IS6110 FAFLP PCR, a tool for genomic mapping enabling investigation of evolutionary relationships of *Mycobacterium tuberculosis* in resource poor settings

Kartykayan Moganeradj

Doctor of Philosophy

UCL
Declaration

I, Kartykayan Moganeradj, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Signed:
Abstract

Tuberculosis (TB) is an important communicable disease affecting the human population world-wide. Despite the efforts of the scientific community, national governments and WHO in controlling the disease, TB still remains a major killer in resource poor settings. New rapid assays and techniques that are simple and cost-effective are urgently needed to identify, treat and understand pathogenesis including the geographical distribution of the disease. The aim of the thesis is to develop a novel genomic mapping tool using Insertion Element, IS6110 that could aid in epidemiological studies of *Mycobacterium tuberculosis* complex (MTBC) in low and middle income countries.

IS6110, a bacterial transposon, plays an essential role in changing the physical and biochemical traits of MTBC. Due to their transposition in TB genomes, they are used as epidemiological markers for differentiation of TB organisms and the mapping of these elements could also shed light on the putative altered function of adjacent genes. In the era of Whole Genome Sequencing (WGS) where repeat elements are difficult to sequence with short read technologies, a rapid and simple method of insertion site mapping using IS6110 FAFLP PCR was developed. This work is aimed at developing a rapid, cost-effective and robust genomic tool box exploiting the IS6110 FAFLP PCR assay that can both identify and characterise the TB genotypes / genetic lineages in any geographical location.

For the first time using the assay above, TB samples from Nepal were categorised into different genetic lineages. Fifty-five percent of the samples
analysed belong to Principal Genetic Group 1 (PGG1), Beijing and Central Asian strains. Also, new primers were designed targeting the Beijing and the T- groups using the FAFLP derived data that gave rise to the development of rapid lineage specific PCR assays. In addition, it was noticed that 3.9% of the Nepalese strains tested in this research work were likely multi-drug resistant (MDR-TB) using PCR targeting the Rifampicin-resistance-determining region (RRDR) of the rpoB region.

It is demonstrated here that IS6110 FAFLP methodology could easily characterise the TB samples into different genetic lineages provided they have more than four IS6110 copies. In addition, lineage specific PCR does not need any expensive instruments or reagents except for PCR blocks and gel visualisers, and could be very effective in the rapid identification of different TB genotypes within hours. These data also add to knowledge about the circulating strains of TB in Nepal, currently a poorly characterised region of the world in this regard, and could help in contact tracing studies by epidemiologists. The IS6110 FAFLP technique thus can be employed in any geographical location to map TB genetic lineages where there is little or no information available on the prevailing TB strains.
# Table of Contents

Declaration .................................................................................................................. i
Abstract ....................................................................................................................... ii
Table of Contents ........................................................................................................ iv
List of Tables ................................................................................................................ ix
List of figures ............................................................................................................... xi
Abbreviations .............................................................................................................. xv
Glossary of Key terms ................................................................................................. xviii
Outline of the thesis ..................................................................................................... xix
My Role in the thesis .................................................................................................... xxi
Publications and conference presentations ............................................................... xxii
Acknowledgements .................................................................................................... xxiv

## Chapter 1 Introduction.

1.1 *Mycobacterium tuberculosis* complex (MTBC) .............................................. 2
1.2 TB Timeline .......................................................................................................... 3
1.3 Morphology of *Mycobacterium tuberculosis* .................................................... 6
1.4 Immunopathogenesis of Tuberculosis ................................................................. 7
1.5 Clinical signs and symptoms .............................................................................. 9
1.6 Diagnosis of active tuberculosis ......................................................................... 12
  1.6.1 Chest radiographs ........................................................................................... 13
  1.6.2 Tuberculin Skin test (TST) or Mantoux test ................................................. 13
  1.6.3 Interferon- Gamma release assay (IGRA) ..................................................... 14
  1.6.4 Sputum microscopy ....................................................................................... 14
  1.6.5 Sputum culture .............................................................................................. 15
  1.6.6 Molecular identification test and drug susceptibility tests ...................... 17
  1.6.7 Diagnosis of Latent tuberculosis ................................................................. 18
1.7 TB treatment ......................................................................................................... 20
1.7.1 Rifampicin (RIF) ..............................................................................................................21
1.7.2 Isoniazid (INH) ...........................................................................................................21
1.7.3 Pyrazinamide (PZA) .....................................................................................................22
1.7.4 Ethambutol (EMB) .......................................................................................................22
1.7.5 Second line drugs ..........................................................................................................22
1.7.6 Vaccination ..................................................................................................................23

