 10 *M. tuberculosis* strains by whole-genome DNA-DNA microarray analysis.

This particular group of strains were not related epidemiologically to those in the outbreak in Leicester despite most of the strains having an indistinguishable VNTR profile 42235 (Rajakumar *et al.*, 2004).

7.3. METHODS

7.3.1. Selected strains

Strains were selected on the basis on the five locus ETR profile (Section 2.6.1). Clinical strains 8088, 9375, 9866, and isolate 6947 were isolated from patients who originated from South Asia and are now resident in Leeds and Bradford. The other clinical strains (0135, 2566, and 3242) from South Asian patients and two other strains (1339 and 7009) were from Birmingham. The VNTR profiles of these clinical strains are shown in Table 7.2. To validate the comparison of strains causing disease in a mixed population, additional clinical strains representing VNTR profiles that have been associated with patients from different ethnic origins were also studied. This included VNTR profile 32433 (strain 6947 from a South Asian patient) and VNTR profile 32333. These two VNTR profiles represent strains that are not members of the CAS spoligotype clade. Profile 32333 is strongly associated with the European Haarlem spoligotype clade (Sola *et al.*, 2005). The South Indian (SI) clinical isolate (TMC120) ATCC 35811 was also included in this study with *M. tuberculosis* H37Rv used as the reference strain.

Strain designation	ETR	Global origin of patient	Originating laboratory
8088	42235	South Asia	Leeds, Yorkshire
0135	42235	South Asia	Birmingham, West Midlands
2566	42235	South Asia	Birmingham, West Midlands
3242	42235	South Asia	Birmingham, West Midlands
9375	02235	South Asia	Leeds, Yorkshire
9866	42234	South Asia	Leeds, Yorkshire
SI (TMC 120)	-	South India	South India
6947	32433	Europe	Leeds, Yorkshire
1339	32333	Europe	Birmingham, West Midlands
7009	32333	Europe	Birmingham, West Midlands

Table 7.2. Strain collection analysed by DNA microarray hybridisation.

7.3.2. DNA microarray analysis and confirmation of deleted regions

Genomic DNA extraction and microarray hybridization procedures were performed as previously described (Stewart *et al.*, 2002;Frota *et al.*, 2004) (Section 2.10.1). Briefly, 2-3 mg of DNA was labelled by incorporation of Cy3 and Cy5 dCTP during DNA polymerisation (Section 2.10.2). DNA microarray hybridisation and detection was carried out as described in Section 2.10.

Genes were only considered to be deleted when the p-value was <0.05. Deleted regions were compared to those published for the Beijing/W family (Tsolaki *et al.*, 2005;Stavrum *et al.*, 2008), strain CH (the Leicester outbreak index isolate) (Rajakumar *et al.*, 2004) and those reported by Tsolaki and colleagues (Tsolaki *et al.*, 2004). The function of each deleted region was identified by comparison to the TubercuList Web Server (http://genolist.pasteur.fr/TubercuList/). Clustering of disrupted ORFs was analysed by importation of categorical data into BioNumerics (Applied Maths, St Martens-Latem, Belgium).

7.3.3. Confirmation of deletions by PCR and sequencing

PCR and DNA sequencing was used to confirm the relevant deletions (Section 2.10.7 and oligonucleotides listed in Table 7.3).

Gene or genomic region affected by deletion	Primer	Sequence (5'-3')
Rv1519	CM1518F CM1520R CM1518Fb CM1520Rb	TTCTCACCTGGTTGATCGTG GTCCAGTAATCGTCGCCTTC CGTTTTGAGGATCCCAGTGT GGAATGCCAAATACCGTGAG
RD3 region*	RD3 intF RD3 intR	TTATCTTGGCGTTGACGATG CATATAAGGGTGCCCGCTAC
plcD/*cut*1 and Rvd2 region	CM1755exF CM1758R ORF3Rvd3F ORF1Rvd2R	CAGTTCGCTGATGTGACGAT ATTGCCTCCGCTAGAACAGA GATTGCGTTTGTTTTGCTGA TGGTCGCAGTGTTTCCAATA
Rv1917c	CM1916F CM1917exR CM1917R CM1917F	ATGACCCTGATCCACCTCTG TCGATTCCTAAAGCGGCTAA CCACCAGAGATCAACTC CGCCACTGTTGAAGAAG
Rv3017c→22c region	CM3017exF CM3022exR CM3022R CM3019exF IS-R ISup ISdown ISdw	TGGCTGTTCGTCAGTAGGTG GGAACCTTCACTCGTACACCA TTGCAGAGTGCGGTGGGGTTT CGCTAGCGGAATCAATGTG AGTTTGGTCATCAGCCGTTC TACCTCCTCGATGAACCACC CTCTACCAGTACTGCGGCGACG CTGCCTACTACGCTCAAC
Rv3135	CM3135F CM3136R	CATATCGCTTGACCCACAGA TCGCTGTTTGCTGTGTCTTT
Rv3516-17	CM3515F CM3518R CM3515Fc CM3515Rb	CCTTGTGTTTGTGGATCGTG TTCGCATGTGTCTCAAGAGG ACCTTGTCGTCCTTTTGCAC CGAAATCCAAACAGCCACTT
Rv3738c-39c	CM3737F CM3739exR	GAGTTCCTCGCCTCACCAT TCAGTTGACTGACCGGCTTT

Table 7.3. Oligonucleotides used for confirmation by PCR and DNA sequencing of deletions detected by DNA microarrays.

7.4. RESULTS

7.4.1. Global lineage assignation

Deletions were identified in this study that assigned these strains to one of the six main phylogenetic global lineages previously described (Table 7.4) (Gagneux *et al.*, 2006). All six of the SA-strains possessed a deleted Rv1519 (LSP RD750) which assigned these strains to the East-African-Indian lineage. The South Indian clinical isolate ATCC 35811 was assigned to the Indo-Oceanic lineage by a deletion in Rv3651 (LSP RD239) which includes ‘‘ancestral’’ strains (Brosch *et al.*, 2002). The three non-SA strains were all assigned to the Euro-American lineage by virtue of deleted Rv2270-80 (LSP RD182) in strains 1339 and 7009 and deleted Rv2313c-15c (LSP RD183) in strain 6947.

Strain designation	ETR	Global origin of patient	Lineage assigning deletion	Global lineage
8088	42235	South Asia	Rv1519 (RD750)	East-African-Indian
0135	42235	South Asia	Rv1519 (RD750)	East-African-Indian
2566	42235	South Asia	Rv1519 (RD750)	East-African-Indian
3242	42235	South Asia	Rv1519 (RD750)	East-African-Indian
9375	02235	South Asia	Rv1519 (RD750)	East-African-Indian
9866	42234	South Asia	Rv1519 (RD750)	East-African-Indian
SI (TMC 120)	-	South India	Rv3651 (RD239)	Indo-Oceanic
6947	32433	Europe	Rv2313c-15c (RD183)	Euro-American
1339	32333	Europe	Rv2270-80 (RD182)	Euro-American
7009	32333	Europe	Rv2270-80 (RD182)	Euro-American

Table 7.4. Assignation of the strains in this collection to previously defined global lineages.

Strains were assigned to the LSP framework previous described (Gagneux *et al.*, 2006).

7.4.2. Microarray studies of clinical strains

The total number of deletions in the analysed SA-strains ranged from 10 to 24 deletions. The median number of deleted genes was 12 (range 10-14) when mobile elements were not considered. Only one South Asian strain (8088) had an RD3 region deletion (prophage phiRV1) accounting for 14 genes. Excluding the RD3 deletion in strain 8088, strain 9375 had more deletions (14 genes) than any other SA-strain (Table 7.5). The South Indian reference strain (ATCC 35811) and strain 6947 possessed a high number of deletions with 51 and 58 deleted genes respectively. However, most of the deletions were disruptions caused by mobile elements (Figure 7.2). Deleted regions were confirmed by PCR and partial or total sequencing in the SA-strains 8088, 9375 and 9866. The validation of the microarray results was extended to the rest of the strains of the SA group (0135, 2566 and 3242) to confirm commonly deleted genes.

Construction of a clustering dendrogram based on deleted regions revealed that the deletion profiles of the South Asian strains are clustered together when compared to the other non-South Asian strains and that there are two families of disrupted genes within this study (Figure 7.3). The first group of deletions is variable between isolates and contains genes that are deleted in most of the South Asian strains. The second group is much larger and contains two groups of deletions; ones caused by an insertion/mobile element only and non-IS mediated deletions.

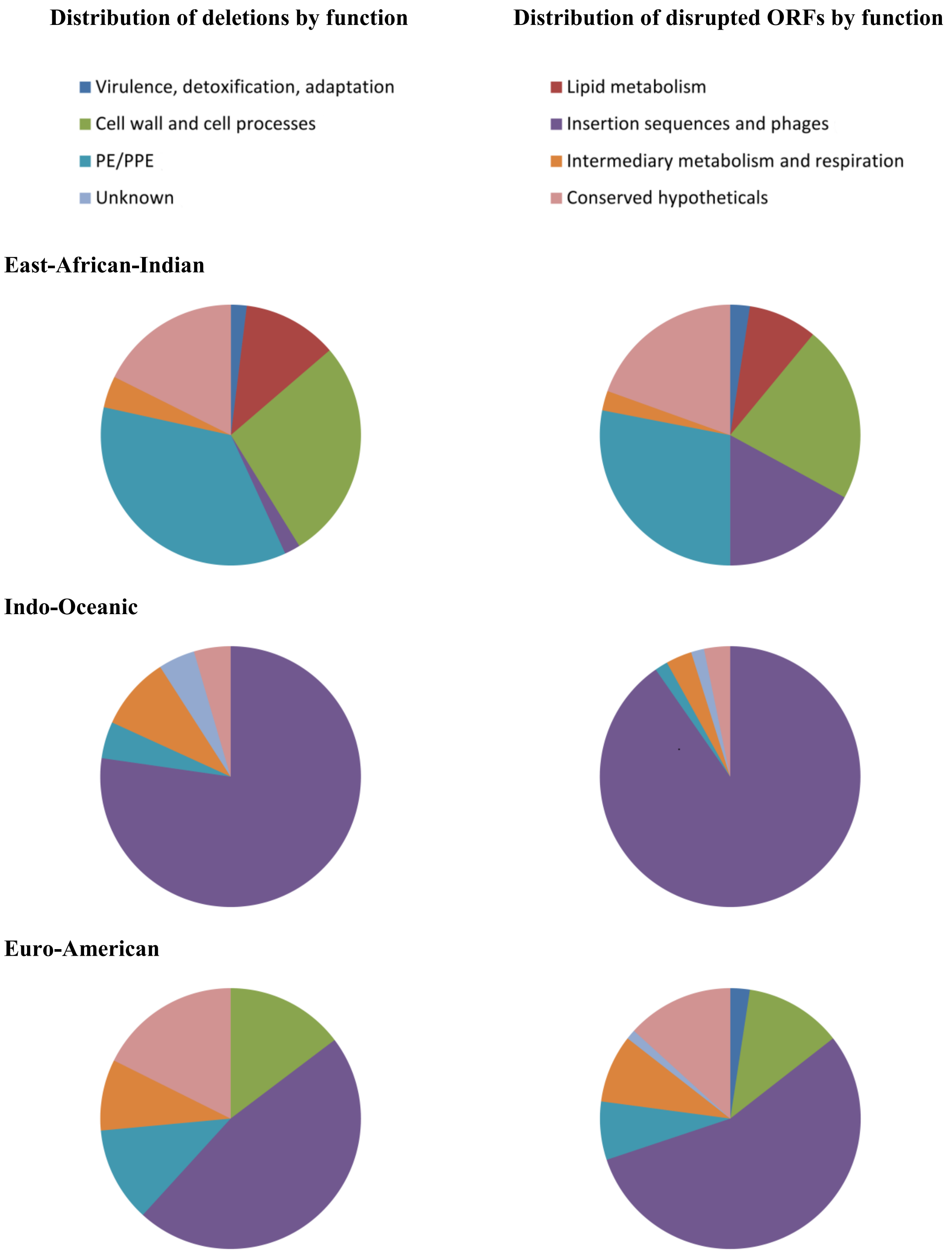


Figure 7.2 Proportions of deleted regions and disrupted ORFs by functional class.

Functional categories are derived from Tuberculist. Figure A details the number and proportion of individual identified deletions. Deleted ORFs (B) are the total numbers of deleted ORFs such that one deleted region (A) can disrupt multiple ORFs (B).

Global Lineage	East-African-Indian						Indo-Oceanic	Euro-American		
VNTR profile	42235				02235	42234	-	32433	32333	
Strain (Origin)	8088 (L)	0135 (B)	2566 (B)	3242 (B)	9375 (L)	9866 (L)	SI (South India)	6947 (L)	1339 (B)	7009 (B)
Rv0064									■	
Rv0180c					■					
Rv0795-96							▨	▨		
Rv0963c										■
Rv1354c									■	
Rv1355c-56c									■	■
Rv1369c-70c							▨	▨		
Rv1519	■	■	■	■	■	■				
Rv1524-25							■			
Rv1573→86c	▨						▨	▨		
Rv1735c						■				
Rv1755c		■								
Rv1756c-57c							▨	▨		
Rv1758			■		■					
Rv1759					■					
Rv1760→62					■					
Rv1763-64							▨	▨		
Rv1802						■				
Rv1895										■
Rv1917c					■					
Rv1947				■						
Rv2105-06							▨	▨		
Rv2167c-68c							▨	▨		
Rv2271→77c									■	■
Rv2278-79							▨	▨		
Rv2314c-15c								■		
Rv2354-55							▨	▨		
Rv2479c-80c							▨	▨		
Rv2595			■							
Rv2645→2647c							▨			
Rv2648-49							▨	▨		
Rv2650→ 59c							▨			
Rv2814c-15c							▨	▨		
Rv3017c						■				
Rv3018c						■				
Rv3019c	■	■	■		■	■			■	
Rv3020c	■	■	■	■	■	■			■	■
Rv3021c	■			■					■	
Rv3022c	■									
Rv3135	■	■	■	■	■	■	■	■	■	■
Rv3184→87							▨	▨		
Rv3203							■			
Rv3325							▨	▨		
Rv3381c							▨	▨		
Rv3382c→84c				■						
Rv3424c→28c								■		
Rv3474-75							▨	▨		
Rv3516-17	■	■	■	■	■	■				
Rv3651							■			
Rv3738c-39c	■	■	■	■	■					
Rv3741		■								
Rv3786c		■								
Rv3887c →89c								■		

Table 7.5. Distribution of deleted genes identified by microarray experiments.

The deleted sequences belonging to insertion or mobile elements are shown in grey. The VNTR profiles of the strains and Rv number, corresponding to *M. tuberculosis* H37Rv genome, are indicated. Only significant deletions are shown (p<0.05). Isolates were either isolated in Birmingham (B) or Leeds (L) in the UK.

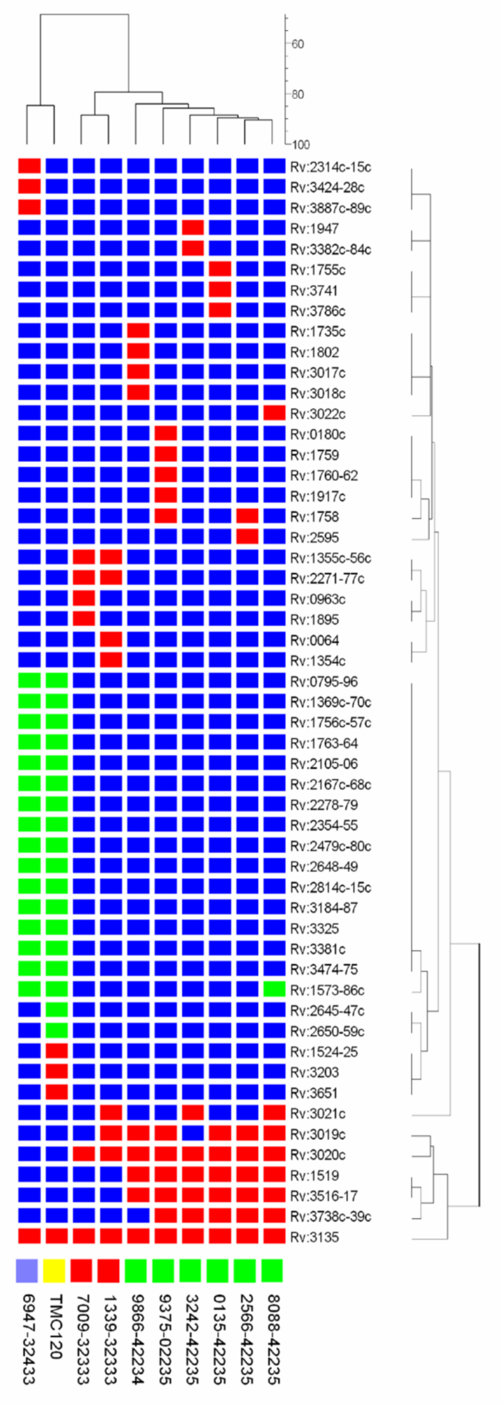


Figure 7.3. Deleted regions in the ten *M. tuberculosis* strains analysed.

The dendrogram was constructed by the unweighted pair group method (UPGMA) clustering using a categorical co-efficient. Red squares indicate deleted regions and green squares indicate regions disrupted by mobile elements.

7.4.3. Common deletions within this study and strain CH

The analysis of the deletion events in the SA-strains investigated showed that these strains possessed some common deletions present in all SA-strains (Rv1519, Rv3516-Rv3517 and Rv3738c-Rv3739c). PCR and sequencing showed that these deletions were indistinguishable to some of the deletions reported in the Leicester CH strain (Rajakumar *et al.*, 2004). Deletion Rv3738c-Rv3739c (PPE66-PPE67) was identified in all the SA-strains except 9866 (Table 7.5 and Table 7.6). None of these common deletions have been described in Beijing/W strains (Tsolaki *et al.*, 2005;Stavrum *et al.*, 2008) or in a collection of 100 clinical strains (Tsolaki *et al.*, 2004).

		South Asian strains (East African Indian lineage)					
	CH	**8088**	**0135**	**2566**	**3242**	**9375**	**9866**
VNTR	**42235**	**42235**	**42235**	**42235**	**42235**	**02235**	**42234**
Deleted genes							
Rv0180c	-	+	+	+	+	-	+
Rv1519 (RD750)	-	-	-	-	-	-	-
Rv1995-96	-	+	+	+	+	+	+
Rv3017c	+	+	+	+	+	+	-
Rv3018c	+	+	+	+	+	+	-
Rv3019c	-	-	-	-	+	-	-
Rv3020c	-	-	-	-	-	-	-
Rv3021c	+	-	+	+	-	+	+
Rv3022c	+	-	+	+	-	+	+
Rv3135	+	-	-	-	-	-	-
Rv3516-17	-	-	-	-	-	-	-
Rv3738c-39c	-	-	-	-	-	-	+

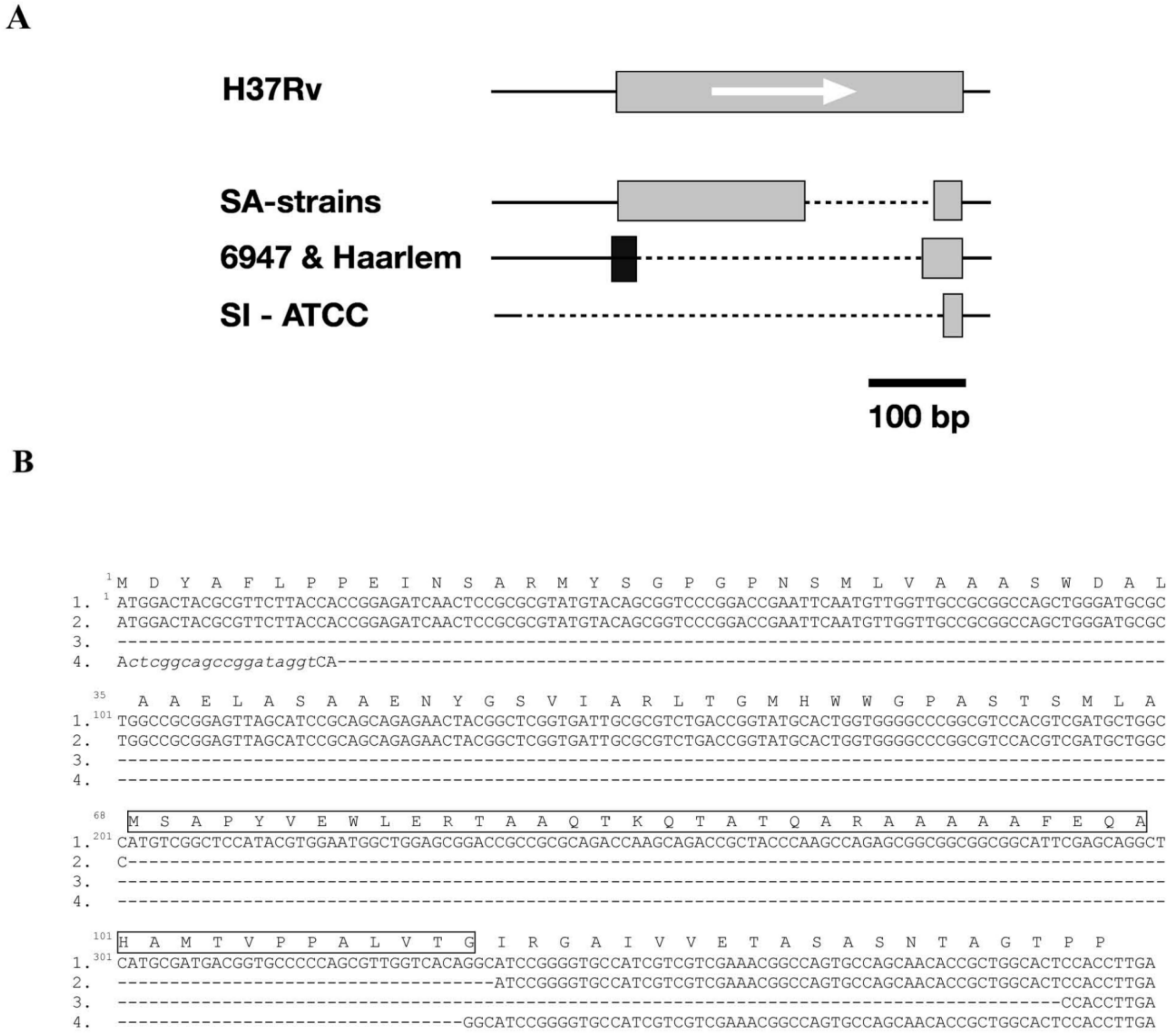
Table 7.6. Genes disrupted by deletions in SA-strains and the CH strain.

Deletions in the CH strain were previously reported by (Rajakumar *et al.*, 2004). Plus (+): gene present. Minus (-): gene deleted.

7.4.4. Polymorphic deletions

The SA-strains also showed additional deletions which distinguished them from the CH strain and from other strain families (Rajakumar *et al.*, 2004;Tsolaki *et al.*, 2004). These additional deletions provided genetic variability in this specific group of strains (Table 7.5 and Table 7.6). The sites of genetic variation among *M. tuberculosis* isolates could contribute to putative differences in the phenotypic and biological properties of the strains.

Previous studies have shown polymorphism in the deletions of some PE/PPE genes and were also found in the *M. tuberculosis* isolates studied here (Marmiesse *et al.*, 2004;Musser *et al.*, 2000). Four polymorphic deletions in genes or regions were analysed in the SA strains: Rv3135 (PPE50), Rv1917c (PPE34), Rv1755c-Rv1758 (plcC-cut1) region, and Rv3017c-Rv3022c region (*esxQ*PPE46-PE27A-*esxR*-*esxS*-PPE47-PPE48). Deletions in Rv3135 were observed in all of the strains examined in this study. A characteristic partial deletion was identified in all of the SA strains studied and the deletion differed from those present in other strains (Figure 7.4). This deletion was not found in the strain CH (Rajakumar *et al.*, 2004).



```
    1 M  D  Y  A  F  L  P  P  E  I  N  S  A  R  M  Y  S  G  P  G  P  N  S  M  L  V  A  A  A  S  W  D  A  L
1. 1 ATGGACTACGCGTTCTTACCACCGGAGATCAACTCCGCGCGTATGTACAGCGGTCCCGGACCGAATTCAATGTTGGTTGCCGCGGCCAGCTGGGATGCGC
2.   ATGGACTACGCGTTCTTACCACCGGAGATCAACTCCGCGCGTATGTACAGCGGTCCCGGACCGAATTCAATGTTGGTTGCCGCGGCCAGCTGGGATGCGC
3.   ----------------------------------------------------------------------------------------------------
4.   ActcggcagccggataggtCA-------------------------------------------------------------------------------

   35  A  A  E  L  A  S  A  A  E  N  Y  G  S  V  I  A  R  L  T  G  M  H  W  W  G  P  A  S  T  S  M  L  A
1.101 TGGCCGCGGAGTTAGCATCCGCAGCAGAGAACTACGGCTCGGTGATTGCGCGTCTGACCGGTATGCACTGGTGGGGCCCGGCGTCCACGTCGATGCTGGC
2.   TGGCCGCGGAGTTAGCATCCGCAGCAGAGAACTACGGCTCGGTGATTGCGCGTCTGACCGGTATGCACTGGTGGGGCCCGGCGTCCACGTCGATGCTGGC
3.   ----------------------------------------------------------------------------------------------------
4.   ----------------------------------------------------------------------------------------------------

   68  M  S  A  P  Y  V  E  W  L  E  R  T  A  A  Q  T  K  Q  T  A  T  Q  A  R  A  A  A  A  A  F  E  Q  A
1.201 CATGTCGGCTCCATACGTGGAATGGCTGGAGCGGACCGCCGCGCAGACCAAGCAGACCGCTACCCAAGCCAGAGCGGCGGCGGCGGCATTCGAGCAGGCT
2.   C---------------------------------------------------------------------------------------------------
3.   ----------------------------------------------------------------------------------------------------
4.   ----------------------------------------------------------------------------------------------------

  101 H  A  M  T  V  P  P  A  L  V  T  G  I  R  G  A  I  V  V  E  T  A  S  A  S  N  T  A  G  T  P  P
1.301 CATGCGATGACGGTGCCCCCAGCGTTGGTCACAGGCATCCGGGGTGCCATCGTCGTCGAAACGGCCAGTGCCAGCAACACCGCTGGCACTCCACCTTGA
2.   ------------------------------------ATCCGGGGTGCCATCGTCGTCGAAACGGCCAGTGCCAGCAACACCGCTGGCACTCCACCTTGA
3.   ------------------------------------------------------------------------------------------CCACCTTGA
4.   ---------------------------------GGCATCCGGGGTGCCATCGTCGTCGAAACGGCCAGTGCCAGCAACACCGCTGGCACTCCACCTTGA
```

Figure 7.4. Variation in the Rv3135 PPE gene.

A. The gene in H37Rv is shown with a grey box and the orientation of the gene is indicated by an arrow. The line with dots represents the deleted sequence in the strains compared to H37Rv. The new sequence in some of the strains is shown as a black box.

B. Sequence comparison of different Rv3135 variants: 1. H37Rv; 2. South Asian strains (clinical strains: 8088, 0135, 2566, 3242, 9375 and 9866); and 3. SI-ATCC; and 4. 6947-Erdman (Acc. no. AE007137)-CDC1551 (Acc. no. Y17598). The deleted nucleotides are indicated by dashes. The new sequence with respect to H37Rv is in lower case. The deleted amino acids in SA-strains are indicated by a box in the sequence of H37Rv. Figure created by Carmen Menéndez

The p-value obtained for this deletion was variable among SA-strains (p=0.0271-0.168) but presence of the deletion was validated by PCR and sequencing which revealed the identical partial deletion in all of the strains. The deletion of 135 nucleotides in the sequence of this gene caused the loss of 45 amino acids in the sequence of the protein (Figure 7.4). In order to confirm the variability of this gene, the deletion of Rv3135 was also confirmed in the three Euro-American strains and the South Indian strain ATCC 35811, both of them with consistent statistical significance ($p<0.05$). The Rv3135 gene in the three Euro-American strains had an insertion of 20 nucleotides at the 5' end of the deleted genomic sequence. The sequences in these strains were 100% homologous with those in the Erdman and CDC1551 strains of *M. tuberculosis*. The clinical isolate South Indian ATCC 35811 possessed a large deletion in Rv3135 with almost the entire region deleted including the promoter sequences for this region.

Isolate 9375 (VNTR profile 02235) was the only isolate that possessed a partial deletion in Rv1917c (PPE34) that was detected by the microarray. Rv1917c encodes a surface exposed protein, which is highly polymorphic in clinical isolates (Sampson *et al.*, 2001). Most of the polymorphism detected in this gene is caused by variation in VNTR loci present in this region (Sampson *et al.*, 2001). However, no evidence of the ETR-A region was detected in Rv1917c of the 02235 strain with the deleted region replaced by IS*6110*. Subsequent analyses by PCR and sequencing also showed disruption of this region in SA-strains 8088 (VNTR profile 42235) and 9866 (VNTR profile 42234) that was not detected by microarray (Figure 7.5). The orientation of IS*6110* in all three strains was opposite that of the direction of gene transcription. The insertion of an IS*6110* element in the Rv1917c was not detected in the CH strain (Yesilkaya *et al.*, 2006;Rajakumar *et al.*, 2004).

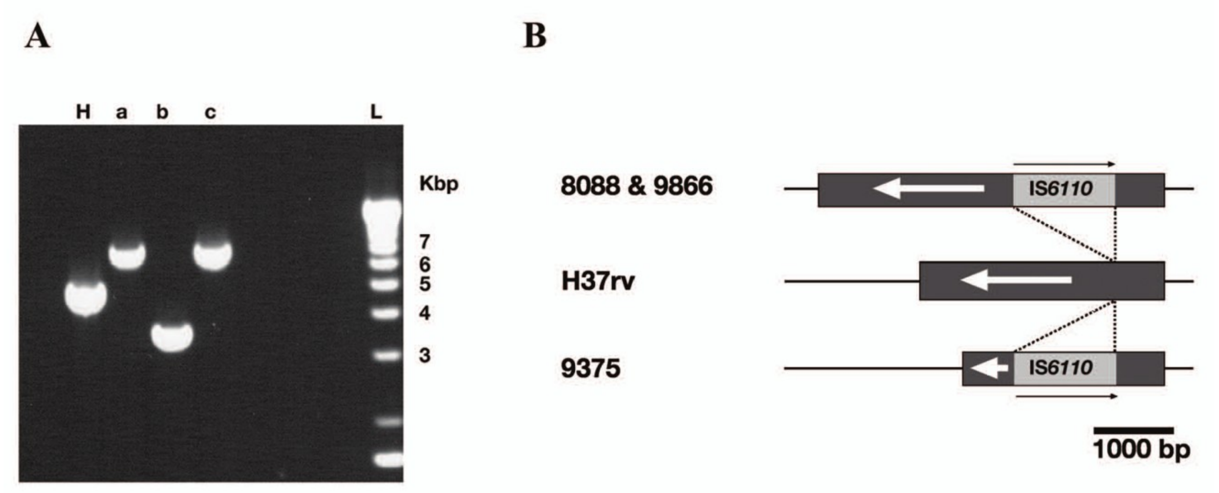


Figure 7.5. Polymorphism in the Rv1917c gene.

A. Rv1917c region amplified by PCR. H, H37Rv; (a) clinical strain 8088; (b) clinical strain 9375 and (c) clinical strain 9866.

B. Graphical representation of the location of the insertion of IS*6110* in Rv1917c. The gene is represented by a dark grey box and the IS*6110* are indicated by a light grey box. The insertion points are indicated with lines. The scale in base pairs is indicated. Figure created by Carmen Menéndez.

A polymorphic region was identified between Rv3017c and Rv3022c in the three SA-strains. All three SA strains have at least *esxR* (Rv3019c) and *esxS* (Rv3020c) deleted, with the deletion of adjacent genes in this region variable between strains (Figure 7.6). This region also contains PE/PPE genes with highly repetitive sequences (Rv3018c (PPE46), Rv3018A (PE27A), Rv3021c (PPE47), Rv3022c (PPE48), and Rv3022a (PE29)), and three genes encoding ESAT-6 like genes *esxQ* (Rv3017c), *esxR* (Rv3019c) and *esxS* (Rv3020c). One IS*6110* sequence was inserted in the intergenic region of Rv3018c and Rv3019c of the CH strain but the deletion of Rv3019c-Rv3020c located by microarrays was not confirmed (Rajakumar *et al.*, 2004;Yesilkaya *et al.*, 2006). The variability of the deletions in this genomic region has previously been shown in some *M. tuberculosis* and *M. microti* strains but this chapter was the first report of replacement by an insertion element or transposase in this region (Figure 7.6) in closely related strains (Marmiesse *et al.*, 2004).

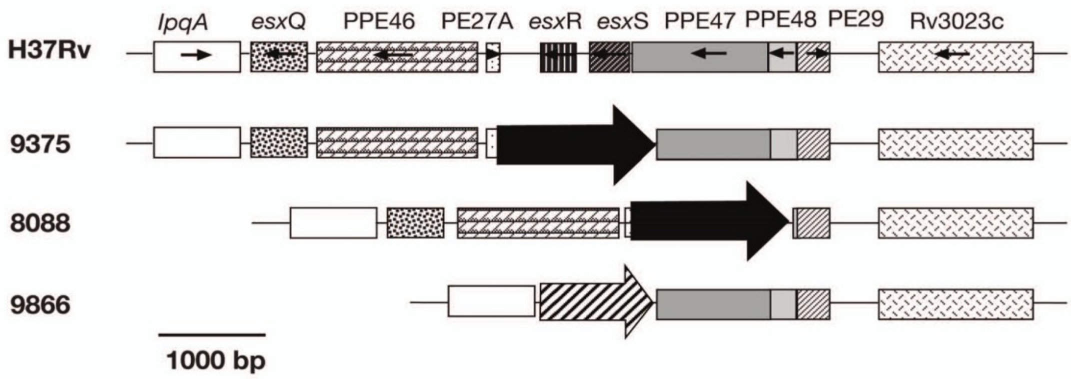


Figure 7.6. Polymorphism between Rv3017c and Rv3022c.

The orientation of the genes and the genes are shown. The insertion of IS*6110* is indicated by a black arrow, the insertion of the transposase (second ORF of the IS*6110*) is represented by a striped arrow. The scale in base pairs is indicated. Figure created by Carmen Menéndez.

The region between Rv1755c (*plcD*, phospholipase) and Rv1758 (*cut1*, cutinase) was analysed by PCR and DNA sequencing. These genes are interrupted by a single IS*6110* element in H37Rv and corresponded to the RvD2 region that is deleted in H37Rv but present in *M. bovis*, H37Ra and other clinical strains (Gordon *et al.*, 1999;Brosch *et al.*, 1999). Insertion and deletion events in this region have resulted in notably high diversity among strains (Ho *et al.*, 2000). The region Rvd2 is located downstream of the *plcD* gene (Rv1755c) and is present in strains 9375 and 9866 but not in 8088. In strains 8088 and 9866, *plcD* (Rv1755c) and cut1 (Rv1758) genes were not interrupted by a copy of IS*6110*, whereas they are in the *M. tuberculosis* H37Rv genome (Figure 7.7).

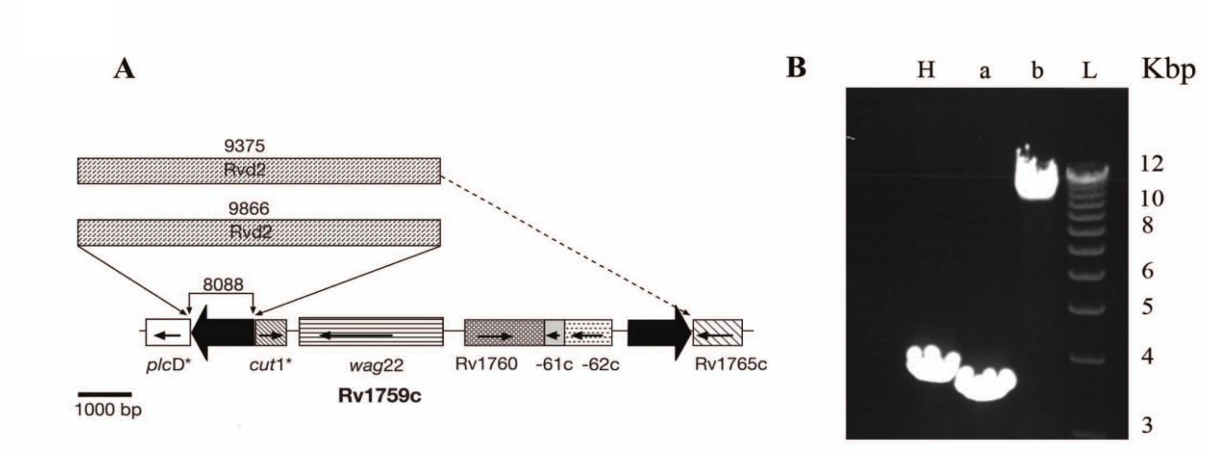


Figure 7.7. Rv1755c-Rv1765c region.

The orientation of the genes is represented by arrows. The IS*6110* in H37Rv are shown by black arrows. The deleted sequences in each strain were confirmed by PCR. The assumed substitution by Rvd2 in the strains 9375 and 9866 is shown in the figure. The asterisks indicate the truncated genes in H37Rv by IS*6110* insertion. (B) Region amplified by PCR. H, H37Rv; (a) clinical strain 8088; (b) clinical strain 9866 and L, ladder. Figure created by Carmen Menéndez.

The *plcD* (Rv1755c) gene is also interrupted by a copy of IS*6110* in the genome of the CH strain (Yesilkaya *et al.*, 2006). Rv1759c (wag22, member of the PE-Polymorphic GC-rich Repetitive Sequence (PGRS) family) is only absent in isolate 9375 where the deletion extends to the adjacent genes. The size of the PCR fragment seems to indicate the presence of the region Rvd2 in this strain and sequencing using the internal primers of Rvd2 confirmed its presence. The deletion or presence of prophages is variable in clinical strains of *M. tuberculosis*. Strain 8088 possessed a deletion in the RD3 region (phiRV1 prophage). This prophage is present in the genome of *M. tuberculosis* H37Rv but deleted in *M. bovis* BCG and some strains of *M. tuberculosis* (Gordon *et al.*, 1999). This deletion also appeared in the Euro-American strain 6947 and in the South Indian clinical isolate ATCC 35811 (our data) and has been described in Beijing/W strains (Tsolaki *et al.*, 2005).

7.4.5. Accession numbers

The nucleotide sequences of deleted regions of SA-strains are in the EMBL Data Bank with the accession numbers: AJ878456; AJ878457; AJ878458; AJ878459; AJ878460; AJ878461; AJ879166; AJ879167; AJ879168; AJ879169; AJ879170; AJ879171; AJ879172; AJ879173; AJ879174; AJ879175; AJ879176; AJ879177; AJ879178; AJ879179; AJ879180; and AJ879181.

7.5. DISCUSSION

In this chapter, microarray-based comparative genomics were utilised to analyse the distribution of deleted regions in the genome of six clinical strains of *M. tuberculosis* that originated from South Asia in comparison to the previously reported CH strain that also originates from South Asia. Examples of other common clades were also included. Our results indicate that strains of the South Asian group have a common evolutionary origin similar to the strain CH but are distinct from members of the Beijing clade.

7.5.1. Global phylogeny of strains in this study

There were characteristic deletions in each of the 10 strains in our study that assigned each strain to one of six global phylogenetic lineages as previously described (Gagneux *et al.*, 2006). All six of the SA-strains were assigned to the East-African-Indian lineage. The South Indian clinical isolate ATCC 35811 was assigned to the Indo-Oceanic lineage and the three non-SA strains were all assigned to the Euro-American lineage. These deletion based lineages equate to various spoligotype clades: East-African-Indian equates to CAS; Indo-Oceanic is the ancestral EAI spoligotype; and the Euro-American lineage contains Haarlem, LAM, T, and X (Gagneux and Small, 2007).

Presence of the RD750 deletion defined SA-strains as members of the East-African-Indian global lineage (Gagneux *et al.*, 2006). Recent studies examined the immunological relevance of this deleted region in the CH strain and found that the deletion in RD750 conferred a phenotype that could subvert the host immune response (Newton *et al.*, 2006).

A subsequent microarray study examined Beijing and non-Beijing strains (including one CAS) present in Myanmar. Comparison of our data with the data from Myanmar identified three common deletions between Beijing strains in Myanmar and CAS strains studied in this chapter with multiple other variable deletions identified as well (Stavrum *et al.*, 2008). Comparative genomics and population-based spoligotyping studies indicate that Beijing and CAS may share a common ancestry but have evolved into two currently distinct clades.

7.5.2. The origin of two major global clades: CAS and the Beijing strain

CAS is a member of PGG 1 (Sreevatsan *et al.*, 1997;Kulkarni *et al.*, 2005) but does not possess the TbD1 (TbD1-) region (Gutierrez *et al.*, 2006). The Beijing and EAI families are also located in PGG1 but Beijing is TbD1- with EAI TbD1+ (Gutierrez *et al.*, 2006). Therefore, the EAI clade was present before the appearance of CAS and Beijing. Both Beijing and CAS are located in the same principal phylogenetic group and share the presence of spacers 35-43. However, the Beijing clade has lost spacers 1-22 that are present in CAS (Banu *et al.*, 2004). Therefore, the CAS family could have been one of the possible ancestors of the Beijing strain (Sola *et al.*, 2003;Stavrum *et al.*, 2008). It is hypothesised that the Manu spoligotype (with only two deleted spacers at 33 and 34) is the probable ancestor of both the CAS and EAI clade (Singh *et al.*, 2004).