1.8 Molecular markers in TB epidemiology ..........................................................................23
1.8.1 IS6110 and IS6110 RFLP ..........................................................................................24
1.8.2 PGRS- RFLP ................................................................................................................31
1.8.3 Spoligotyping ...............................................................................................................32
1.8.4 MIRU-VNTR ...............................................................................................................36
1.8.5 AFLP and FAFLP ..........................................................................................................38

1.9 Global TB lineages .........................................................................................................43
1.9.1 Major Genetic Groups of MTBC and its global distribution ........................................46
1.9.2 Global distribution of TB genetic lineages including Co-Evolution of TB and Humans ...........................................................................................................................................51
1.9.3 Molecular Clocks in the evolution of M.tuberculosis ..................................................53
1.9.4 Importance of genetic polymorphism ...........................................................................55

1.10 Molecular Epidemiology of TB in resource poor settings ..........................................55
1.11 Currently used epidemiological tools, its suitability and constraints ..........................55

1.12 Nepal TB epidemiology ................................................................................................56
1.12.1 TB in Nepal .................................................................................................................56
1.12.2 Current diagnostic procedures in Nepal ......................................................................58
1.12.3 Constraints in diagnosis of TB in Nepal .....................................................................59

1.13 Aims and Objectives ......................................................................................................61
1.14 Hypothesis .....................................................................................................................61

Chapter 2 General Materials and Methods .......................................................................63
2.1 Bacterial samples ................................................................. 64
  2.1.1 Sample collection centres .............................................. 64
  2.1.2 Ethics .......................................................................... 65
  2.1.3 IS6110 Fluorescent Amplified Fragment Length Polymorphism (FAFLP) ................................................................. 65
  2.1.4 Genomic DNA Restriction Endonuclease digestion .......... 67
  2.1.5 Ligation of digested DNA to double stranded adapters ...... 68
  2.1.6 PCR amplification of fragments ........................................ 68
  2.1.7 Capillary Electrophoresis of fragments using ABI3730xl Genetic Analyser ................................................................. 70

Chapter 3 Optimisation and Development of Insertion Element IS6110 Fluorescent Amplified Fragment Length Polymorphism (FAFLP) PCR ....... 73
  3.1 Introduction ........................................................................ 74
    3.1.1 Aims ........................................................................... 74
    3.1.2 Objectives ..................................................................... 75
    3.1.3 IS6110 FAFLP Methodology .......................................... 75
  3.2 Materials and Methods .......................................................... 76
  3.3 Results .............................................................................. 80
    3.3.1 Amplification of DNA fragments ...................................... 82
    3.3.2 Validation of the PCR reagents ....................................... 82
    3.3.3 H37Rv and Nepal TB DNA study samples ..................... 89
  3.4 Discussion .......................................................................... 90
  3.5 Summary ............................................................................ 91

Chapter 4 Mapping of Insertion sites (IS6110) in the M. tuberculosis H37Rv reference genome and rapid definition of genetic lineages .................. 92
  4.1 Introduction ........................................................................ 93
    4.1.1 Aims and Objectives ..................................................... 95
  4.2 Materials and Methods .......................................................... 96
4.2.1 Rapid definition of TB genetic lineages-PGG3 (H37Rv) and PGG1 (Beijing) .......................................................... 97

4.3 Results ................................................................................................................................. 99

4.3.1 Prediction of fragments using Seqbuilder (DNASTAR Lasergene version 8, USA) software (in silico analysis) .................................................. 99

4.3.2 Identification of fragments using Peak Scanner (Thermofisher Scientific, UK) software (in vitro analysis) .......................................................... 99

4.3.3 Mapping the insertion site IS6110 position in the genome of M. tuberculosis. .......................................................................................... 100

4.3.4 Rapid definition of TB genetic lineages- PGG3 and PGG1 .... 103

4.4 Discussion ............................................................................................................................ 105

4.5 Summary ............................................................................................................................ 107

Chapter 5 Classification of Nepalese TB clinical isolates into different TB genetic lineages ...................................................................................... 108

5.1 Introduction .......................................................................................................................... 109

5.1.1 Aims and Objectives ...................................................................................................... 111

5.1 Materials and Methods ..................................................................................................... 111

5.1.1 Strains ............................................................................................................................. 111

5.1.2 IS6110 FAFLP PCR, Fragment Sizing & Analysis ......................................................... 112

5.2 Results .................................................................................................................................. 112