Spoligotyping has been widely used for population-based phylogenetic studies. However, it has been shown that spoligotyping is not as robust for constructing phylogenetic frameworks when compared to large-scale LSP or SNP-based data as significant homoplasy may exist (Comas *et al.*, 2009;Kato-Maeda *et al.*, 2011). Studies using a minimal number of SNPs further strengthened the early spoligotype evidence that EAI is an ancestral clade followed by

CAS (Filliol *et al.*, 2006). More recent studies have refined spoligotype clades using minimal SNP sets whilst almost being as informative as large-scale sequencing projects (Abadia *et al.*, 2010). The ancestral EAI clade is more predominant in the South of India with the CAS clade more predominant in the North of India (Gutierrez *et al.*, 2006;Singh *et al.*, 2007).

There may be a striking correlation between the distribution of the human population in North and South India and the distribution of *M. tuberculosis* strains. This may indicate that *M. tuberculosis* has co-evolved with the human population within India. A recent study of human genetics in North and South India (and two Pakistani populations) studied the distribution of 560,000 SNPs in the whole genomes of 132 individuals from 25 groups representing the breadth of social, language and geographic variation in India. Most of the Indian populations sampled are mixtures of two groups which were named the Ancestral North Indians (ANI) and Ancestral South Indians (ASI). The ratio of ANI:ASI varies from 39% to 79% in a cline from south to north. The ANI human population represents an older clade displaced from the North by the Indo-European speaking ANI in approximately 1500BC (Reich *et al.*, 2009). This geographical distribution of human populations closely resembles the distribution of the EAI and CAS clades in India.

Specific human genotyping studies examining the relationship between *M. tuberculosis* and host genotypes in India have not yet been undertaken. However, human genetic studies in other geographical areas have shown that there are associations between both the genotypes of the host immune system and *M. tuberculosis* strain (Ma *et al.*, 2007). Moreover, specificity host-mycobacterial interactions have been identified in Toll-Like Receptors of patients infected with *Mycobacterium leprae* in the ISC (Wong *et al.*, 2010).

7.5.3. Functional implications of identified deletions

It is interesting to note that some of the deleted regions in the SA-strains studied here were also found in the index strain CH from the outbreak in Leicester in 2001, where the strains were isolated from patients that were also of predominantly South Asia origin (Rajakumar *et al.*, 2004). The deletion of LSP designated RD750 involved the Rv1519 and Rv1520 genes that encode conserved hypothetical proteins with unknown function. This deletion has been found to be associated with the persistence in human populations of the CH strain (Newton *et al.*, 2006). An understanding of the pathogenicity of this group of strains might be aided by an in-depth study of genetic similarities and differences of an expanded collection of strains with similar epidemiological and genomic characteristics.

7.5.4. Variability in PE/PPE genes

The results in this chapter show major variability in some of the PE/PPE genes of SA-strains that represent a major portion of the *M. tuberculosis* H37Rv genome (Cole *et al.*, 1998). The biological function of most of these proteins is unknown. It has been suggested that they are involved in antigenic variation or interfere with host immune responses (Cole *et al.*, 1998;Banu *et al.*, 2002;Brennan *et al.*, 2001). Moreover, genes of the ESAT-6 family (Cole *et al.*, 1998), which are closely linked in the genome to genes encoding PE/PPEs (Tekaia *et al.*, 1999), have been demonstrated to encode several immunodominant molecules that are strongly recognized by the immune system in different animal models of TB as well as by T-cells from humans exposed to *M. tuberculosis* (Skjot *et al.*, 2002). This family of proteins has also been shown to be immunogenic in a human peripheral blood mononuclear cell model (Skeiky *et al.*, 2000). Some cell-surface proteins present in *M. tuberculosis*, such as PE/PPE genes, can influence the type of host response and can alter the composition of the infiltrates and the cellular composition of lesions (Cosma *et al.*, 2003). These potential antigens for host immunity may play an important role during host-pathogen interactions. The antigenic variability in the analysed strains could contribute, with a combination of other factors, to a special interaction with specific hosts, in this case people originating from South Asia.

Another major source of polymorphism among *M. tuberculosis* isolates appears to be defined by deletions in Rv3135, as the gene is larger in Beijing strains than in H37Rv (Musser *et al.*, 2000;Mokrousov *et al.*, 2002). The commonly deleted regions of Beijing/W strains do not extend to this family of proteins (Tsolaki *et al.*, 2005). The Rv3135 gene, which encodes the PPE50 protein, was found to be uniquely variable and could be a useful marker to differentiate epidemiologically related groups of strains, since there was a common deletion of this gene in the SA-strains (Musser *et al.*, 2000). However, the full extent of variability within this region needs to be determined in a larger collection of isolates as this gene was not affected by deletion in the CH strain (Rajakumar *et al.*, 2004). Despite having an unknown function, Rv3135 is considered an essential gene and is closely linked to a two-component system (*dosR*/*dosS*) in the genome of H37Rv (Sassetti *et al.*, 2003). However, the effect of this deletion in these strains remains unknown, as a defect in this PPE gene may be complemented by another gene absent in the reference strain H37Rv or by another PPE due to the functional similarities among these members.

The presence of an insertion in Rv1917c (PPE34) could contribute to genetic variability, which is not only important in genomic evolution but also in the expression of these genes and in antigenic variation. The similar point of insertion of the IS*6110* among our SA-strains and the CH strain suggest the presence of a hot spot in this gene. Polymorphisms in this gene are used as an epidemiological tool because it contains one of the more discriminatory loci (ETR-A) in VNTR typing (Sola *et al.*, 2003).

From the five years of prospective typing presented in Chapter 5, 530/5,731 (9%) strains possessed a non-amplifiable ETR-A locus. Of the six global clades, the East African Indian Clade had the highest number of strains with non-amplifiable ETR-A (217/1,888) which equates to 11% of all strains in this clade. Confirmation of disruption caused by IS*6110* would need to be carried out by PCR and DNA sequencing.

7.5.5. Variability in ESAT-6-like genes

A notable microarray result was the detection of one polymorphic region from Rv3017c to Rv3022c among the three SA-strains which encodes for ESAT-6-like genes. Deletion of ESAT-6-like genes might confer a selective advantage during certain stages of infection or transmission. The role provided by the selective pressure of the host's immunological system in the deletion of some genes is not well understood; however it has been suggested that there is limited selective pressure because of little genetic diversity in a large number of genes encoding essential antigens which are recognized by the host immune system (Musser *et al.*, 2000). The tight binding capability of the closely related ESAT-6 proteins has been reported (Lightbody *et al.*, 2004), but the loss of two of these genes (Rv3019c and Rv3020c) could decrease the opportunity for plasticity in the SA-strains. Rv3019 was also identified as a deletion in a comparative genomic hybridisation study of Beijing and non-Beijing strains in Myanmar. This region was identified as being present in all of the LAM strains tested, present in 15/28 Beijing strains tested, present in 5/9 EAI strains tested but deleted in the single CAS isolate tested (Stavrum *et al.*, 2008). The strain 9866 (with VNTR 42234) is the only strain that possessed a deletion in three of the *esx* genes (Rv3017c, Rv3019c and Rv3020c). Further analysis is necessary to understand the implications that this deletion may have.

The insertion elements that replaced deletions in SA-strains could modify the expression of adjacent PE/PPE genes in the region from Rv3017c to Rv3022c. Taking into account that IS*6110* can upregulate downstream genes (Safi *et al.*, 2004;Soto *et al.*, 2004), the replacement event in this region demonstrated the important role of insertion elements in the evolution of the genome and in the contribution to phenotypic diversity. The presence of insertion elements could contribute to this variability, which is not only important in genomic evolution but also in expression of these genes and in antigenic variation. In *M. microti* the deletion MiD4 removes the *esx* genes Rv3019c and Rv3020c, but the strains which possess this deletion and RD5 (deletion in the *plcC* region) were still able to produce disease (Garcia-Pelayo *et al.*, 2004). However, the protective efficacy of Rv3019c and its suitability as promising candidate TB vaccine suggests that the deficiency of members of the ESAT-6 gene family may affect the immunological response in the patient (Hogarth *et al.*, 2005). In strain 8088, one IS*6110* element was inserted in the *plcC* gene (data not shown) but this gene was not interrupted by insertion elements as was the case in strains 9375 or 9866. Mutations in the *plcC* gene attenuates strains (Raynaud *et al.*, 2002), with further studies required to confirm this effect in SA-strains. In addition, the Rvd2 region has been found in two of the SA strains in this study. Located close to this region is Rv1759c (*wag22*), which encodes a member of the PE-PGRS glycine rich protein family (Espitia *et al.*, 1999). This gene is expressed in tuberculosis infections and the protein is recognized by sera from patients and seems to have a biological role in the interaction of the bacillus with the host (Espitia *et al.*, 1999). The phospholipase genomic region has already been shown to be a preferential locus for IS*6110* transposition (Vera-Cabrera *et al.*, 2001). However, in the SA-strains Rv1755c (*plcD*) was not interrupted by IS*6110*. The polymorphic deletions detected in SA-strains in this study (Rv3135, Rv1917c, Rv1755c-Rv1758 and Rv3017c-Rv3022c) were also polymorphic among previously analysed clinical strains of *M. tuberculosis* and are not specific to the SA-strains (Brosch *et al.*, 1999;Marmiesse *et al.*, 2004;Musser *et al.*, 2000;Sampson *et al.*, 2001). Rv3135 has been recently assigned to a group of large-sequence polymorphisms (LSP Group C) by Alland and colleagues in which it has been suggested that sequence alterations in this region occur under selective pressure (Alland *et al.*, 2007).

This chapter indicates that both SA strains 8088 and 9375 (VNTRs 42235 and 02235, respectively) showed the same deletion in the Rv3738-Rv3739c region which was not present in strain 9866 (VNTR profile 42234). This genomic deletion involved a partial removal of the Rv3737 gene, a probable conserved transmembrane protein close to a transcriptional regulatory protein (Rv3736) belonging to the *AraC*/*XylS* family (Cole *et al.*, 1998). The implication that deletions of Rv3737, Rv3738 and Rv3739 affect expression of Rv3736 will require further investigation.

7.5.6. Phenotypic and clinical associations identified by comparative genomics in previous studies

It may be that further studies of the deletions identified in this chapter may reveal significant associations with phenotypic and clinical characteristics of *M. tuberculosis*. The aforementioned RD750 deletion in the CH strain enables this strain to subvert the host immune response which could contribute to the transmission of this strain (Newton *et al.*, 2006). Since this deletion defines a global lineage, all strains within the East-African-Indian lineage may exhibit the same immune subverting phenotype as the CH strain.

An initial study of 19 clinical strains identified an association between the amount of deleted genomic DNA and cavitary disease. As the amount of deleted genomic DNA sequence increased, the probability of pulmonary cavitations decreased; suggesting that the accumulation of mutations diminished strain pathogenicity (Kato-Maeda *et al.*, 2001). Within specific clades such as the Beijing strain, the presence of deletions in five specific RDs were significantly associated with extrathoracic tuberculosis (Kong *et al.*, 2006).

7.5.7. Limitations of microarrays

There are two principal limitations of microarray-based deletion analysis. Microarrays will only identify deletions relative to the reference strain which will not contain all of the genome of all members of that particular species. *M. tuberculosis* H37Rv lacks at least five regions that are present in clinical isolates of and other members of the *M. tuberculosis* complex (Brosch *et al.*, 1999). This limitation has been overcome with the latest array designs which also include regions from *M. tuberculosis* H37Rv, *M. tuberculosis* CDC1551, and *M. bovis* 2122/97 (Stavrum *et al.*, 2008).

Microarrays also cannot detect genetic rearrangements involving duplication of DNA which has been shown to have played a significant part in the evolution of *M. tuberculosis*. An *in silico* analysis of the *M. tuberculosis* proteome showed that at least 50% of mycobacterial proteins arose from gene duplications or rearrangement including genes involved in fatty acid metabolism, regulation of gene expression, and PE/PPE proteins (Tekaia *et al.*, 1999).

DNA microarrays only provide data on large gene disruptions within specific ORFs. Further resolution on a genome-wide scale can now be provided by WGS, which is able to provide data on variations caused by SNPs and insertions or deletions. Through the analysis of read coverage, the presence or absence of RDs/LSPs can be inferred.

7.6. CONCLUSIONS

Variability in the genome of strains of *M. tuberculosis* which are associated with patients of South Asian origin has been demonstrated in this chapter. The reasons for the host selection of this group of strains cannot be explained only by the deletion pattern, but are probably a combination of events which are involved in the interaction with the cell and the host. Other studies have shown the presence of deleted genes in strain families present around the world, but it has been shown that the Beijing/W strain lack the deletion in genes involved in antigenic variation such as PE/PPE members or ESAT-6 like proteins (Tsolaki *et al.*, 2005). Some of these genes, such as Rv3135 or Rv3738-Rv3739, are very closely linked to genes known to be important during infection. The high levels of variability in particular regions like Rv3017c-Rv3022c in the *M. tuberculosis* SA strains need further investigation by analysis of multiple strains within the CAS lineage and examination of protein expression in order to understand the differences in the pathogenesis, virulence and development of this prevalent global strain.

8. SUMMARY AND FUTURE WORK

8.1. SUMMARY AND CONCLUSIONS

This project was an investigation of *M. tuberculosis* genotypes in the West and East Midlands. The main aim of this project was to improve our knowledge and understanding of the molecular epidemiology and transmission of *M. tuberculosis* in the Midlands.

The specific objectives of this PhD were to:

- Evaluate IS*6110* RFLP data in comparison to MIRU-VNTR data to identify if MIRU-VNTR typing could be a rapid alternative method to IS*6110* RFLP.
- Develop automated analysis of MIRU-VNTR loci using non-dHPLC which would enable the analysis of all *M. tuberculosis* strains in the Midlands and enhanced the knowledge and understanding of the epidemiology and transmission of *M. tuberculosis* in the Midlands.
- Identify the prevalent global lineages present in the Midlands and the continental origin of infected patients using non-dHPLC, MIRU-VNTR, and a novel computer software programme called Origins.
- Identify, confirm, and analyse the distribution of the single most prevalent MIRU-VNTR profile in the Midlands.
- Undertake a whole-genome analysis of prevalent strains in the Midlands using DNA-DNA microarray hybridisation to identify properties unique to those strains and the genetic origin of these strains.

For these objectives, MIRU-VNTR typing was evaluated against IS*6110* typing and shown to provide comparable clustering when epidemiological links were known. Non-dHPLC was shown to be a feasible automated option of MIRU-VNTR typing. The distribution of all global clades present in the Midlands were analysed and the predominance of two global clades was shown. Finally, whole genome array analysis of strains that originated from the Indian-Subcontinent showed that strains within the East African Indian lineage do have a common origin.

The clonality of all strains in the Midlands was analysed and from this, the single most prevalent strain was identified and investigated further by conventional epidemiological methods. The Mercian strain represents one of the world's longest running and largest TB outbreaks. Within the study period analysed in this thesis, the Mercian strains was identified in a total of 214 patients between 1995 and 2008. *M. tuberculosis* causes active disease in only 10% of patients infected (Sutherland, 1976). From the 214 patients diagnosed with active disease and infected with the Mercian strain, there could be more than 2,000 individuals latently infected with the Mercian strain in the West Midlands.

8.1.1. Limitations

The main limitation of this research is that the presented genotyping data between 2004 and 2009 used 15 MIRU-VNTR loci and the optimal 24 loci set was used only on selected strains. This was mainly due to the additional funding required to analyse the additional nine loci which was not obtained until the end of 2009. This has enabled the commencement of 24 loci typing of all isolates at the start of 2010 with 1,376 strains analysed prospectively by 24 loci so far.

As part of the previously published evaluation of IS*6110* RFLP, candidate MIRU-VNTR loci, and spoligotyping (Supply *et al.*, 2006), it was identified that optimal performance for cluster identification was provided by the combination of data from all three methods. For this thesis, RFLP and spoligotyping was carried out on selected strains. Alternative methods for IS*6110* RFLP and spoligotyping were developed and evaluated during this thesis. Selected *M. tuberculosis* strains analysed by IS*6110* RFLP within the Birmingham Laboratory were analysed by an alternative method (IS*6110* fAFLP) which has been developed by Dr Cath Arnold (HPA Centre for Infections, London). fAFLP determines the insertion position and copy number of IS*6110* in *M. tuberculosis* strains using a DNA sequencer which generates a digital sequence that details the position of each IS*6110* element. This method has been shown to generate comparable clustering when compared to IS*6110* fingerprints obtained using the conventional Southern hybridisation method (Thorne *et al.*, 2007a). IS*6110* fAFLP analysis of all available Mercian strains is currently being carried out. Spoligotyping using a liquid suspension array system was evaluated using a system available at the University of Birmingham as part of another PhD project and further implementation of this method awaits the procurement of a Luminex system by the HPA laboratory.

8.2. HYPOTHESES GENERATED AND FUTURE WORK

The data obtained from DNA fingerprinting of all *M. tuberculosis* isolates in the Midlands from 2004 onwards as part of this thesis will form the basis for multiple future projects.

8.2.1. In-depth molecular characterisation of the Mercian strain

The widespread transmission of the Mercian strain may be caused by bacterial, host, or environmental factors. Various epidemiological investigations have been undertaken and the only environmental factors identified so far were a significant association with alcohol or drug use in Wolverhampton. No specific location has been identified. The only putative host factor identified so far was that a cohort of eight patients infected with the Mercian strain in Wolverhampton continued to experience weight loss even after initiation of therapy with no treatment adherence issues. However, patients infected with the Mercian strain completed treatment with a similar rate as the rest of the patient population. A further hypothesis to test from this data would be that the Mercian strain interacts differently with the host immune response. This could be tested by analysing the response of an *in vitro* THP-1 macrophage model to infection with the Mercian strain.

From the 24 locus MIRU-VNTR and IS*6110* RFLP data, it appears that the Mercian strain has spread by clonal expansion in three different cities across the West Midlands. What we do not know is whether the Mercian strain appeared in one city with two subsequent major transmission events to the other cities and subsequent transmission contained to those cities or whether continuous transmission between the three cities is still occurring. The hypothesis that the Mercian strain is truly clonal will be tested by whole genome sequencing which has been has been used to elucidate the transmission route of the Harlingen strain in the

Netherlands (Schurch *et al.*, 2010a). The hypothesis that the Mercian strain has transmitted widely because it possesses unique genetic properties that enhances virulence and transmission will be tested by analysis of the whole genomic content by DNA microarrays.

8.2.2. Prospective national 24 locus typing

Prior to 2010, typing was funded on a regional level which resulted in variation in the proportion of *M. tuberculosis* strains typed. The HPA has funded a National TB Typing Service initially for three years (2010-2012) which includes universal, prospective typing of all *M. tuberculosis* isolates in England by 24 MIRU-VNTR loci (Health Protection Agency, 2011). The impact of 24 loci typing on a national scale will be evaluated by various parameters including the number of false positive cultures identified and the enhancement of conventional epidemiological data when combined with molecular data.

From Chapter 5 and the identification of two predominant global clades in the Midlands, the next hypothesis to accurately determine would be the determination of what proportion of strains that originate from Southern Asia are directly imported and what proportion are the result of recent transmission within the Midlands. This will be determined by adding the extra nine loci retrospectively to clusters in Birmingham and the West Midlands that were identified using 15 loci between 2007 and 2009. This will create a four year period of 24 locus strain typing data (2007-2010) that achieves the temporal requirements for the construction of an epidemiologically relevant transmission model for the city of Birmingham (Vynnycky *et al.*, 2001). This would be the first such long-term transmission model constructed for in England and will identify in which population groups transmission (UK-born and non-UK-born) is occurring and associations between clinical presentation, clustering, and strain genotype. There have been two previous studies in England that have estimated transmission rates (Love *et al.*, 2009;Maguire *et al.*, 2002). However, the durations of these studies were only 12 and 24 months respectively.

8.2.3. Social Network Analysis

Currently, clusters identified by 24 loci data obtained prospectively are analysed on a monthly basis by the Birmingham Chest Clinic led by Dr. Martin Dedicoat. The aim is to confirm or identify social links between patients. These cluster reviews also determine if further screening for latent TB infection is required. To improve the identification of links between patients, structured social network analysis using a standardised patient questionnaire and epidemiological analysis software such as UCINet will be carried out (Borgatti *et al.*, 1999). Social network analysis identifies patients or locations (nodes) that have the highest centrality scores between all nodes within a cluster. This would enable the identification of patients that have exhibited the highest rate of contact with all other patients in a cluster and could potentially be a "super-spreader". A location or social activity may be identified as a central node which could focus further screening for latent infection onto that particular location. The Mercian strain will be investigated by social network analysis with the aim of uncovering previously identified social links and evaluate the hypothesis that the Mercian strain is the most prevalent strain because there is a significant environmental factor associated with transmission.

8.2.4. *In vitro* infection of THP-1 monocytes

As part of a PhD project carried out by Helen Smith at Birmingham University, phenotypic differences within and between clades that originate from the ISC are currently being analysed by investigating the interaction with THP-1 macrophages in an *in vitro* infection model.

8.2.5. Whole Genome Sequencing

Strains analysed in this thesis will be selected for analysis by large-scale WGS as part of the UK Clinical Research Collaboration Modernising Medical Microbiology Consortium which is funded by the Wellcome Trust and MRC. The three collaborative institutions in this project are the University of Oxford, HPA and the Wellcome Trust Sanger Institute. The aim of this project is to establish how new technologies such as high-throughput next generation WGS can be optimally integrated into clinical microbiology. To calibrate the rate of variation in the DNA sequence of *M. tuberculosis*, clusters ranging in size from two or more isolates from the same patient to large complex clusters with linking conventional epidemiological data will be sequenced. After the retrospective analysis of isolates, it is planned to undertake prospective WGS as patients are identified with active TB disease. It is anticipated that WGS will be able to provide even greater resolution in defining epidemiological links between patients and could determine the exact transmission routes of large, complex outbreaks.

APPENDIX I. REAGENTS AND SOLUTIONS

CTAB/NaCl:

- 4.1 g NaCl in 80 ml sterile distilled H_2O
- 10 g N-cetyl-N,N,N,-trimethyl ammonium bromide (CTAB)
- 80 ml sterile distilled H_2O
- Heat the solution to 65°C and adjust the volume to 100 ml with sterile distilled H_2O

20X SSC:

- 3 M NaCl
- 0.3 M Na-citrate
- 1 L sterile distilled H_2O
- Adjust the pH to 7.0 with 5 M NaOH

Stringency Wash Solution 1:

- 100 ml 2X SSC
- 0.1% SDS

Stringency Wash Solution 2:

- 100 ml 0.1X SSC
- 0.1% SDS

50X TAE:

- 2 M tris
- 0.05 M EDTA
- Adjust to pH 8.0 with ~57 ml glacial acetic acid.
- Make up to 1 l with sterile distilled H_2O

1X TBE:

- 89 mM tris
- 89 mM boric acid
- 2.5 mM EDTA
- Adjust the pH to 8.2.
- 1 l sterile distilled H_2O

10X TE:

- 100 mM Tris-HCl, pH 8.0
- 10 mM EDTA
- 1 l sterile distilled H_2O

1.5X TMAC Hybridisation Solution:

- 4.5 M TMAC
- 0.15% sarkosyl solution
- 75 mM Tris-HCl, pH 8.0
- 6 mM EDTA, pH 8.0
- 250 ml sterile distilled H_2O

1X TMAC Hybridisation Solution:

- 3 M TMAC
- 0.1% sarkosyl solution
- 50 mM Tris-HCl, pH 8.0
- 4 mM EDTA, pH 8.0
- 250 ml sterile distilled H_2O

APPENDIX II. EPIDEMIOLOGY FOR TABLE 3.4

Cluster	Patient	Conventional Epidemiological Data of each Patient	Extra nine loci profile	Concordance between molecular and conventional epidemiological data using	
				ETR+MIRU	VNTR+
1	1	Same county as C5 but different town.	Orphan	No	Yes
	2	Same county as C5 but different town. Same town as P3.	344423692	No	Yes
	3	Same county as C5 but different town. Same town as P2.	344423692	No	Yes
	4	Same county as C5 but different town.	Orphan	No	Yes
2	1	No known links.	Orphan	No	Yes
	2	No known links. Different city from P8.	442423384	No	No
	3	No known links.	Orphan	No	Yes
	4	No known links.	Orphan	No	Yes
	5	No known links.	Orphan	No	Yes
	6	No known links.	Orphan	No	Yes
	7	No known links.	Orphan	No	Yes
	8	No known links. Different city from P2.	442423384	No	No
	9	No known links.	Orphan	No	Yes
3	1	Member of family 1	456443362	Yes	Yes
	2	No known links	Orphan	No	Yes
	3	Contact with Family 1	456443362	Yes	Yes
	4	Public bars, no contact with family 1.	Orphan	No	Yes
	5	Contact with family 1	456443362	Yes	Yes
	6	Public bars, no contact with family 1.	Orphan	No	Yes
	7	Previous TB outside UK.	Orphan	No	Yes
	8	Public bars, known secondary contact with family 1.	456443362	Yes	Yes
4	1	Contact with three families.	434443183	Yes	Yes
	2	Contact with two patients.	434443183	Yes	Yes
	3	Hospital worker - no known links.	434443183	No	No
	4	Public bars, no specific contacts.	434443183	No	No
	5	Contact with patient.	434443183	Yes	Yes
	6	Contact with patient.	434443183	Yes	Yes
	7	No known links.	434443183	No	No
	8	Contact of family.	434443183	Yes	Yes
	9	Contact with P5.	434443183	Yes	Yes
	10	No known links.	Orphan	No	Yes
	11	No known links.	434443183	No	No

Cluster	Patient	Conventional Epidemiological Data of each Patient	Extra nine loci profile	Concordance between molecular and conventional epidemiological data using	
				ETR+MIRU	VNTR+
5	1	Member of family 2	443443153	Yes	Yes
	2	Hostel	443443153	Yes	Yes
	3	Hostel	443443153	Yes	Yes
	4	Hostel	443443153	Yes	Yes
	5	Member of family 2	443443153	Yes	Yes
	6	Member of family 2	443443153	Yes	Yes
	7	Public bar and partner was culture +ve	443443153	Yes	Yes
	8	Hostel	443443153	Yes	Yes
	9	Elderly patient	443443153	No	No
	10	Hostel – family 3	443443153	Yes	Yes
	11	Public bars – no specific links	Orphan	No	Yes
	12	Hostel	443443153	Yes	Yes
	13	Member of family 3	443443153	Yes	Yes
	14	Member of family 3	443443153	Yes	Yes
	15	Father was culture positive	443443153	Yes	Yes
	16	Household contact of P22 and P24.	443443153	Yes	Yes
	17	Contact of culture positive patient in same public bar.	443443153	Yes	Yes
	18	Hostel	443443153	Yes	Yes
	19	No known links	Orphan	No	Yes
	20	No known links	Orphan	No	Yes
	21	Worked with P23& ex-partner was culture positive.	442443151	Yes	Yes
	22	Household contact of P16+P24	443443153	Yes	Yes
	23	Worked with P21.	442443151	Yes	Yes
	24	Household contact of P16+P22.	443443153	Yes	Yes
6	1	Warehouse	236423252	Yes	Yes
	2	Warehouse	236423252	Yes	Yes
	3	Warehouse	236423252	Yes	Yes
	4	Warehouse	236423252	Yes	Yes
7	1	No known link	Orphan	No	Yes
	2	Pubs	442423382	Yes	Yes
	3	Prison	442423382	Yes	Yes
	4	No known link	Orphan	No	Yes
	5	Member of family 4.	442423382	Yes	Yes
	6	Known contact of family 4	442423382	Yes	Yes
Concordance between molecular and conventional epidemiological data for all seven clusters			Yes	37	59
			No	27	7

REFERENCES

Abadia,E., Zhang,J., dos Vultos,T., Ritacco,V., Kremer,K., Aktas,E. *et al.* (2010) Resolving lineage assignation on *Mycobacterium tuberculosis* clinical isolates classified by spoligotyping with a new high-throughput 3R SNPs based method. *Infect Genet Evol* **10**: 1066-1074.

Abou-Zeid,C., Garbe,T., Lathigra,R., Wiker,H.G., Harboe,M., Rook,G.A., and Young,D.B. (1991) Genetic and immunological analysis of *Mycobacterium tuberculosis* fibronectin-binding proteins. *Infect Immun* **59**: 2712-2718.

Abubakar,I., Moore,J., Drobniewski,F., Kruijshaar,M., Brown,T., Yates,M. *et al.* (2009) Extensively drug-resistant tuberculosis in the UK: 1995 to 2007. *Thorax* **64**: 512-515.

Akhtar,P., Singh,S., Bifani,P., Kaur,S., Srivastava,B.S., and Srivastava,R. (2009) Variable-number tandem repeat 3690 polymorphism in Indian clinical isolates of *Mycobacterium tuberculosis* and its influence on transcription. *J Med Microbiol* **58**: 798-805.

Al-Hajoj,S.A., Zozio,T., Al-Rabiah,F., Mohammad,V., Al-Nasser,M., Sola,C., and Rastogi,N. (2007) First insight into the population structure of *Mycobacterium tuberculosis* in Saudi Arabia. *J Clin Microbiol* **45**: 2467-2473.

Alexander,K.A., Laver,P.N., Michel,A.L., Williams,M., van Helden,P.D., Warren,R.M., and Gey van Pittius,N.C. (2010) Novel *Mycobacterium tuberculosis* complex pathogen, M. mungi. *Emerg Infect Dis* **16**: 1296-1299.

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

Alland,D., Lacher,D.W., Hazbon,M.H., Motiwala,A.S., Qi,W., Fleischmann,R.D., and Whittam,T.S. (2007) Role of large sequence polymorphisms (LSPs) in generating genomic diversity among clinical isolates of *Mycobacterium tuberculosis* and the utility of LSPs in phylogenetic analysis. *J Clin Microbiol* **45**: 39-46.

Allix,C., Supply,P., and Fauville-Dufaux,M. (2004) Utility of fast mycobacterial interspersed repetitive unit-variable number tandem repeat genotyping in clinical mycobacteriological analysis. *Clin Infect Dis* **39**: 783-789.

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

Allix-Beguec,C., Harmsen,D., Weniger,T., Supply,P., and Niemann,S. (2008b) Evaluation and strategy for use of MIRU-VNTRplus, a multifunctional database for online analysis of genotyping data and phylogenetic identification of *Mycobacterium tuberculosis* complex isolates. *J Clin Microbiol* **46**: 2692-2699.

Almeida,D., Rodrigues,C., Ashavaid,T.F., Lalvani,A., Udwadia,Z.F., and Mehta,A. (2005) High incidence of the Beijing genotype among multidrug-resistant isolates of *Mycobacterium tuberculosis* in a tertiary care center in Mumbai, India. *Clin Infect Dis* **40**: 881-886.

Alonso,M., Borrell,S., Lirola,M.M., Bouza,E., and Garcia de Viedma,D. (2008) A proposal for applying molecular markers as an aid to identifying potential cases of imported tuberculosis in immigrants. *Tuberculosis (Edinb)* **88**: 641-647.

Alonso-Rodriguez,N., Martinez-Lirola,M., Herranz,M., Sanchez-Benitez,M., Barroso,P., Bouza,E., and Garcia de Viedma,D. (2008) Evaluation of the new advanced 15-loci MIRU-VNTR genotyping tool in *Mycobacterium tuberculosis* molecular epidemiology studies. *BMC Microbiol* **8**: 34.

Alonso-Rodriguez,N., Martinez-Lirola,M., Sanchez,M.L., Herranz,M., Penafiel,T., Bonillo,M.C. *et al.* (2009) Prospective universal application of mycobacterial interspersed repetitive-unit-variable-number tandem-repeat genotyping to characterize *Mycobacterium tuberculosis* isolates for fast identification of clustered and orphan cases. *J Clin Microbiol* **47**: 2026-2032.

Andersen,P., Munk,M.E., Pollock,J.M., and Doherty,T.M. (2000) Specific immune-based diagnosis of tuberculosis. *Lancet* **356**: 1099-1104.

Aranaz,A., Liebana,E., Gomez-Mampaso,E., Galan,J.C., Cousins,D., Ortega,A. *et al.* (1999) *Mycobacterium tuberculosis* subsp. *caprae* subsp. nov.: a taxonomic study of a new member of the *Mycobacterium tuberculosis* complex isolated from goats in Spain. *Int J Syst Bacteriol* **49**: 1263-1273.

Arnold,C., Thorne,N., Underwood,A., Baster,K., and Gharbia,S. (2006) Evolution of short sequence repeats in *Mycobacterium tuberculosis*. *FEMS Microbiol Lett* **256**: 340-346.

Baboolal,S., Millet,J., Akpaka,P.E., Ramoutar,D., and Rastogi,N. (2009) First insight into *Mycobacterium tuberculosis* epidemiology and genetic diversity in Trinidad and Tobago. *J Clin Microbiol* **47**: 1911-1914.

Baker,L., Brown,T., Maiden,M.C., and Drobniewski,F. (2004) Silent nucleotide polymorphisms and a phylogeny for *Mycobacterium tuberculosis*. *Emerg Infect Dis* **10**: 1568-1577.

Balasubramanian,S., and Bentley,D. Polynucleotide arrays and their use in sequencing. PCT/GB2001/000407(WO/2001/057248). 31/01/2001. Great Britain.

Balganesh,M., Kuruppath,S., Marcel,N., Sharma,S., Nair,A., and Sharma,U. (2010) Rv1218c, an ABC transporter of *Mycobacterium tuberculosis* with implications in drug discovery. *Antimicrob Agents Chemother* **54**: 5167-5172.

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

Banu,S., Honore,N., Saint-Joanis,B., Philpott,D., Prevost,M.C., and Cole,S.T. (2002) Are the PE-PGRS proteins of *Mycobacterium tuberculosis* variable surface antigens? *Mol Microbiol* **44**: 9-19.

Barlow,R.E., Gascoyne-Binzi,D.M., Gillespie,S.H., Dickens,A., Qamer,S., and Hawkey,P.M. (2001) Comparison of variable number tandem repeat and IS*6110*-restriction fragment length polymorphism analyses for discrimination of high- and low-copy-number IS*6110 Mycobacterium tuberculosis* isolates. *J Clin Microbiol* **39**: 2453-2457.

Barnes,P.F., and Cave,M.D. (2003) Molecular epidemiology of tuberculosis. *N Engl J Med* **349**: 1149-1156.

Barrangou,R., Fremaux,C., Deveau,H., Richards,M., Boyaval,P., Moineau,S. *et al.* (2007) CRISPR provides acquired resistance against viruses in prokaryotes. *Science* **315**: 1709-1712.

Barry,C.E., III, Wilson,M., Lee,R., and Schoolnik,G.K. (2000) DNA microarrays and combinatorial chemical libraries: tools for the drug discovery pipeline. *Int J Tuberc Lung Dis* **4**: S189-S193.

Behr,M.A. (2002) BCG--different strains, different vaccines? *Lancet Infect Dis* **2**: 86-92.

Behr,M.A., Wilson,M.A., Gill,W.P., Salamon,H., Schoolnik,G.K., Rane,S., and Small,P.M. (1999) Comparative genomics of BCG vaccines by whole-genome DNA microarray. *Science* **284**: 1520-1523.

Bentley,D.R. (2006) Whole-genome re-sequencing. *Curr Opin Genet Dev* **16**: 545-552.

Berry,M.P., Graham,C.M., McNab,F.W., Xu,Z., Bloch,S.A., Oni,T. *et al.* (2010) An interferon-inducible neutrophil-driven blood transcriptional signature in human tuberculosis. *Nature* **466**: 973-977.

Betts,J.C., Dodson,P., Quan,S., Lewis,A.P., Thomas,P.J., Duncan,K., and McAdam,R.A. (2000) Comparison of the proteome of *Mycobacterium tuberculosis* strain H37Rv with clinical isolate CDC 1551. *Microbiology* **146**: 3205-3216.

Betts,J.C., McLaren,A., Lennon,M.G., Kelly,F.M., Lukey,P.T., Blakemore,S.J., and Duncan,K. (2003) Signature gene expression profiles discriminate between isoniazid-, thiolactomycin-, and triclosan-treated *Mycobacterium tuberculosis*. *Antimicrob Agents Chemother* **47**: 2903-2913.

Bhanu,N.V., van Soolingen,D., van Embden,J.D., Dar,L., Pandey,R.M., and Seth,P. (2002) Predominace of a novel *Mycobacterium tuberculosis* genotype in the Delhi region of India. *Tuberculosis (Edinb)* **82**: 105-112.

Bifani,P., Moghazeh,S., Shopsin,B., Driscoll,J., Ravikovitch,A., and Kreiswirth,B.N. (2000) Molecular characterization of *Mycobacterium tuberculosis* H37Rv/Ra variants: distinguishing the mycobacterial laboratory strain. *J Clin Microbiol* **38**: 3200-3204.

Bifani,P.J., Mathema,B., Liu,Z., Moghazeh,S.L., Shopsin,B., Tempalski,B. *et al.* (1999) Identification of a W variant outbreak of *Mycobacterium tuberculosis* via population-based molecular epidemiology. *JAMA* **282**: 2321-2327.

Bifani,P.J., Plikaytis,B.B., Kapur,V., Stockbauer,K., Pan,X., Lutfey,M.L. *et al.* (1996) Origin and interstate spread of a New York City multidrug-resistant *Mycobacterium tuberculosis* clone family. *JAMA* **275**: 452-457.

Bikandi,J., San,M.R., Rementeria,A., and Garaizar,J. (2004) In silico analysis of complete bacterial genomes: PCR, AFLP-PCR and endonuclease restriction. *Bioinformatics* **20**: 798-799.

Blackwood,K.S., Wolfe,J.N., and Kabani,A.M. (2004) Application of mycobacterial interspersed repetitive unit typing to Manitoba tuberculosis cases: can restriction fragment length polymorphism be forgotten? *J Clin Microbiol* **42**: 5001-5006.

Boehme,C.C., Nabeta,P., Hillemann,D., Nicol,M.P., Shenai,S., Krapp,F. *et al.* (2010) Rapid molecular detection of tuberculosis and rifampin resistance. *N Engl J Med* **363**: 1005-1015.

Borgatti,S.P., Everett,M.G., and Reeman,L.C. (1999) UCINET Analytic Technologies.

Borgdorff,M.W., Nagelkerke,N.J., de Haas,P.E., and van Soolingen,D. (2001) Transmission of *Mycobacterium tuberculosis* depending on the age and sex of source cases. *Am J Epidemiol* **154**: 934-943.

Brandli,O. (1998) The clinical presentation of tuberculosis. *Respiration* **65**: 97-105.

Brennan,M.J., Delogu,G., Chen,Y., Bardarov,S., Kriakov,J., Alavi,M., and Jacobs,W.R., Jr. (2001) Evidence that mycobacterial PE_PGRS proteins are cell surface constituents that influence interactions with other cells. *Infect Immun* **69**: 7326-7333.

Breslauer,D.N., Maamari,R.N., Switz,N.A., Lam,W.A., and Fletcher,D.A. (2009) Mobile phone based clinical microscopy for global health applications. *PLoS One* **4**: e6320.

Brosch,R., Gordon,S.V., Garnier,T., Eiglmeier,K., Frigui,W., Valenti,P. *et al.* (2007) Genome plasticity of BCG and impact on vaccine efficacy. *Proc Natl Acad Sci U S A* **104**: 5596-5601.

Brosch,R., Gordon,S.V., Marmiesse,M., Brodin,P., Buchrieser,C., Eiglmeier,K. *et al.* (2002) A new evolutionary scenario for the *Mycobacterium tuberculosis* complex. *Proc Natl Acad Sci U S A* **99**: 3684-3689.

Brosch,R., Philipp,W.J., Stavropoulos,E., Colston,M.J., Cole,S.T., and Gordon,S.V. (1999) Genomic analysis reveals variation between *Mycobacterium tuberculosis* H37Rv and the attenuated *M. tuberculosis* H37Ra strain. *Infect Immun* **67**: 5768-5774.

Brosch,R., Pym,A.S., Gordon,S.V., and Cole,S.T. (2001) The evolution of mycobacterial pathogenicity: clues from comparative genomics. *Trends Microbiol* **9**: 452-458.

Brown,T., Nikolayevskyy,V., Velji,P., and Drobniewski,F. (2010) Associations between *Mycobacterium tuberculosis* strains and phenotypes. *Emerg Infect Dis* **16**: 272-280.

Brudey,K., Driscoll,J.R., Rigouts,L., Prodinger,W.M., Gori,A., Al-Hajoj,S.A. *et al.* (2006) *Mycobacterium tuberculosis* complex genetic diversity: mining the fourth international spoligotyping database (SpolDB4) for classification, population genetics and epidemiology. *BMC Microbiol* **6**: 23.

Brudey,K., Gordon,M., Mostrom,P., Svensson,L., Jonsson,B., Sola,C. *et al.* (2004) Molecular epidemiology of *Mycobacterium tuberculosis* in western Sweden. *J Clin Microbiol* **42**: 3046-3051.

Bull,T.J., Sidi-Boumedine,K., McMinn,E.J., Stevenson,K., Pickup,R., and Hermon-Taylor,J. (2003) Mycobacterial interspersed repetitive units (MIRU) differentiate *Mycobacterium avium* subspecies *paratuberculosis* from other species of the *Mycobacterium avium* complex. *Mol Cell Probes* **17**: 157-164.