5.2.1 Analysis of Data using BioNumerics software v6.1 ..................................................... 112

5.3 Discussion ............................................................................................................................ 117

5.4 Summary ............................................................................................................................ 118

Chapter 6 Rifampicin Resistance status in Nepalese TB isolates from clinical samples ...................................................................................... 119

6.1 Introduction .......................................................................................................................... 120

6.1.1 Drug Resistant TB in Nepal (adapted from NTC, 2015) ........................................... 120

6.1.2 Aims and Objectives ...................................................................................................... 121
Chapter 6

6.2 Materials and Methods .................................................................................. 121

6.2.1 rpoB Analysis .............................................................................................. 121

6.3 Results .............................................................................................................. 122

6.4 Discussion ........................................................................................................ 123

6.5 Summary .......................................................................................................... 125

Chapter 7

7.1 Introduction ...................................................................................................... 127

7.2 Discussion of findings in this study ................................................................. 127

7.2.1 Optimisation and Development of Insertion Element IS6110 Fluorescent Amplified Fragment Length Polymorphism (FAFLP) PCR 127

7.2.2 Mapping of Insertion sites (IS6110) in the M. tuberculosis H37Rv reference genome and rapid definition of genetic lineages (published work - (179)) ................................................................. 128

7.2.3 Classification of Nepalese TB clinical isolates into different TB genetic lineages (published work - (188)) ................................................................. 133

7.2.4 Rifampicin Resistance status in Nepalese TB isolates from clinical samples (published work - (188,195)) ................................................................. 137

7.3 Future work ..................................................................................................... 141

7.4 Summary .......................................................................................................... 141

7.5 Conclusions ..................................................................................................... 146

References ............................................................................................................ 147

Appendices ............................................................................................................ 181
List of Tables

Table 1-1 TB timeline showcasing the important discoveries and events from the 19th century to the 21st century................................................................. 4
Table 1-2 Acid Fast Bacilli (AFB) Smear classification .................................. 15
Table 1-3 Summary of recommendations for testing for latent tuberculosis infection (LTBI)........................................................................................................ 19
Table 1-4 Composition of the different MIRU-VNTR sets widely used........ 37
Table 1-5 List of some common microorganisms studied using FAFLP analysis......................................................................................................................... 40
Table 1-6 Table showing the distribution of M. tuberculosis complex (MTBC) strains with respect to the Principal Genetic Groups (PGGs) and the TB spoligotypes families / lineages................................................................. 48
Table 1-7 Estimated TB burden in Nepal in 2014 ............................................ 58
Table 2-1 List of TaqI fluorescently labelled selective primers....................... 70
Table 2-2 Run Module settings for capillary electrophoresis of FAFLP fragments.................................................................................................................... 71
Table 4-1 List of unlabelled two base selective primers................................. 96
Table 4-2 showing the selected samples and their IS6110 FAFLP derived lineages ............................................................................................................... 98
Table 4-3 Mapping of H37Rv genome using 4-dye FAFLP PCR, DNA sequencing including BLAST results and in silico analysis with sequence orientation 5’-3’ according to Figure 3-1 Method Schematic....................... 100
Table 4-4 Mapping of H37Rv genome using 4-dye FAFLP PCR, DNA sequencing including BLAST results and in silico analysis with sequence orientation 3’-5’ according to Figure 3-1 Method Schematic....................... 101
Table 5-1 Common fragments identified using IS6110 FAFLP PCR in TB genetic lineages of the 176 bacterial DNA isolates in Nepal.................. 114

Table 6-1 List of mutations seen in rpoB Rifampicin Resistance- Determining Region (RRDR) of rifampicin resistant M. tuberculosis isolates from Nepal. .................................................................................................................................................. 122

Table 7-1 Brief summary of the advantages and disadvantages of IS6110 FAFLP ........................................................................................................................................................................... 140

Table 7-2 Summary of the results from all chapters in this thesis........... 142
List of figures

Figure 1-1 Estimated rates of incidence of new TB cases world-wide, 2015. 5

Figure 1-2 Milestones set by WHO to control TB by adopting End TB strategy ................................................................................................................................................... 6

Figure 1-3 Acid fast Ziehl Neelsen staining showing rod-shaped Mycobacterium tuberculosis bacteria ......................................................................................................................... 6

Figure 1-4 Unique Cell wall structure of Mycobacterium tuberculosis .......... 7

Figure 1-5 Pathophysiology of Tuberculosis infection .................................. 8