Candia,N., Lopez,B., Zozio,T., Carrivale,M., Diaz,C., Russomando,G. *et al.* (2007) First insight into *Mycobacterium tuberculosis* genetic diversity in Paraguay. *BMC Microbiol* **7:75.**: 75.

Canetti,G., Fox,W., Khomenko,A., Mahler,H.T., Menon,N.K., Mitchison,D.A. *et al.* (1969) Advances in techniques of testing mycobacterial drug sensitivity, and the use of sensitivity tests in tuberculosis control programmes. *Bull World Health Organ* **41**: 21-43.

Cardoso Oelemann,M., Gomes,H.M., Willery,E., Possuelo,L., Batista Lima,K.V., Allix-Béguec,C. *et al* (2011). The forest behind the tree: phylogenetic exploration of a dominant *Mycobacterium tuberculosis* strain lineage from a high tuberculosis burden country. *PLoS One* **6**: e18256.

Centers for Disease Control and Prevention (CDC). (2006) Notice to Readers: Revised Definition of Extensively Drug-Resistant Tuberculosis. *MMWR Morb Mortal Wkly Rep* **55**: 1176.

Chatterjee,A., D'Souza,D., Vira,T., Bamne,A., Ambe,G.T., Nicol,M.P. *et al.* (2010) Strains of *Mycobacterium tuberculosis* from western Maharashtra, India, exhibit a high degree of diversity and strain-specific associations with drug resistance, cavitary disease, and treatment failure. *J Clin Microbiol* **48**: 3593-3599.

Cheah,E.S., Malkin,J., Free,R.C., Lee,S.M., Perera,N., Woltmann,G. *et al.* (2010) A two-tube combined TaqMan/SYBR Green assay to identify mycobacteria and detect single global lineage-defining polymorphisms in *Mycobacterium tuberculosis*. *J Mol Diagn* **12**: 250-256.

Choy,Y.S., Dabora,S.L., Hall,F., Ramesh,V., Niida,Y., Franz,D. *et al.* (1999) Superiority of denaturing high performance liquid chromatography over single-stranded conformation and conformation-sensitive gel electrophoresis for mutation detection in TSC2. *Ann Hum Genet* **63**: 383-391.

Christianson,S., Wolfe,J., Orr,P., Karlowsky,J., Levett,P.N., Horsman,G.B. *et al.* (2010) Evaluation of 24 locus MIRU-VNTR genotyping of *Mycobacterium tuberculosis* isolates in Canada. *Tuberculosis (Edinb)* **90**: 31-38.

Colditz,G.A., Brewer,T.F., Berkey,C.S., Wilson,M.E., Burdick,E., Fineberg,H.V., and Mosteller,F. (1994) Efficacy of BCG vaccine in the prevention of tuberculosis. Meta-analysis of the published literature. *JAMA* **271**: 698-702.

Cole,S.T., Brosch,R., Parkhill,J., Garnier,T., Churcher,C., Harris,D. *et al.* (1998) Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* **393**: 537-544.

Cole,S.T., Eiglmeier,K., Parkhill,J., James,K.D., Thomson,N.R., Wheeler,P.R. *et al.* (2001) Massive gene decay in the leprosy bacillus. *Nature* **409**: 1007-1011.

Collins,H.L., and Kaufmann,S.H. (2001) Prospects for better tuberculosis vaccines. *Lancet Infect Dis* **1**: 21-28.

Collyns,T.A., Gascoyne-Binzi,D.M., and Hawkey,P.M. (2002) Molecular fingerprinting of *Mycobacterium tuberculosis*: does it help in understanding the epidemiology of tuberculosis? *Rev Med Microbiol* **13**: 119-127.

Comas,I., Homolka,S., Niemann,S., and Gagneux,S. (2009) Genotyping of genetically monomorphic bacteria: DNA sequencing in *Mycobacterium tuberculosis* highlights the limitations of current methodologies. *PLoS One* **4**: e7815.

Cooksey,R.C., Limor,J., Morlock,G.P., and Crawford,J.T. (2003) Identifying Mycobacterium species and strain typing using a microfluidic labchip instrument. *Biotechniques* **35**: 786-794.

Cooksey,R.C., Morlock,G.P., Holloway,B.P., Limor,J., and Hepburn,M. (2002) Temperature-mediated heteroduplex analysis performed by using denaturing high-performance liquid chromatography to identify sequence polymorphisms in *Mycobacterium tuberculosis* complex organisms. *J Clin Microbiol* **40**: 1610-1616.

Coronado,V.G., Beck-Sague,C.M., Hutton,M.D., Davis,B.J., Nicholas,P., Villareal,C. *et al.* (1993) Transmission of multidrug-resistant *Mycobacterium tuberculosis* among persons with human immunodeficiency virus infection in an urban hospital: epidemiologic and restriction fragment length polymorphism analysis. *J Infect Dis* **168**: 1052-1055.

Cosma,C.L., Sherman,D.R., and Ramakrishnan,L. (2003) The secret lives of the pathogenic mycobacteria. *Annu Rev Microbiol* **57**: 641-676.

Cowan,L.S., and Crawford,J.T. (2002) Genotype analysis of *Mycobacterium tuberculosis* isolates from a sentinel surveillance population. *Emerg Infect Dis* **8**: 1294-1302.

Cowan,L.S., Diem,L., Brake,M.C., and Crawford,J.T. (2004) Transfer of a *Mycobacterium tuberculosis* genotyping method, Spoligotyping, from a reverse line-blot hybridization, membrane-based assay to the Luminex multianalyte profiling system. *J Clin Microbiol* **42**: 474-477.

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

Cowan,S.T., and Steele,K.J. (1993) Characters of Gram-positive bacteria. In Manual for the identification of medical bacteria. Barrow,G.I., and Feltham,R.K.A. (eds). Cambridge University Press, pp. 50-93.

Crawford,J.T., and Bates,J.H. (1979) Isolation of plasmids from mycobacteria. *Infect Immun* **24**: 979-981.

Crofts,J.P., Gelb,D., Andrews,N., Delpech,V., Watson,J.M., and Abubakar,I. (2008) Investigating tuberculosis trends in England. *Public Health* **122**: 1302-1310.

Davies,P.D. (2009) Why universal BCG in UK was deemed not necessary. *BMJ* **338**: b192.

Davis,E.O., Sedgwick,S.G., and Colston,M.J. (1991) Novel structure of the recA locus of *Mycobacterium tuberculosis* implies processing of the gene product. *J Bacteriol* **173**: 5653-5662.

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

de Jong,B.C., Antonio,M., Awine,T., Ogungbemi,K., de Jong,Y.P., Gagneux,S. *et al.* (2009) Use of spoligotyping and large sequence polymorphisms to study the population structure of the *Mycobacterium tuberculosis* complex in a cohort study of consecutive smear-positive tuberculosis cases in The Gambia. *J Clin Microbiol* **47**: 994-1001.

de Vries,G., van Hest,R.A., and Richardus,J.H. (2007) Impact of mobile radiographic screening on tuberculosis among drug users and homeless persons. *Am J Respir Crit Care Med* **176**: 201-207.

Devaney,J.M., Girard,J.E., and Marino,M.A. (2000) DNA microsatellite analysis using ion-pair reversed-phase high-performance liquid chromatography. *Anal Chem* **72**: 858-864.

Dinnes,J., Deeks,J., Kunst,H., Gibson,A., Cummins,E., Waugh,N. *et al.* (2007) A systematic review of rapid diagnostic tests for the detection of tuberculosis infection. *Health Technol Assess* **11**: 1-196.

Dong,H., Liu,Z., Lv,B., Zhang,Y., Liu,J., Zhao,X. *et al.* (2010) Spoligotypes of *Mycobacterium tuberculosis* from different Provinces of China. *J Clin Microbiol* **48**: 4102-4106.

Donoghue,H.D. (2009) Human tuberculosis--an ancient disease, as elucidated by ancient microbial biomolecules. *Microbes Infect* **11**: 1156-1162.

dos Vultos T., Mestre,O., Rauzier,J., Golec,M., Rastogi,N., Rasolofo,V. *et al.* (2008) Evolution and diversity of clonal bacteria: the paradigm of *Mycobacterium tuberculosis*. *PLoS One* **3**: e1538.

Dou,H.Y., Lu,J.J., Lin,C.W., Chang,J.R., Sun,J.R., and Su,I.J. (2009) Utility and evaluation of new variable-number tandem-repeat systems for genotyping mycobacterial tuberculosis isolates. *J Microbiol Methods* **77**: 127-129.

Drobniewski,F., Rusch-Gerdes,S., and Hoffner,S. (2007) Antimicrobial susceptibility testing of *Mycobacterium tuberculosis* (EUCAST document E.DEF 8.1)--report of the Subcommittee on Antimicrobial Susceptibility Testing of *Mycobacterium tuberculosis* of the European Committee for Antimicrobial Susceptibility Testing (EUCAST) of the European Society of Clinical Microbiology and Infectious Diseases (ESCMID). *Clin Microbiol Infect* **13**: 1144-1156.

Duchene,V., Ferdinand,S., Filliol,I., Guegan,J.F., Rastogi,N., and Sola,C. (2004) Phylogenetic reconstruction of *Mycobacterium tuberculosis* within four settings of the Caribbean region: tree comparative analyse and first appraisal on their phylogeography. *Infect Genet Evol* **4**: 5-14.

Eaves,D.J., Liebana,E., Woodward,M.J., and Piddock,L.J. (2002) Detection of *gyrA* mutations in quinolone-resistant *Salmonella enterica* by denaturing high-performance liquid chromatography. *J Clin Microbiol* **40**: 4121-4125.

Edlin,B.R., Tokars,J.I., Grieco,M.H., Crawford,J.T., Williams,J., Sordillo,E.M. *et al.* (1992) An outbreak of multidrug-resistant tuberculosis among hospitalized patients with the acquired immunodeficiency syndrome. *N Engl J Med* **326**: 1514-1521.

Eisenach,K.D., Crawford,J.T., and Bates,J.H. (1988) Repetitive DNA sequences as probes for *Mycobacterium tuberculosis*. *J Clin Microbiol* **26**: 2240-2245.

Eldholm,V., Matee,M., Mfinanga,S.G., Heun,M., and Dahle,U.R. (2006) A first insight into the genetic diversity of *Mycobacterium tuberculosis* in Dar es Salaam, Tanzania, assessed by spoligotyping. *BMC Microbiol* **6**: 76.

Espitia,C., Laclette,J.P., Mondragon-Palomino,M., Amador,A., Campuzano,J., Martens,A. *et al.* (1999) The PE-PGRS glycine-rich proteins of *Mycobacterium tuberculosis*: a new family of fibronectin-binding proteins? *Microbiology* **145**: 3487-3495.

European Centre for Disease Prevention and Control/WHO Regional Office for Europe. Tuberculosis surveillance in Europe 2009. 1-3-2011. Stockholm, European Centre for Disease Prevention and Control. 3-5-2011. Available from http://www.ecdc.europa.eu/en/publications/Publications/1103_TB_SUR_2009.pdf

Evaluations and Standards Laboratory. Investigation of specimens for *Mycobacterium* species. National Standard Method BSOP 40 Issue 5. 2006. Health Protection Agency. 3-6-2011.
Available from http://www.hpa-standardmethods.org.uk/documents/bsop/pdf/bsop40.pdf

Evans,J.T., Gibson,A., Khanom,S., Overton-Lewis,C., Smith,E.G., and Hawkey,P.M. (2011) Introduction of 24-loci MIRU-VNTR typing in a region-wide scheme considerably reduces the number of cases requiring epidemiological investigation. *Clin Microbiol Infect* **17**: In press.

Evans,J.T., Hawkey,P.M., Smith,E.G., Boese,K.A., Warren,R.E., and Hong,G. (2004) Automated high-throughput mycobacterial interspersed repetitive unit typing of

Mycobacterium tuberculosis strains by a combination of PCR and nondenaturing high-performance liquid chromatography. *J Clin Microbiol* **42**: 4175-4180.

Evans,J.T., Smith,E.G., Banerjee,A., Smith,R.M., Dale,J., Innes,J.A. *et al.* (2007) Cluster of human tuberculosis caused by *Mycobacterium bovis*: evidence for person-to-person transmission in the UK. *Lancet* **369**: 1270-1276.

Ewer,K., Deeks,J., Alvarez,L., Bryant,G., Waller,S., Andersen,P. *et al.* (2003) Comparison of T-cell-based assay with tuberculin skin test for diagnosis of *Mycobacterium tuberculosis* infection in a school tuberculosis outbreak. *Lancet* **361**: 1168-1173.

Fabre,M., Koeck,J.L., Le,F.P., Simon,F., Herve,V., Vergnaud,G., and Pourcel,C. (2004) High genetic diversity revealed by variable-number tandem repeat genotyping and analysis of hsp65 gene polymorphism in a large collection of "*Mycobacterium canettii*" strains indicates that the *M. tuberculosis* complex is a recently emerged clone of "*M. canettii*". *J Clin Microbiol* **42**: 3248-3255.

Filliol,I., Driscoll,J.R., van Soolingen,D., Kreiswirth,B.N., Kremer,K., Valetudie,G. *et al.* (2002) Global distribution of *Mycobacterium tuberculosis* spoligotypes. *Emerg Infect Dis* **8**: 1347-1349.

Filliol,I., Motiwala,A.S., Cavatore,M., Qi,W., Hazbon,M.H., Bobadilla del Valle,M. *et al.* (2006) Global phylogeny of *Mycobacterium tuberculosis* based on single nucleotide polymorphism (SNP) analysis: insights into tuberculosis evolution, phylogenetic accuracy of other DNA fingerprinting systems, and recommendations for a minimal standard SNP set. *J Bacteriol* **188**: 759-772.

Fischer,S.G., and Lerman,L.S. (1979) Length-independent separation of DNA restriction fragments in two-dimensional gel electrophoresis. *Cell* **16**: 191-200.

Fleischmann,R.D., Alland,D., Eisen,J.A., Carpenter,L., White,O., Peterson,J. *et al.* (2002) Whole-genome comparison of *Mycobacterium tuberculosis* clinical and laboratory strains. *J Bacteriol* **184**: 5479-5490.

Fok,A., Numata,Y., Schulzer,M., and FitzGerald,M.J. (2008) Risk factors for clustering of tuberculosis cases: a systematic review of population-based molecular epidemiology studies. *Int J Tuberc Lung Dis* **12**: 480-492.

Fomukong,N.G., Tang,T.H., al-Maamary,S., Ibrahim,W.A., Ramayah,S., Yates,M. *et al.* (1994) Insertion sequence typing of *Mycobacterium tuberculosis*: characterization of a widespread subtype with a single copy of IS*6110*. *Tuber Lung Dis* **75**: 435-40.

Ford,C.B., Lin,P.L., Chase,M.R., Shah,R.R., Iartchouk,O., Galagan,J. *et al.* (2011) Use of whole genome sequencing to estimate the mutation rate of *Mycobacterium tuberculosis* during latent infection. *Nat Genet* **43**: 482-486.

Fox,W., Ellard,G.A., and Mitchison,D.A. (1999) Studies on the treatment of tuberculosis undertaken by the British Medical Research Council tuberculosis units, 1946-1986, with relevant subsequent publications. *Int J Tuberc Lung Dis* **3**: S231-S279.

Freeman,R., Kato-Maeda,M., Hauge,K.A., Horan,K.L., Oren,E., Narita,M. *et al.* (2005) Use of rapid genomic deletion typing to monitor a tuberculosis outbreak within an urban homeless population. *J Clin Microbiol* **43**: 5550-5554.

Freidlin,P.J., Goldblatt,D., Kaidar-Shwartz,H., and Rorman,E. (2009) Polymorphic exact tandem repeat A (PETRA): a newly defined lineage of *Mycobacterium tuberculosis* in israel originating predominantly in Sub-Saharan Africa. *J Clin Microbiol* **47**: 4006-4020.

Frota,C.C., Hunt,D.M., Buxton,R.S., Rickman,L., Hinds,J., Kremer,K. *et al.* (2004) Genome structure in the vole bacillus, *Mycobacterium microti*, a member of the *Mycobacterium tuberculosis* complex with a low virulence for humans. *Microbiology* **150**: 1519-1527.

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

Frothingham,R., Strickland,P.L., Bretzel,G., Ramaswamy,S., Musser,J.M., and Williams,D.L. (1999) Phenotypic and genotypic characterization of *Mycobacterium africanum* isolates from West Africa. *J Clin Microbiol* **37**: 1921-1926.

Gagneux,S., DeRiemer,K., Van,T., Kato-Maeda,M., de Jong,B.C., Narayanan,S. *et al.* (2006) Variable host-pathogen compatibility in *Mycobacterium tuberculosis*. *Proc Natl Acad Sci U S A* **103**: 2869-2873.

Gagneux,S., and Small,P.M. (2007) Global phylogeography of *Mycobacterium tuberculosis* and implications for tuberculosis product development. *Lancet Infect Dis* **7**: 328-337.

Garcia de Viedma,D., Alonso,R.N., Andres,S., Martinez,L.M., Ruiz Serrano,M.J., and Bouza,E. (2006) Evaluation of alternatives to RFLP for the analysis of clustered cases of tuberculosis. *Int J Tuberc Lung Dis* **10**: 454-459.

Garcia-Pelayo,M.C., Caimi,K.C., Inwald,J.K., Hinds,J., Bigi,F., Romano,M.I. *et al.* (2004) Microarray analysis of *Mycobacterium microti* reveals deletion of genes encoding PE-PPE proteins and ESAT-6 family antigens. *Tuberculosis (Edinb)* **84**: 159-166.

Gardy,J.L., Johnston,J.C., Ho Sui,S.J., Cook,V.J., Shah,L., Brodkin,E. *et al.* (2011) Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. *N Engl J Med* **364**: 730-739.

Garnier,T., Eiglmeier,K., Camus,J.C., Medina,N., Mansoor,H., Pryor,M. *et al.* (2003) The complete genome sequence of *Mycobacterium bovis*. *Proc Natl Acad Sci U S A* **100**: 7877-7882.

Gascoyne-Binzi,D.M., Barlow,R.E., Essex,A., Gelletlie,R., Khan,M.A., Hafiz,S. *et al.* (2002) Predominant VNTR family of strains of *Mycobacterium tuberculosis* isolated from South Asian patients. *Int J Tuberc Lung Dis* **6**: 492-496.

Gascoyne-Binzi,D.M., Barlow,R.E., Frothingham,R., Robinson,G., Collyns,T.A., Gelletlie,R., and Hawkey,P.M. (2001) Rapid identification of laboratory contamination with

Mycobacterium tuberculosis using variable number tandem repeat analysis. *J Clin Microbiol* **39**: 69-74.

Genewein,A., Telenti,A., Bernasconi,C., Mordasini,C., Weiss,S., Maurer,A.M. *et al.* (1993) Molecular approach to identifying route of transmission of tuberculosis in the community. *Lancet* **342**: 841-844.

Gibson,A., Evans,J.T., Gardiner,S., Smith,E.G., and Hawkey,P.M. (2010) Analysis of an additional 9-loci MIRU-VNTR confirms a large single clone of *Mycobacterium tuberculosis* in the West Midlands, United Kingdom. *Clin Microbiol Infect* **16**: S605.

Gibson,A.L., Huard,R.C., Gey van Pittius,N.C., Lazzarini,L.C., Driscoll,J., Kurepina,N. *et al.* (2008) Application of sensitive and specific molecular methods to uncover global dissemination of the major RDRio Sublineage of the Latin American-Mediterranean *Mycobacterium tuberculosis* spoligotype family. *J Clin Microbiol* **46**: 1259-1267.

Gilbert,R.L., Antoine,D., French,C.E., Abubakar,I., Watson,J.M., and Jones,J.A. (2009) The impact of immigration on tuberculosis rates in the United Kingdom compared with other European countries. *Int J Tuberc Lung Dis* **13**: 645-651.

Gillespie,S.H., Dickens,A., and McHugh,T.D. (2000) False molecular clusters due to nonrandom association of IS*6110* with *Mycobacterium tuberculosis*. *J Clin Microbiol* **38**: 2081-2086.

Glynn,J.R., Alghamdi,S., Mallard,K., McNerney,R., Ndlovu,R., Munthali,L. *et al.* (2010) Changes in *Mycobacterium tuberculosis* genotype families over 20 years in a population-based study in Northern Malawi. *PLoS One* **5**: e12259.

Godfrey-Faussett,P., Mortimer,P.R., Jenkins,P.A., and Stoker,N.G. (1992) Evidence of transmission of tuberculosis by DNA fingerprinting. *BMJ* **305**: 221-223.

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

Goguet de la Salmoniere YO, Kim,C.C., Tsolaki,A.G., Pym,A.S., Siegrist,M.S., and Small,P.M. (2004) High-throughput method for detecting genomic-deletion polymorphisms. *J Clin Microbiol* **42**: 2913-2918.

Gordon,S.V., Brosch,R., Billault,A., Garnier,T., Eiglmeier,K., and Cole,S.T. (1999) Identification of variable regions in the genomes of tubercle bacilli using bacterial artificial chromosome arrays. *Mol Microbiol* **32**: 643-655.

Goyal,M., Saunders,N.A., van Embden,J.D., Young,D.B., and Shaw,R.J. (1997) Differentiation of *Mycobacterium tuberculosis* isolates by spoligotyping and IS*6110* restriction fragment length polymorphism. *J Clin Microbiol* **35**: 647-651.

Grange,J.M., Collins,C.H., and McSwiggan,D. (1976) Bacteriophage typing of *Mycobacterium tuberculosis* strains isolated in South East England. *Tubercle* **57**: 59-66.

Groathouse,N.A., Rivoire,B., Kim,H., Lee,H., Cho,S.N., Brennan,P.J., and Vissa,V.D. (2004) Multiple polymorphic loci for molecular typing of strains of *Mycobacterium leprae*. *J Clin Microbiol* **42**: 1666-1672.

Groenen,P.M., Bunschoten,A.E., van Soolingen,D., and van Embden,J.D. (1993) Nature of DNA polymorphism in the direct repeat cluster of *Mycobacterium tuberculosis*; application for strain differentiation by a novel typing method. *Mol Microbiol* **10**: 1057-1065.

Gross,E., Arnold,N., Goette,J., Schwarz-Boeger,U., and Kiechle,M. (1999) A comparison of BRCA1 mutation analysis by direct sequencing, SSCP and DHPLC. *Hum Genet* **105**: 72-78.

Gutacker,M.M., Mathema,B., Soini,H., Shashkina,E., Kreiswirth,B.N., Graviss,E.A., and Musser,J.M. (2006) Single-nucleotide polymorphism-based population genetic analysis of *Mycobacterium tuberculosis* strains from 4 geographic sites. *J Infect Dis* **193**: 121-128.

Gutierrez,M.C., Ahmed,N., Willery,E., Narayanan,S., Hasnain,S.E., Chauhan,D.S. *et al.* (2006) Predominance of ancestral lineages of *Mycobacterium tuberculosis* in India. *Emerg Infect Dis* **12**: 1367-1374.

Gutierrez,M.C., Brisse,S., Brosch,R., Fabre,M., Omais,B., Marmiesse,M. *et al.* (2005) Ancient origin and gene mosaicism of the progenitor of *Mycobacterium tuberculosis*. *PLoS Pathog* **1**: e5.

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

Hanekom,M., van der Spuy,G.D., Gey van Pittius,N.C., McEvoy,C.R., Hoek,K.G., Ndabambi,S.L. *et al.* (2008) 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 of tuberculosis. *J Clin Microbiol* **46**: 3338-3345.

Hannachi-M'Zali,F., Ambler,J.E., Taylor,C.F., and Hawkey,P.M. (2002) Examination of single and multiple mutations involved in resistance to quinolones in *Staphylococcus aureus* by a combination of PCR and denaturing high-performance liquid chromatography (DHPLC). *J Antimicrob Chemother* **50**: 649-655.

Hardy,K.J., Ussery,D.W., Oppenheim,B.A., and Hawkey,P.M. (2004) Distribution and characterization of staphylococcal interspersed repeat units (SIRUs) and potential use for strain differentiation. *Microbiology* **150**: 4045-4052.

Hasan,Z., Tanveer,M., Kanji,A., Hasan,Q., Ghebremichael,S., and Hasan,R. (2006) Spoligotyping of *Mycobacterium tuberculosis* isolates from Pakistan reveals predominance of Central Asian Strain 1 and Beijing isolates. *J Clin Microbiol* **44**: 1763-1768.

Hawkey,P.M., and Kerr,K.G. (2003) Laboratory investigation of healthcare-associated infection. In Medical Bacteriology: A Practical Approach. Hawkey,P.M., and Lewis,D. (eds). Oxford: Oxford University Press, pp. 331-354.

Hawkey,P.M., Smith,E.G., Evans,J.T., Monk,P., Bryan,G., Mohamed,H.H. *et al.* (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.

Health Protection Agency (2010) Tuberculosis in the UK: Annual report on tuberculosis surveillance and control in the UK 2010. Health Protection Agency. Available from http://www.hpa.org.uk/web/HPAweb&HPAwebStandard/HPAweb_C/1287143581697

Health Protection Agency. The HPA National Tuberculosis Strain Typing Project. 1-2-2011. 14-5-2011. Available from http://www.hpa.org.uk/web/HPAwebFile/HPAweb_C/1296684114826

Helb,D., Jones,M., Story,E., Boehme,C., Wallace,E., Ho,K. *et al.* (2010) Rapid detection of *Mycobacterium tuberculosis* and rifampin resistance by use of on-demand, near-patient technology. *J Clin Microbiol* **48**: 229-237.

Hermans,P.W., van Soolingen,D., Bik,E.M., de Haas,P.E., Dale,J.W., and van Embden,J.D. (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.

Hermans,P.W., van Soolingen,D., Dale,J.W., Schuitema,A.R., McAdam,R.A., Catty,D., and van Embden,J.D. (1990) Insertion element IS986 from *Mycobacterium tuberculosis*: a useful tool for diagnosis and epidemiology of tuberculosis. *J Clin Microbiol* **28**: 2051-2058.

Hershberg,R., Lipatov,M., Small,P.M., Sheffer,H., Niemann,S., Homolka,S. *et al.* (2008) High functional diversity in *Mycobacterium tuberculosis* driven by genetic drift and human demography. *PLoS Biol* **6**: e311.

Hillemann,D., Rusch-Gerdes,S., and Richter,E. (2009) Feasibility of the GenoType MTBDRsl assay for fluoroquinolone, amikacin-capreomycin, and ethambutol resistance testing of *Mycobacterium tuberculosis* strains and clinical specimens. *J Clin Microbiol* **47**: 1767-1772.

Hillemann,D., Weizenegger,M., Kubica,T., Richter,E., and Niemann,S. (2005) Use of the genotype MTBDR assay for rapid detection of rifampin and isoniazid resistance in *Mycobacterium tuberculosis* complex isolates. *J Clin Microbiol* **43**: 3699-3703.

Hirsh,A.E., Tsolaki,A.G., DeRiemer,K., Feldman,M.W., and Small,P.M. (2004) Stable association between strains of *Mycobacterium tuberculosis* and their human host populations. *Proc Natl Acad Sci U S A* **101**: 4871-4876.

Ho,T.B., Robertson,B.D., Taylor,G.M., Shaw,R.J., and Young,D.B. (2000) Comparison of *Mycobacterium tuberculosis* genomes reveals frequent deletions in a 20 kb variable region in clinical isolates. *Yeast* **17**: 272-282.

Hogarth,P.J., Logan,K.E., Vordermeier,H.M., Singh,M., Hewinson,R.G., and Chambers,M.A. (2005) Protective immunity against *Mycobacterium bovis* induced by vaccination with Rv3109c--a member of the ESAT-6 gene family. *Vaccine* **23**: 2557-2564.

Honisch,C., Mosko,M., Arnold,C., Gharbia,S.E., Diel,R., and Niemann,S. (2010) Replacing Reverse Line-Blot Hybridization Spoligotyping of the *Mycobacterium tuberculosis* Complex. *J Clin Microbiol*.

Huard,R.C., Fabre,M., de Haas,P., Lazzarini,L.C., van Soolingen,D., Cousins,D., and Ho,J.L. (2006) Novel genetic polymorphisms that further delineate the phylogeny of the *Mycobacterium tuberculosis* complex. *J Bacteriol* **188**: 4271-87.

Hughes,A.L., Friedman,R., and Murray,M. (2002) Genomewide pattern of synonymous nucleotide substitution in two complete genomes of *Mycobacterium tuberculosis*. *Emerg Infect Dis* **8**: 1342-1346.

Hung,N.V., Sy,D.N., Anthony,R.M., Cobelens,F.G., and van Soolingen,D. (2007) Fluorescence microscopy for tuberculosis diagnosis. *Lancet Infect Dis* **7**: 238-239.

Hunter,P.R., and Gaston,M.A. (1988) Numerical index of the discriminatory ability of typing systems: an application of Simpson's index of diversity. *J Clin Microbiol* **26**: 2465-2466.

Ichikawa,K., Yagi,T., Inagaki,T., Moriyama,M., Nakagawa,T., Uchiya,K. *et al.* (2010) Molecular typing of *Mycobacterium intracellulare* using multilocus variable-number of tandem-repeat analysis: identification of loci and analysis of clinical isolates. *Microbiology* **156**: 496-504.

Ioerger,T.R., Feng,Y., Chen,X., Dobos,K.M., Victor,T.C., Streicher,E.M. *et al.* (2010) The non-clonality of drug resistance in Beijing-genotype isolates of *Mycobacterium tuberculosis* from the Western Cape of South Africa. *BMC Genomics* **11:670.**: 670.

Ioerger,T.R., Koo,S., No,E.G., Chen,X., Larsen,M.H., Jacobs,W.R., Jr. *et al.* (2009) Genome analysis of multi- and extensively-drug-resistant tuberculosis from KwaZulu-Natal, South Africa. *PLoS One* **4**: e7778.

Jackson,A.D., Seagar,A.L., Reid,M.E., Doig,C., Forbes,K.J., Laurenson,I.F., and McMenamin,J. (2009) Characterising transmission of a tuberculosis genotype in Scotland: a qualitative approach to social network enquiry. *Int J Tuberc Lung Dis* **13**: 486-493.

Jackson,P.J., Walthers,E.A., Kalif,A.S., Richmond,K.L., Adair,D.M., Hill,K.K. *et al.* (1997) Characterization of the variable-number tandem repeats in vrrA from different *Bacillus anthracis* isolates. *Appl Environ Microbiol* **63**: 1400-1405.

Jacobsen,M., Repsilber,D., Gutschmidt,A., Neher,A., Feldmann,K., Mollenkopf,H.J. *et al.* (2007) Candidate biomarkers for discrimination between infection and disease caused by *Mycobacterium tuberculosis*. *J Mol Med* **85**: 613-621.

Jagielski,T., Augustynowicz-Kopec,E., Zozio,T., Rastogi,N., and Zwolska,Z. (2010) Spoligotype-based comparative population structure analysis of multidrug-resistant and isoniazid-monoresistant *Mycobacterium tuberculosis* complex clinical isolates in Poland. *J Clin Microbiol* **48**: 3899-3909.

Jansen,R., Embden,J.D., Gaastra,W., and Schouls,L.M. (2002) Identification of genes that are associated with DNA repeats in prokaryotes. *Mol Microbiol* **43**: 1565-1575.

Jeffreys,A.J., Wilson,V., and Thein,S.L. (1985) Hypervariable 'minisatellite' regions in human DNA. *Nature* **314**: 67-73.

Kamerbeek,J., Schouls,L., Kolk,A., van Agterveld,M., van Soolingen,D., Kuijper,S. *et al.* (1997) Simultaneous detection and strain differentiation of *Mycobacterium tuberculosis* for diagnosis and epidemiology. *J Clin Microbiol* **35**: 907-914.

Kato-Maeda,M., Gagneux,S., Flores,L.L., Kim,E.Y., Small,P.M., Desmond,E.P., and Hopewell,P.C. (2011) Strain classification of *Mycobacterium tuberculosis*: congruence between large sequence polymorphisms and spoligotypes. *Int J Tuberc Lung Dis* **15**: 131-133.

Kato-Maeda,M., Rhee,J.T., Gingeras,T.R., Salamon,H., Drenkow,J., Smittipat,N., and Small,P.M. (2001) Comparing genomes within the species *Mycobacterium tuberculosis*. *Genome Res* **11**: 547-554.

Kaufmann,S.H. (2001) How can immunology contribute to the control of tuberculosis? *Nat Rev Immunol* **1**: 20-30.

Kaufmann,S.H. (2002) Protection against tuberculosis: cytokines, T cells, and macrophages. *Ann Rheum Dis* **61**: ii54-ii58.

Kaufmann,S.H., Hussey,G., and Lambert,P.H. (2010) New vaccines for tuberculosis. *Lancet* **375**: 2110-2119.

Kibiki,G.S., Mulder,B., Dolmans,W.M., de Beer,J.L., Boeree,M., Sam,N. *et al.* (2007) *M. tuberculosis* genotypic diversity and drug susceptibility pattern in HIV-infected and non-HIV-infected patients in northern Tanzania. *BMC Microbiol* **7**: 51.

Kik,S.V., Verver,S., van Soolingen,D., de Haas,P.E., Cobelens,F.G., Kremer,K. *et al.* (2008) Tuberculosis outbreaks predicted by characteristics of first patients in a DNA fingerprint cluster. *Am J Respir Crit Care Med* **178**: 96-104.

Kivi,M., Liu,X., Raychaudhuri,S., Altman,R.B., and Small,P.M. (2002) Determining the genomic locations of repetitive DNA sequences with a whole-genome microarray: IS*6110* in *Mycobacterium tuberculosis*. *J Clin Microbiol* **40**: 2192-2198.

Koksalan,O.K. (2005) Low positive predictive values and specificities of spoligotyping and mycobacterial interspersed repetitive unit-variable-number tandem repeat typing methods for performing population-based molecular epidemiology studies of tuberculosis. *J Clin Microbiol* **43**: 3031-3032.

Kong,Y., Cave,M.D., Zhang,L., Foxman,B., Marrs,C.F., Bates,J.H., and Yang,Z.H. (2006) Population-based study of deletions in five different genomic regions of *Mycobacterium tuberculosis* and possible clinical relevance of the deletions. *J Clin Microbiol* **44**: 3940-3946.

Kostic,T., Weilharter,A., Rubino,S., Delogu,G., Uzzau,S., Rudi,K. *et al.* (2007) A microbial diagnostic microarray technique for the sensitive detection and identification of pathogenic bacteria in a background of nonpathogens. *Anal Biochem* **360**: 244-254.

Kremer,K., Arnold,C., Cataldi,A., Gutierrez,M.C., Haas,W.H., Panaiotov,S. *et al.* (2005a) Discriminatory power and reproducibility of novel DNA typing methods for *Mycobacterium tuberculosis* complex strains. *J Clin Microbiol* **43**: 5628-5638.

Kremer,K., Au,B.K., Yip,P.C., Skuce,R., Supply,P., Kam,K.M., and van Soolingen,D. (2005b) Use of variable-number tandem-repeat typing to differentiate *Mycobacterium tuberculosis* Beijing family isolates from Hong Kong and comparison with IS*6110* restriction fragment length polymorphism typing and spoligotyping. *J Clin Microbiol* **43**: 314-320.

Kremer,K., Glynn,J.R., Lillebaek,T., Niemann,S., Kurepina,N.E., Kreiswirth,B.N. *et al.* (2004) Definition of the Beijing/W lineage of *Mycobacterium tuberculosis* on the basis of genetic markers. *J Clin Microbiol* **42**: 4040-4049.

Kremer,K., van Soolingen,D., Frothingham,R., Haas,W.H., Hermans,P.W., Martin,C. *et al.* (1999) Comparison of methods based on different molecular epidemiological markers for typing of *Mycobacterium tuberculosis* complex strains: interlaboratory study of discriminatory power and reproducibility. *J Clin Microbiol* **37**: 2607-2618.

Kruijshaar,M.E., and Abubakar,I. (2009) Increase in extrapulmonary tuberculosis in England and Wales 1999-2006. *Thorax* **64**: 1090-1095.

Kulkarni,S., Sola,C., Filliol,I., Rastogi,N., and Kadival,G. (2005) Spoligotyping of *Mycobacterium tuberculosis* isolates from patients with pulmonary tuberculosis in Mumbai, India. *Res Microbiol* **156**: 588-596.

Lalvani,A., Pathan,A.A., Durkan,H., Wilkinson,K.A., Whelan,A., Deeks,J.J. *et al.* (2001) Enhanced contact tracing and spatial tracking of *Mycobacterium tuberculosis* infection by enumeration of antigen-specific T cells. *Lancet* **357**: 2017-2021.

Lambregts-van Weezenbeek,C.S., Sebek,M.M., van Gerven,P.J., de,V.G., Verver,S., Kalisvaart,N.A., and van Soolingen,D. (2003) Tuberculosis contact investigation and DNA fingerprint surveillance in The Netherlands: 6 years' experience with nation-wide cluster feedback and cluster monitoring. *Int J Tuberc Lung Dis* **7**: S463-S470.

Lari,N., Rindi,L., Bonanni,D., Rastogi,N., Sola,C., Tortoli,E., and Garzelli,C. (2007) Three-year longitudinal study of genotypes of *Mycobacterium tuberculosis* isolates in Tuscany, Italy. *J Clin Microbiol* **45**: 1851-1857.

Lazzarini,L.C., Huard,R.C., Boechat,N.L., Gomes,H.M., Oelemann,M.C., Kurepina,N. *et al.* (2007) Discovery of a novel *Mycobacterium tuberculosis* lineage that is a major cause of tuberculosis in Rio de Janeiro, Brazil. *J Clin Microbiol* **45**: 3891-3902.

Lazzarini,L.C., Spindola,S.M., Bang,H., Gibson,A.L., Weisenberg,S., da Silva,C.W. *et al.* (2008) RDRio *Mycobacterium tuberculosis* infection is associated with a higher frequency of cavitary pulmonary disease. *J Clin Microbiol* **46**: 2175-2183.

Le Fleche,P., Fabre,M., Denoeud,F., Koeck,J.L., and Vergnaud,G. (2002) High resolution, on-line identification of strains from the *Mycobacterium tuberculosis* complex based on tandem repeat typing. *BMC Microbiol* **2:37. Epub;%2002 Nov 27.**: 37.

Leaman,A.M., Rysdale,E., and Webber,R. (2006) Use of the emergency department by Polish migrant workers. *Emerg Med J* **23**: 918-919.

Lee,J.S., Krause,R., Schreiber,J., Mollenkopf,H.J., Kowall,J., Stein,R. *et al.* (2008) Mutation in the transcriptional regulator PhoP contributes to avirulence of *Mycobacterium tuberculosis* H37Ra strain. *Cell Host Microbe* **3**: 97-103.

Levinson,G., and Gutman,G.A. (1987) Slipped-strand mispairing: a major mechanism for DNA sequence evolution. *Mol Biol Evol* **4**: 203-221.

Levy-Frebault,V.V., and Portaels,F. (1992) Proposed minimal standards for the genus Mycobacterium and for description of new slowly growing Mycobacterium species. *Int J Syst Bacteriol* **42**: 315-323.

Liebana,E., Aranaz,A., Francis,B., and Cousins,D. (1996) Assessment of genetic markers for species differentiation within the *Mycobacterium tuberculosis* complex. *J Clin Microbiol* **34**: 933-938.

Lightbody,K.L., Renshaw,P.S., Collins,M.L., Wright,R.L., Hunt,D.M., Gordon,S.V. *et al.* (2004) Characterisation of complex formation between members of the *Mycobacterium tuberculosis* complex CFP-10/ESAT-6 protein family: towards an understanding of the rules governing complex formation and thereby functional flexibility. *FEMS Microbiol Lett* **238**: 255-262.

Ling,D.I., Zwerling,A.A., and Pai,M. (2008) GenoType MTBDR assays for the diagnosis of multidrug-resistant tuberculosis: a meta-analysis. *Eur Respir J* **32**: 1165-1174.

Liu,W., Smith,D.I., Rechtzigel,K.J., Thibodeau,S.N., and James,C.D. (1998) Denaturing high performance liquid chromatography (DHPLC) used in the detection of germline and somatic mutations. *Nucleic Acids Res* **26**: 1396-1400.

Love,J., Sonnenberg,P., Glynn,J.R., Gibson,A., Gopaul,K., Fang,Z. *et al.* (2009) Molecular epidemiology of tuberculosis in England, 1998. *Int J Tuberc Lung Dis* **13**: 201-207.

Ma,X., Liu,Y., Gowen,B.B., Graviss,E.A., Clark,A.G., and Musser,J.M. (2007) Full-exon resequencing reveals toll-like receptor variants contribute to human susceptibility to tuberculosis disease. *PLoS One* **2**: e1318.

Maes,M., Kremer,K., van Sooligen,D., Takiff,H., and de Waard,J.H. (2008) 24-locus MIRU-VNTR genotyping is a useful tool to study the molecular epidemiology of tuberculosis among Warao Amerindians in Venezuela. *Tuberculosis (Edinb)* **88**: 490-494.

Maguire,H., Dale,J.W., McHugh,T.D., Butcher,P.D., Gillespie,S.H., Costetsos,A. *et al.* (2002) Molecular epidemiology of tuberculosis in London 1995-7 showing low rate of active transmission. *Thorax* **57**: 617-622.

Mahairas,G.G., Sabo,P.J., Hickey,M.J., Singh,D.C., and Stover,C.K. (1996) Molecular analysis of genetic differences between Mycobacterium bovis BCG and virulent M. bovis. *J Bacteriol* **178**: 1274-1282.

Maiden,M.C., Bygraves,J.A., Feil,E., Morelli,G., Russell,J.E., Urwin,R. *et al.* (1998) Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc Natl Acad Sci U S A* **95**: 3140-3145.

Mallard,K., Sharaf Eldin,G.S., and McNerney,R. (2009) ScreenTape as a tool for the rapid differentiation of *Mycobacterium tuberculosis* isolates. *J Med Microbiol* **58**: 1266-1268.

Marmiesse,M., Brodin,P., Buchrieser,C., Gutierrez,C., Simoes,N., Vincent,V. *et al.* (2004) Macro-array and bioinformatic analyses reveal mycobacterial 'core' genes, variation in the ESAT-6 gene family and new phylogenetic markers for the *Mycobacterium tuberculosis* complex. *Microbiology* **150**: 483-496.

Marsh,J.W., O'Leary,M.M., Shutt,K.A., Pasculle,A.W., Johnson,S., Gerding,D.N. *et al.* (2006) Multilocus variable-number tandem-repeat analysis for investigation of *Clostridium difficile* transmission in Hospitals. *J Clin Microbiol* **44**: 2558-2566.

Mazars,E., Lesjean,S., Banuls,A.L., Gilbert,M., Vincent,V., Gicquel,B. *et al.* (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.

Mazurek,G.H., Jereb,J., Lobue,P., Iademarco,M.F., Metchock,B., and Vernon,A. (2005) Guidelines for using the QuantiFERON-TB Gold test for detecting *Mycobacterium tuberculosis* infection, United States. *MMWR Recomm Rep* **54**: 49-55.

Mazurek,G.H., LoBue,P.A., Daley,C.L., Bernardo,J., Lardizabal,A.A., Bishai,W.R. *et al.* (2001) Comparison of a whole-blood interferon gamma assay with tuberculin skin testing for detecting latent *Mycobacterium tuberculosis* infection. *JAMA* **286**: 1740-1747.

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

McDonough,K.A., Kress,Y., and Bloom,B.R. (1993) Pathogenesis of tuberculosis: interaction of *Mycobacterium tuberculosis* with macrophages. *Infect Immun* **61**: 2763-2773.

McEvoy,C.R., Falmer,A.A., Gey van Pittius,N.C., Victor,T.C., van Helden,P.D., and Warren,R.M. (2007) The role of IS*6110* in the evolution of *Mycobacterium tuberculosis*. *Tuberculosis (Edinb)* **87**: 393-404.

McHugh,T.D., Batt,S.L., Shorten,R.J., Gosling,R.D., Uiso,L., and Gillespie,S.H. (2005) *Mycobacterium tuberculosis* lineage: a naming of the parts. *Tuberculosis (Edinb)* **85**: 127-136.

McHugh,T.D., Newport,L.E., and Gillespie,S.H. (1997) IS*6110* homologs are present in multiple copies in mycobacteria other than tuberculosis-causing mycobacteria. *J Clin Microbiol* **35**: 1769-1771.

McMurray,C.L., Hardy,K.J., and Hawkey,P.M. (2010) Rapid, automated epidemiological typing of methicillin-resistant Staphylococcus aureus. *J Microbiol Methods* **80**: 109-111.

McNabb,S.J., Kammerer,J.S., Hickey,A.C., Braden,C.R., Shang,N., Rosenblum,L.S., and Navin,T.R. (2004) Added epidemiologic value to tuberculosis prevention and control of the investigation of clustered genotypes of *Mycobacterium tuberculosis* isolates. *Am J Epidemiol* **160**: 589-597.

McNerney,R. (1999) TB: the return of the phage. A review of fifty years of mycobacteriophage research. *Int J Tuberc Lung Dis* **3**: 179-184.

Mestre,O., Luo,T., dos Vultos,T., Kremer,K., Murray,A., Namouchi,A. *et al.* (2011) Phylogeny of *Mycobacterium tuberculosis* Beijing strains constructed from polymorphisms in genes involved in DNA replication, recombination and repair. *PLoS One* **6**: e16020.

Miller,M.B., and Tang,Y.W. (2009) Basic concepts of microarrays and potential applications in clinical microbiology. *Clin Microbiol Rev* **22**: 611-633.

Millet,J., Baboolal,S., Akpaka,P.E., Ramoutar,D., and Rastogi,N. (2009) Phylogeographical and molecular characterization of an emerging *Mycobacterium tuberculosis* clone in Trinidad and Tobago. *Infect Genet Evol* **9**: 1336-1344.

Mokrousov,I., Narvskaya,O., Otten,T., Vyazovaya,A., Limeschenko,E., Steklova,L., and Vyshnevskyi,B. (2002) Phylogenetic reconstruction within *Mycobacterium tuberculosis* Beijing genotype in northwestern Russia. *Res Microbiol* **153**: 629-637.

Mokrousov,I., Valcheva,V., Sovhozova,N., Aldashev,A., Rastogi,N., and Isakova,J. (2009) Penitentiary population of *Mycobacterium tuberculosis* in Kyrgyzstan: exceptionally high prevalence of the Beijing genotype and its Russia-specific subtype. *Infect Genet Evol* **9**: 1400-1405.

Moore,D.A., Evans,C.A., Gilman,R.H., Caviedes,L., Coronel,J., Vivar,A. *et al.* (2006) Microscopic-observation drug-susceptibility assay for the diagnosis of TB. *N Engl J Med* **355**: 1539-1550.

Morgan,M., Kalantri,S., Flores,L., and Pai,M. (2005) A commercial line probe assay for the rapid detection of rifampicin resistance in *Mycobacterium tuberculosis*: a systematic review and meta-analysis. *BMC Infect Dis* **5**: 62.

Moser,I., Prodinger,W.M., Hotzel,H., Greenwald,R., Lyashchenko,K.P., Bakker,D. *et al.* (2008) *Mycobacterium pinnipedii*: transmission from South American sea lion (*Otaria byronia*) to Bactrian camel (*Camelus bactrianus bactrianus*) and Malayan tapirs (*Tapirus indicus*). *Vet Microbiol* **127**: 399-406.

Moss,A.R., Alland,D., Telzak,E., Hewlett,D., Jr., Sharp,V., Chiliade,P. *et al.* (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.

Mulenga,C., Shamputa,I.C., Mwakazanga,D., Kapata,N., Portaels,F., and Rigouts,L. (2010) Diversity of *Mycobacterium tuberculosis* genotypes circulating in Ndola, Zambia. *BMC Infect Dis* **10**: 177.

Murray,S.J., Barrett,A., Magee,J.G., and Freeman,R. (2003) Optimisation of acid fast smears for the direct detection of mycobacteria in clinical samples. *J Clin Pathol* **56**: 613-615.

Musser,J.M., Amin,A., and Ramaswamy,S. (2000) Negligible genetic diversity of *Mycobacterium tuberculosis* host immune system protein targets: evidence of limited selective pressure. *Genetics* **155**: 7-16.

National Institute for Clinical Excellence. Tuberculosis: Clinical diagnosis and management of tuberculosis, and measures for its prevention and control. 2011. Available from http://guidance.nice.org.uk/CG117

Newton,S.M., Smith,R.J., Wilkinson,K.A., Nicol,M.P., Garton,N.J., Staples,K.J. *et al.* (2006) A deletion defining a common Asian lineage of *Mycobacterium tuberculosis* associates with immune subversion. *Proc Natl Acad Sci U S A* **103**: 15594-15598.

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

Niemann,S., Koser,C.U., Gagneux,S., Plinke,C., Homolka,S., Bignell,H. *et al.* (2009) Genomic diversity among drug sensitive and multidrug resistant isolates of *Mycobacterium tuberculosis* with identical DNA fingerprints. *PLoS One* **4**: e7407.

Nikolayevskyy,V., Gopaul,K., Balabanova,Y., Brown,T., Fedorin,I., and Drobniewski,F. (2006) Differentiation of tuberculosis strains in a population with mainly Beijing-family strains. *Emerg Infect Dis* **12**: 1406-1413.

Oefner,P.J., and Underhill,P.A. (1995) Comparative DNA sequencing by denaturing high-performance liquid chromatography (dHPLC). *Am J Hum Genet* **57**: A266.

Oelemann,M.C., Diel,R., Vatin,V., Haas,W., Rusch-Gerdes,S., Locht,C. *et al.* (2007) Assessment of an optimized mycobacterial interspersed repetitive- unit-variable-number tandem-repeat typing system combined with spoligotyping for population-based molecular epidemiology studies of tuberculosis. *J Clin Microbiol* **45**: 691-697.

Oeltmann,J.E., Oren,E., Haddad,M.B., Lake,L., Harrington,T.A., Ijaz,K., and Narita,M. (2006) Tuberculosis outbreak in marijuana users, Seattle, Washington, 2004. *Emerg Infect Dis* **12**: 1156-1159.

Office for National Statistics. Census 2001 - Ethnicity and religion in England and Wales. 2003. Available from http://www.statistics.gov.uk/census2001/profiles/commentaries/ethnicity.asp

Office for National Statistics. West Midlands: Population and Migration. 2010. Available from http://www.statistics.gov.uk/cci/nugget.asp?id=2233

Ojo,O.O., Sheehan,S., Corcoran,D.G., Nikolayevsky,V., Brown,T., O'Sullivan,M. *et al.* (2010) Molecular epidemiology of *Mycobacterium tuberculosis* clinical isolates in Southwest Ireland. *Infect Genet Evol* **10**: 1110-1116.

Olsen,I., Balasingham,S.V., Davidsen,T., Debebe,E., Rodland,E.A., van,S.D. *et al.* (2009) Characterization of the major formamidopyrimidine-DNA glycosylase homolog in *Mycobacterium tuberculosis* and its linkage to variable tandem repeats. *FEMS Immunol Med Microbiol* **56**: 151-161.

Orita,M., Iwahana,H., Kanazawa,H., Hayashi,K., and Sekiya,T. (1989) Detection of polymorphisms of human DNA by gel electrophoresis as single-strand conformation polymorphisms. *Proc Natl Acad Sci U S A* **86**: 2766-2770.

Otlu,B., Durmaz,R., Gunal,S., Sola,C., Zozio,T., and Rastogi,N. (2009) Beijing/W and major spoligotype families of *Mycobacterium tuberculosis* strains isolated from tuberculosis patients in Eastern Turkey. *New Microbiol* **32**: 255-263.

Pallen,M.J., Loman,N.J., and Penn,C.W. (2010) High-throughput sequencing and clinical microbiology: progress, opportunities and challenges. *Curr Opin Microbiol* **13**: 625-631.

Pareek,M., Watson,J.P., Ormerod,L.P., Kon,O.M., Woltmann,G., White,P.J. *et al.* (2011) Screening of immigrants in the UK for imported latent tuberculosis: a multicentre cohort study and cost-effectiveness analysis. *Lancet Infect Dis* **11**: 435-444.

Park,H., Jang,H., Song,E., Chang,C.L., Lee,M., Jeong,S. *et al.* (2005) Detection and genotyping of Mycobacterium species from clinical isolates and specimens by oligonucleotide array. *J Clin Microbiol* **43**: 1782-1788.

Park,H., Song,E.J., Song,E.S., Lee,E.Y., Kim,C.M., Jeong,S.H. *et al.* (2006) Comparison of a conventional antimicrobial susceptibility assay to an oligonucleotide chip system for detection of drug resistance in *Mycobacterium tuberculosis* isolates. *J Clin Microbiol* **44**: 1619-1624.

Parwati,I., van,C.R., and van,S.D. (2010) Possible underlying mechanisms for successful emergence of the *Mycobacterium tuberculosis* Beijing genotype strains. *Lancet Infect Dis* **10**: 103-111.

Perez-Perez,F.J., and Hanson,N.D. (2002) Detection of plasmid-mediated *AmpC* beta-lactamase genes in clinical isolates by using multiplex PCR. *J Clin Microbiol* **40**: 2153-2162.

Perlman,D.C., Perkins,M.P., Paone,D., Kochems,L., Salomon,N., Friedmann,P., and Des Jarlais,D.C. (1997) "Shotgunning" as an illicit drug smoking practice. *J Subst Abuse Treat* **14**: 3-9.

Pfyffer,G.E., Welscher,H.M., Kissling,P., Cieslak,C., Casal,M.J., Gutierrez,J., and Rusch-Gerdes,S. (1997) Comparison of the Mycobacteria Growth Indicator Tube (MGIT) with radiometric and solid culture for recovery of acid-fast bacilli. *J Clin Microbiol* **35**: 364-368.

Platt,S. VNTR Diversity and Confidence Extractor (V-DICE). 2011. 7-1-2011. Available from http://www.hpa-bioinformatics.org.uk/cgi-bin/DICI/DICI.pl

Pottenger,F.M. (1948) Public health significance of rare tubercle bacilli in sputum. *Am Rev Tuberc* **58**: 314-321.

Prodinger,W.M., Brandstatter,A., Naumann,L., Pacciarini,M., Kubica,T., Boschiroli,M.L. *et al.* (2005) Characterization of *Mycobacterium caprae* isolates from Europe by mycobacterial interspersed repetitive unit genotyping. *J Clin Microbiol* **43**: 4984-4992.

Pym,A.S., Brodin,P., Brosch,R., Huerre,M., and Cole,S.T. (2002) Loss of RD1 contributed to the attenuation of the live tuberculosis vaccines Mycobacterium bovis BCG and *Mycobacterium microti*. *Mol Microbiol* **46**: 709-717.

Pym,A.S., and Brosch,R. (2000) Tools for the population genomics of the tubercle bacilli. *Genome Res* **10**: 1837-1839.

Radhakrishnan,I., MY,K., Kumar,R.A., and Mundayoor,S. (2001) Implications of low frequency of IS*6110* in fingerprinting field isolates of *Mycobacterium tuberculosis* from Kerala, India. *J Clin Microbiol* **39**: 1683.

Rado,T.A., Bates,J.H., Engel,H.W., Mankiewicz,E., Murohashi,T., Mizuguchi,Y., and Sula,L. (1975) World Health Organization studies on bacteriophage typing of mycobacteria. Subdivision of the species *Mycobacterium tuberculosis*. *Am Rev Respir Dis* **111**: 459-468.

Rahim,Z., Zaman,K., van der Zanden,A.G., Mollers,M.J., van,S.D., Raqib,R. *et al.* (2007) Assessment of population structure and major circulating phylogeographical clades of *Mycobacterium tuberculosis* complex in Bangladesh suggests a high prevalence of a specific subclade of ancient *M. tuberculosis* genotypes. *J Clin Microbiol* **45**: 3791-3794.

Rajakumar,K., Shafi,J., Smith,R.J., Stabler,R.A., Andrew,P.W., Modha,D. *et al.* (2004) Use of genome level-informed PCR as a new investigational approach for analysis of outbreak-associated *Mycobacterium tuberculosis* isolates. *J Clin Microbiol* **42**: 1890-1896.

Raynaud,C., Guilhot,C., Rauzier,J., Bordat,Y., Pelicic,V., Manganelli,R. *et al.* (2002) Phospholipases C are involved in the virulence of *Mycobacterium tuberculosis*. *Mol Microbiol* **45**: 203-217.

Reed,M.B., Pichler,V.K., McIntosh,F., Mattia,A., Fallow,A., Masala,S. *et al.* (2009) Major *Mycobacterium tuberculosis* lineages associate with patient country of origin. *J Clin Microbiol* **47**: 1119-1128.

Reich,D., Thangaraj,K., Patterson,N., Price,A.L., and Singh,L. (2009) Reconstructing Indian population history. *Nature* **461**: 489-494.

Rengarajan,J., Bloom,B.R., and Rubin,E.J. (2005) Genome-wide requirements for *Mycobacterium tuberculosis* adaptation and survival in macrophages. *Proc Natl Acad Sci U S A* **102**: 8327-8332.

Richter,E., Weizenegger,M., Fahr,A.M., and Rusch-Gerdes,S. (2004) Usefulness of the GenoType MTBC assay for differentiating species of the *Mycobacterium tuberculosis* complex in cultures obtained from clinical specimens. *J Clin Microbiol* **42**: 4303-4306.

Richter,E., Weizenegger,M., Rusch-Gerdes,S., and Niemann,S. (2003) Evaluation of genotype MTBC assay for differentiation of clinical *Mycobacterium tuberculosis* complex isolates. *J Clin Microbiol* **41**: 2672-2675.

Riley,R.L. (1957) Aerial dissemination of pulmonary tuberculosis. *Am Rev Tuberc* **76**: 931-941.

Rosales,S., Pineda-Garcia,L., Ghebremichael,S., Rastogi,N., and Hoffner,S.E. (2010) Molecular diversity of *Mycobacterium tuberculosis* isolates from patients with tuberculosis in Honduras. *BMC Microbiol* **10**: 208.

Rovina,N., Karabela,S., Constantoulakis,P., Michou,V., Konstantinou,K., Sgountzos,V. *et al.* (2011) MIRU-VNTR typing of drug-resistant tuberculosis isolates in Greece. *Ther Adv Respir Dis* **5**: 229-236.

Rozen,S., and Skaletsky,H. (2000) Primer3 on the WWW for general users and for biologist programmers. *Methods Mol Biol* **132:365-86.**: 365-386.

Ruddy,M.C., Davies,A.P., Yates,M.D., Yates,S., Balasegaram,S., Drabu,Y. *et al.* (2004) Outbreak of isoniazid resistant tuberculosis in north London. *Thorax* **59**: 279-285.

Safi,H., Barnes,P.F., Lakey,D.L., Shams,H., Samten,B., Vankayalapati,R., and Howard,S.T. (2004) IS*6110* functions as a mobile, monocyte-activated promoter in *Mycobacterium tuberculosis*. *Mol Microbiol* **52**: 999-1012.

Sampson,S.L., Lukey,P., Warren,R.M., van Helden,P.D., Richardson,M., and Everett,M.J. (2001) Expression, characterization and subcellular localization of the *Mycobacterium tuberculosis* PPE gene Rv1917c. *Tuberculosis (Edinb)* **81**: 305-317.

Sassetti,C.M., Boyd,D.H., and Rubin,E.J. (2003) Genes required for mycobacterial growth defined by high density mutagenesis. *Mol Microbiol* **48**: 77-84.

Savine,E., Warren,R.M., van der Spuy,G.D., Beyers,N., van Helden,P.D., Locht,C., and Supply,P. (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.

Schurch,A.C., Kremer,K., Daviena,O., Kiers,A., Boeree,M.J., Siezen,R.J., and van,S.D. (2010a) High-resolution typing by integration of genome sequencing data in a large tuberculosis cluster. *J Clin Microbiol* **48**: 3403-3406.

Schurch,A.C., Kremer,K., Kiers,A., Daviena,O., Boeree,M.J., Siezen,R.J. *et al.* (2010b) The tempo and mode of molecular evolution of *Mycobacterium tuberculosis* at patient-to-patient scale. *Infect Genet Evol* **10**: 108-114.

Schurch,A.C., Kremer,K., Warren,R.M., Hung,N.V., Zhao,Y., Wan,K. *et al.* (2011) Mutations in the regulatory network underlie the recent clonal expansion of a dominant subclone of the *Mycobacterium tuberculosis* Beijing genotype. *Infect Genet Evol* **11**: 587-597.

Scott,A.N., Menzies,D., Tannenbaum,T.N., Thibert,L., Kozak,R., Joseph,L. *et al.* (2005) Sensitivities and specificities of spoligotyping and mycobacterial interspersed repetitive unit-variable-number tandem repeat typing methods for studying molecular epidemiology of tuberculosis. *J Clin Microbiol* **43**: 89-94.

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

Seki,M., Honda,I., Fujita,I., Yano,I., Yamamoto,S., and Koyama,A. (2009) Whole genome sequence analysis of Mycobacterium bovis bacillus Calmette-Guerin (BCG) Tokyo 172: a comparative study of BCG vaccine substrains. *Vaccine* **27**: 1710-1716.

Semaw,S., Simpson,S.W., Quade,J., Renne,P.R., Butler,R.F., McIntosh,W.C. *et al.* (2005) Early Pliocene hominids from Gona, Ethiopia. *Nature* **433**: 301-305.

Shamputa,I.C., Jugheli,L., Sadradze,N., Willery,E., Portaels,F., Supply,P., and Rigouts,L. (2006) Mixed infection and clonal representativeness of a single sputum sample in tuberculosis patients from a penitentiary hospital in Georgia. *Respir Res* **7:99.**: 99.

Shamputa,I.C., Rigouts,L., Eyongeta,L.A., El Aila,N.A., Van,D.A., Salim,A.H. *et al.* (2004) Genotypic and phenotypic heterogeneity among *Mycobacterium tuberculosis* isolates from pulmonary tuberculosis patients. *J Clin Microbiol* **42**: 5528-5536.

Sharaf-Eldin,G.S., Saeed,N.S., Hamid,M.E., Jordaan,A.M., van der Spuy,G.D., Warren,R.M. *et al.* (2002) Molecular analysis of clinical isolates of *Mycobacterium tuberculosis* collected from patients with persistent disease in the Khartoum region of Sudan. *J Infect* **44**: 244-251.

Sharma,S.K., Mohan,A., Sharma,A., and Mitra,D.K. (2005) Miliary tuberculosis: new insights into an old disease. *Lancet Infect Dis* **5**: 415-430.

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

Shlush,L.I., Behar,D.M., Zelazny,A., Keller,N., Lupski,J.R., Beaudet,A.L., and Bercovich,D. (2002) Molecular epidemiological analysis of the changing nature of a meningococcal outbreak following a vaccination campaign. *J Clin Microbiol* **40**: 3565-3571.

Singh,U.B., Arora,J., Suresh,N., Pant,H., Rana,T., Sola,C. *et al.* (2007) Genetic biodiversity of *Mycobacterium tuberculosis* isolates from patients with pulmonary tuberculosis in India. *Infect Genet Evol* **7**: 441-448.

Singh,U.B., Suresh,N., Bhanu,N.V., Arora,J., Pant,H., Sinha,S. *et al.* (2004) Predominant tuberculosis spoligotypes, Delhi, India. *Emerg Infect Dis* **10**: 1138-1142.

Skeiky,Y.A., Ovendale,P.J., Jen,S., Alderson,M.R., Dillon,D.C., Smith,S. *et al.* (2000) T cell expression cloning of a *Mycobacterium tuberculosis* gene encoding a protective antigen associated with the early control of infection. *J Immunol* **165**: 7140-7149.

Skjot,R.L., Brock,I., Arend,S.M., Munk,M.E., Theisen,M., Ottenhoff,T.H., and Andersen,P. (2002) Epitope mapping of the immunodominant antigen TB10.4 and the two homologous proteins TB10.3 and TB12.9, which constitute a subfamily of the ESAT-6 gene family. *Infect Immun* **70**: 5446-5453.

Skuce,R.A., McCorry,T.P., McCarroll,J.F., Roring,S.M., Scott,A.N., Brittain,D. *et al.* (2002) Discrimination of *Mycobacterium tuberculosis* complex bacteria using novel VNTR-PCR targets. *Microbiology* **148**: 519-528.

Small,P.M., Hopewell,P.C., Singh,S.P., Paz,A., Parsonnet,J., Ruston,D.C. *et al.* (1994) The epidemiology of tuberculosis in San Francisco. A population-based study using conventional and molecular methods. *N Engl J Med* **330**: 1703-1709.

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

Smittipat,N., Billamas,P., Palittapongarnpim,M., Thong-On,A., Temu,M.M., Thanakijcharoen,P. *et al.* (2005) Polymorphism of variable-number tandem repeats at multiple loci in *Mycobacterium tuberculosis*. *J Clin Microbiol* **43**: 5034-5043.

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

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

Sola,C., Ferdinand,S., Sechi,L.A., Zanetti,S., Martial,D., Mammina,C. *et al.* (2005) *Mycobacterium tuberculosis* molecular evolution in western Mediterranean Islands of Sicily and Sardinia. *Infect Genet Evol* **5**: 145-156.

Sola,C., Filliol,I., Gutierrez,M.C., Mokrousov,I., Vincent,V., and Rastogi,N. (2001) Spoligotype database of *Mycobacterium tuberculosis*: biogeographic distribution of shared types and epidemiologic and phylogenetic perspectives. *Emerg Infect Dis* **7**: 390-396.

Sola,C., Filliol,I., Legrand,E., Lesjean,S., Locht,C., Supply,P., and Rastogi,N. (2003) Genotyping of the *Mycobacterium tuberculosis* complex using MIRUs: association with VNTR and spoligotyping for molecular epidemiology and evolutionary genetics. *Infect Genet Evol* **3**: 125-133.

Sorek,R., Kunin,V., and Hugenholtz,P. (2008) CRISPR--a widespread system that provides acquired resistance against phages in bacteria and archaea. *Nat Rev Microbiol* **6**: 181-186.

Soto,C.Y., Menendez,M.C., Perez,E., Samper,S., Gomez,A.B., Garcia,M.J., and Martin,C. (2004) IS*6110* mediates increased transcription of the phoP virulence gene in a multidrug-resistant clinical isolate responsible for tuberculosis outbreaks. *J Clin Microbiol* **42**: 212-219.

Sreevatsan,S., Pan,X., Stockbauer,K.E., Connell,N.D., Kreiswirth,B.N., Whittam,T.S., and Musser,J.M. (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.

Stavrum,R., Mphahlele,M., Ovreas,K., Muthivhi,T., Fourie,P.B., Weyer,K., and Grewal,H.M. (2009a) High diversity of *Mycobacterium tuberculosis* genotypes in South Africa and preponderance of mixed infections among ST53 isolates. *J Clin Microbiol* **47**: 1848-1856.

Stavrum,R., Myneedu,V.P., Arora,V.K., Ahmed,N., and Grewal,H.M. (2009b) In-depth molecular characterization of *Mycobacterium tuberculosis* from New Delhi--predominance of drug resistant isolates of the 'modern' (TbD1) type. *PLoS One* **4**: e4540.

Stavrum,R., Valvatne,H., Bo,T.H., Jonassen,I., Hinds,J., Butcher,P.D., and Grewal,H.M. (2008) Genomic diversity among Beijing and non-Beijing *Mycobacterium tuberculosis* isolates from Myanmar. *PLoS One* **3**: e1973.

Steenken,W., Oatway,W.H., and Petroff,S.A. (1934) Biological studies of the tubercle bacillus: III. Dissociation and pathogenicity of the R and S variants of the human tubercle bacillus (H(37)). *J Exp Med* **60**: 515-540.

Stein,C.M. (2011) Genetic epidemiology of tuberculosis susceptibility: impact of study design. *PLoS Pathog* **7**: e1001189.

Stewart,G.R., Wernisch,L., Stabler,R., Mangan,J.A., Hinds,J., Laing,K.G. *et al.* (2002) Dissection of the heat-shock response in *Mycobacterium tuberculosis* using mutants and microarrays. *Microbiology* **148**: 3129-3138.

Stinear,T.P., Seemann,T., Harrison,P.F., Jenkin,G.A., Davies,J.K., Johnson,P.D. *et al.* (2008) Insights from the complete genome sequence of Mycobacterium marinum on the evolution of *Mycobacterium tuberculosis*. *Genome Res* **18**: 729-741.

Stragier,P., Ablordey,A., Meyers,W.M., and Portaels,F. (2005) Genotyping Mycobacterium ulcerans and Mycobacterium marinum by using mycobacterial interspersed repetitive units. *J Bacteriol* **187**: 1639-1647.

Styblo,K. (1980) Recent advances in epidemiological research in tuberculosis. *Adv Tuberc Res* **20**: 1-63.

Supply,P., Allix,C., Lesjean,S., Cardoso-Oelemann,M., Rusch-Gerdes,S., Willery,E. *et al.* (2006) Proposal for standardization of optimized mycobacterial interspersed repetitive unit-variable-number tandem repeat typing of *Mycobacterium tuberculosis*. *J Clin Microbiol* **44**: 4498-4510.

Supply,P., Lesjean,S., Savine,E., Kremer,K., van,S.D., and Locht,C. (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.

Supply,P., Magdalena,J., Himpens,S., and Locht,C. (1997) Identification of novel intergenic repetitive units in a mycobacterial two-component system operon. *Mol Microbiol* **26**: 991-1003.

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

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

Tanveer,M., Hasan,Z., Siddiqui,A.R., Ali,A., Kanji,A., Ghebremicheal,S., and Hasan,R. (2008) Genotyping and drug resistance patterns of *M. tuberculosis* strains in Pakistan. *BMC Infect Dis* **8**: 171.

Tekaia,F., Gordon,S.V., Garnier,T., Brosch,R., Barrell,B.G., and Cole,S.T. (1999) Analysis of the proteome of *Mycobacterium tuberculosis* in silico. *Tuber Lung Dis* **79**: 329-342.

Telenti,A., Honore,N., Bernasconi,C., March,J., Ortega,A., Heym,B. *et al.* (1997) Genotypic assessment of isoniazid and rifampin resistance in *Mycobacterium tuberculosis*: a blind study at reference laboratory level. *J Clin Microbiol* **35**: 719-723.

Thierry,D., Brisson-Noel,A., Vincent-Levy-Frebault,V., Nguyen,S., Guesdon,J.L., and Gicquel,B. (1990a) Characterization of a *Mycobacterium tuberculosis* insertion sequence, IS*6110*, and its application in diagnosis. *J Clin Microbiol* **28**: 2668-2673.

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

Thorne,N., Evans,J.T., Smith,E.G., Hawkey,P.M., Gharbia,S., and Arnold,C. (2007a) An IS*6110*-targeting fluorescent amplified fragment length polymorphism alternative to IS*6110* restriction fragment length polymorphism analysis for *Mycobacterium tuberculosis* DNA fingerprinting. *Clin Microbiol Infect* **13**: 964-970.

Thorne,N., Evans,J.T., Smith,E.G., Hawkey,P.M., Gharbia,S., and Arnold,C. (2007b) An IS*6110*-targeting fluorescent amplified fragment length polymorphism alternative to IS*6110* restriction fragment length polymorphism analysis for *Mycobacterium tuberculosis* DNA fingerprinting. *Clin Microbiol Infect* **13**: 964-970.

Tortoli,E., Benedetti,M., Fontanelli,A., and Simonetti,M.T. (2002) Evaluation of automated BACTEC MGIT 960 system for testing susceptibility of *Mycobacterium tuberculosis* to four major antituberculous drugs: comparison with the radiometric BACTEC 460TB method and the agar plate method of proportion. *J Clin Microbiol* **40**: 607-610.

Tortoli,E., Mariottini,A., and Mazzarelli,G. (2003) Evaluation of INNO-LiPA MYCOBACTERIA v2: improved reverse hybridization multiple DNA probe assay for mycobacterial identification. *J Clin Microbiol* **41**: 4418-4420.

Tsolaki,A.G., Gagneux,S., Pym,A.S., Goguet de la Salmoniere YO, Kreiswirth,B.N., van,S.D., and Small,P.M. (2005) Genomic deletions classify the Beijing/W strains as a distinct genetic lineage of *Mycobacterium tuberculosis*. *J Clin Microbiol* **43**: 3185-3191.

Tsolaki,A.G., Hirsh,A.E., DeRiemer,K., Enciso,J.A., Wong,M.Z., Hannan,M. *et al.* (2004) Functional and evolutionary genomics of *Mycobacterium tuberculosis*: insights from genomic deletions in 100 strains. *Proc Natl Acad Sci U S A* **101**: 4865-4870.

Ulukanligil,M., Aslan,G., and Tasci,S. (2000) A comparative study on the different staining methods and number of specimens for the detection of acid fast bacilli. *Mem Inst Oswaldo Cruz* **95**: 855-858.

Underhill,P.A., Jin,L., Lin,A.A., Mehdi,S.Q., Jenkins,T., Vollrath,D. *et al.* (1997) Detection of numerous Y chromosome biallelic polymorphisms by denaturing high-performance liquid chromatography. *Genome Res* **7**: 996-1005.

United Nations Statistics Division. (2009) United Nations Standard Country and Area Codes Classification Scheme. Available from http://unstats.un.org/unsd/methods/m49/m49regin.htm

Valcheva,V., Mokrousov,I., Narvskaya,O., Rastogi,N., and Markova,N. (2008) Utility of new 24-locus variable-number tandem-repeat typing for discriminating *Mycobacterium tuberculosis* clinical isolates collected in Bulgaria. *J Clin Microbiol* **46**: 3005-3011.

van Belkum,A. (1999) Short sequence repeats in microbial pathogenesis and evolution. *Cell Mol Life Sci* **56**: 729-734.

van Embden,J.D., Cave,M.D., Crawford,J.T., Dale,J.W., Eisenach,K.D., Gicquel,B. *et al.* (1993) Strain identification of *Mycobacterium tuberculosis* by DNA fingerprinting: recommendations for a standardized methodology. *J Clin Microbiol* **31**: 406-409.

van Soolingen,D. (2001) Molecular epidemiology of tuberculosis and other mycobacterial infections: main methodologies and achievements. *J Intern Med* **249**: 1-26.

van Soolingen,D., Borgdorff,M.W., de Haas,P.E., Sebek,M.M., Veen,J., Dessens,M. *et al.* (1999) Molecular epidemiology of tuberculosis in the Netherlands: a nationwide study from 1993 through 1997. *J Infect Dis* **180**: 726-736.

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

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

van Soolingen,D., Borgdorff,M.W., de Haas,P.E., Sebek,M.M., Veen,J., Dessens,M. *et al.* (1999) Molecular epidemiology of tuberculosis in the Netherlands: a nationwide study from 1993 through 1997. *J Infect Dis* **180**: 726-736.

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

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

van Soolingen,D., Hoogenboezem,T., de Haas,P.E., Hermans,P.W., Koedam,M.A., Teppema,K.S. *et al.* (1997) A novel pathogenic taxon of the *Mycobacterium tuberculosis* complex, Canetti: characterization of an exceptional isolate from Africa. *Int J Syst Bacteriol* **47**: 1236-1245.

van Soolingen,D., Qian,L., de Haas,P.E., Douglas,J.T., Traore,H., Portaels,F. *et al.* (1995) Predominance of a single genotype of *Mycobacterium tuberculosis* in countries of east Asia. *J Clin Microbiol* **33**: 3234-3238.

Vera-Cabrera,L., Hernandez-Vera,M.A., Welsh,O., Johnson,W.M., and Castro-Garza,J. (2001) Phospholipase region of *Mycobacterium tuberculosis* is a preferential locus for IS*6110* transposition. *J Clin Microbiol* **39**: 3499-3504.

Viegas,S.O., Machado,A., Groenheit,R., Ghebremichael,S., Pennhag,A., Gudo,P.S. *et al.* (2010) Molecular diversity of *Mycobacterium tuberculosis* isolates from patients with pulmonary tuberculosis in Mozambique. *BMC Microbiol* **10:195.**: 195.

Voskuil,M.I., Schnappinger,D., Visconti,K.C., Harrell,M.I., Dolganov,G.M., Sherman,D.R., and Schoolnik,G.K. (2003) Inhibition of respiration by nitric oxide induces a *Mycobacterium tuberculosis* dormancy program. *J Exp Med* **198**: 705-713.

Vynnycky,E., Nagelkerke,N., Borgdorff,M.W., van,S.D., van Embden,J.D., and Fine,P.E. (2001) The effect of age and study duration on the relationship between 'clustering' of DNA fingerprint patterns and the proportion of tuberculosis disease attributable to recent transmission. *Epidemiol Infect* **126**: 43-62.

Wada,T., Maeda,S., Hase,A., and Kobayashi,K. (2007) Evaluation of variable numbers of tandem repeat as molecular epidemiological markers of *Mycobacterium tuberculosis* in Japan. *J Med Microbiol* **56**: 1052-1057.

Wade,M.M., Volokhov,D., Peredelchuk,M., Chizhikov,V., and Zhang,Y. (2004) Accurate mapping of mutations of pyrazinamide-resistant *Mycobacterium tuberculosis* strains with a scanning-frame oligonucleotide microarray. *Diagn Microbiol Infect Dis* **49**: 89-97.

Warren,R.M., Victor,T.C., Streicher,E.M., Richardson,M., van der Spuy,G.D., Johnson,R. *et al.* (2004) Clonal expansion of a globally disseminated lineage of *Mycobacterium tuberculosis* with low IS*6110* copy numbers. *J Clin Microbiol* **42**: 5774-5782.

Webber,R. (2007) Using names to segment customers by cultural, ethnic or religious origin. *J Direct, Data and Digital Marketing Practice* **8**: 226-242.

Wilson,M., DeRisi,J., Kristensen,H.H., Imboden,P., Rane,S., Brown,P.O., and Schoolnik,G.K. (1999) Exploring drug-induced alterations in gene expression in *Mycobacterium tuberculosis* by microarray hybridization. *Proc Natl Acad Sci U S A* **96**: 12833-12838.

Wilson,S.M., al-Suwaidi,Z., McNerney,R., Porter,J., and Drobniewski,F. (1997) Evaluation of a new rapid bacteriophage-based method for the drug susceptibility testing of *Mycobacterium tuberculosis*. *Nat Med* **3**: 465-468.

Wilson,S.M., Goss,S., and Drobniewski,F. (1998) Evaluation of strategies for molecular fingerprinting for use in the routine work of a Mycobacterium reference unit. *J Clin Microbiol* **36**: 3385-3388.

Wirth,T., Hildebrand,F., Allix-Beguec,C., Wolbeling,F., Kubica,T., Kremer,K. *et al.* (2008) Origin, spread and demography of the *Mycobacterium tuberculosis* complex. *PLoS Pathog* **4**: e1000160.

Wong,S.H., Gochhait,S., Malhotra,D., Pettersson,F.H., Teo,Y.Y., Khor,C.C. *et al.* (2010) Leprosy and the adaptation of human toll-like receptor 1. *PLoS Pathog* **6**: e1000979.

World Health Organisation. Global Tuberculosis Control: A short update to the 2009 report. 2009. Avilable from http://www.who.int/tb/publications/global_report/2009/update/en/index.html

World Health Organisation. Multidrug and extensively drug-resistant TB (M/XDR-TB): 2010 global report on surveillance and response. 2010a. Available from http://whqlibdoc.who.int/publications/2010/9789241599191_eng.pdf

World Health Organisation. Tuberculosis Fact Sheet No. 104. 2010b. 4-2-2011b. Available from http://www.who.int/mediacentre/factsheets/fs104/en/

Wright,A., Zignol,M., Van,D.A., Falzon,D., Gerdes,S.R., Feldman,K. *et al.* (2009) Epidemiology of antituberculosis drug resistance 2002-07: an updated analysis of the Global Project on Anti-Tuberculosis Drug Resistance Surveillance. *Lancet* **373**: 1861-1873.

Yaganehdoost,A., Graviss,E.A., Ross,M.W., Adams,G.J., Ramaswamy,S., Wanger,A. *et al.* (1999) Complex transmission dynamics of clonally related virulent *Mycobacterium tuberculosis* associated with barhopping by predominantly human immunodeficiency virus-positive gay men. *J Infect Dis* **180**: 1245-1251.

Yang,Z.H., de Haas,P.E., van,S.D., van Embden,J.D., and Andersen,A.B. (1994) Restriction fragment length polymorphism *Mycobacterium tuberculosis* strains isolated from Greenland during 1992: evidence of tuberculosis transmission between Greenland and Denmark. *J Clin Microbiol* **32**: 3018-3025.

Yeager,H., Jr., Lacy,J., Smith,L.R., and LeMaistre,C.A. (1967) Quantitative studies of mycobacterial populations in sputum and saliva. *Am Rev Respir Dis* **95**: 998-1004.

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

Yesilkaya,H., Forbes,K.J., Shafi,J., Smith,R., Dale,J.W., Rajakumar,K. *et al.* (2006) The genetic portrait of an outbreak strain. *Tuberculosis (Edinb)* **86**: 357-362.

Yindeeyoungyeon,W., Likitvivatanavong,S., and Palittapongarnpim,P. (2009) Characterization of alpha-isopropylmalate synthases containing different copy numbers of tandem repeats in *Mycobacterium tuberculosis*. *BMC Microbiol* **9**: 122.

Zainuddin,Z.F., and Dale,J.W. (1989) Polymorphic repetitive DNA sequences in *Mycobacterium tuberculosis* detected with a gene probe from a Mycobacterium fortuitum plasmid. *J Gen Microbiol* **135**: 2347-2355.

Zainuddin,Z.F., and Dale,J.W. (1990) Does *Mycobacterium tuberculosis* have plasmids? *Tubercle* **71**: 43-49.

Zaluckyj, S. Mercian: the Anglo-Saxon kingdom of central England. 2001. Woonton, Herefordshire, Logaston Press.

Zhang,J., Abadia,E., Refregier,G., Tafaj,S., Boschiroli,M.L., Guillard,B. *et al.* (2010) *Mycobacterium tuberculosis* complex CRISPR genotyping: improving efficiency, throughput and discriminative power of 'spoligotyping' with new spacers and a microbead-based hybridization assay. *J Med Microbiol* **59**: 285-294.

Zhang,Y., and Young,D. (1994) Strain variation in the *katG* region of *Mycobacterium tuberculosis*. *Mol Microbiol* **14**: 301-308.

Zheng,H., Lu,L., Wang,B., Pu,S., Zhang,X., Zhu,G. *et al.* (2008) Genetic basis of virulence attenuation revealed by comparative genomic analysis of *Mycobacterium tuberculosis* strain H37Ra versus H37Rv. *PLoS One* **3**: e2375.