Figure 1-6 Histopathological features of lung infected by pulmonary tuberculosis showing caseation and formation of granulomas ...................... 11

Figure 1-7 LJ medium showing dry, rough, raised creamy white or yellow colour colonies, characteristic of M. tuberculosis ................................................................. 16

Figure 1-8 Bactec MGIT system for detection of M. tuberculosis drug resistance ................................................................................................................................. 16

Figure 1-9 Translated sequence alignment of ORFa in IS6110 and IS986 both from M. tuberculosis and IS987 from M. bovis BCG ........................................ 25

Figure 1-10 Pictorial representation of IS6110 sequence of M. tuberculosis genome showing Direct (DR in red colour) and inverted repeat (IR in green colour) sequences ........................................................................................................ 27

Figure 1-11. Principle of spoligotyping and the processing of signals......... 33

Figure 1-12 Pictorial representation of SpolDB4 classification analysis of different TB lineages ......................................................................................................................... 35

Figure 1-13 Novel evolutionary scenario of MTBC proposed by Brosch et al in 2002 based on the regions of difference (RD) and sequence polymorphisms within five genes .............................................................................................................................. 45
Figure 1-14 Evolutionary scenario proposed by Sreevatsan et al, in 1997 based on SNP polymorphism

Figure 1-15 Schematic representation of the proposed phylogenetic relationships between members of MTBC

Figure 1-16 The phylogeographical distribution of M. tuberculosis and its lineages

Figure 1-17 Timeline of evolutionary events in Mycobacterium tuberculosis

Figure 1-18 Evolution of Mycobacterium tuberculosis complex along with human population

Figure 1-19 A map of Nepal showing the topography of Nepal interlocked between India (in the south, east and west of Nepal) and China (in the north)

Figure 2-1 Geographical location of the two Nepal TB treatment centres, GENETUP and NTC, where samples were collected

Figure 2-2 Schematic representation of the IS6110 FAFLP methodology

Figure 3-1 Pictorial representation of the IS6110 FAFLP PCR methodology

Figure 3-2 Step-by-step procedure for optimising and developing the IS6110 FAFLP methodology

Figure 3-3 Step-by-step procedure of standardising IS6110 FAFLP methodology including the accepted methodology

Figure 3-4 Example of a Peakscanner screenshot showing the absence of amplification of fragments except for Liz600 marker in orange colour

Figure 3-5 Example of a Peakscanner screenshot showing strong signals of more than 28000 rfu when MyTaq mastermix was used
Figure 3-6 Example of a Peakscanner screenshot showing amplification of fragments with correct signal intensity in a sample when Platinum Pfx Polymerase proof-reading enzyme mastermix was used.................................. 84

Figure 3-7 Second example of the Peakscanner screenshot showing amplification of fragments from the sample above in the previous figure but a mixture of fragments noticed at a single position leading to difficult interpretation of the data when Platinum Pfx Polymerase proof-reading enzyme mastermix was used. .............................................. 85

Figure 3-8 Example of a Peakscanner screenshot showing strong raw signal intensity in H37Rv sample using recombinant Taq Polymerase reagents ... 86

Figure 3-9 Example of a Peakscanner screenshot showing double peaks (green dye- VIC) in H37Rv sample using recombinant Taq Polymerase reagents................................................................. 87

Figure 3-10 Example of a Peakscanner screenshot showing the presence of amplified fragments with single peaks in an H37Rv sample when using recombinant Taq Polymerase reagents. ........................................... 88

Figure 3-11 Peakscanner screenshot showing the absence of amplification of fragments in the negative sample (water) using recombinant Taq Polymerase reagents. ................................................................. 89

Figure 3-12 Example of an electropherogram showing two Nepal samples, N03 and N06, identical to each other having the same fingerprint with well-defined and easily identifiable coloured fragments. ................................. 90

Figure 4-1 Agarose Gel Electrophoresis showing M.tuberculosis strains , N10, N23, N34, N46, N62, N63, N70, H37Rv, amplified using PGG3 (A) and PGG1 (B) specific primers, amplifying 296 bp and 200 bp products respectively ................................................................. 104

Figure 5-1 UPGMA derived dendrogram showing the predominant genetic lineages/spoligotypes of 176 Nepalese Mycobacterium tuberculosis isolates. .......................................................................................... 116
Figure 7-1 A. Distribution of different TB lineages in Nepal from this study and B. the distribution of different lineages seen in India, China and other countries near the Indian subcontinent.
Abbreviations

AFLP- Amplified Fragment Length Polymorphism

CAS- Central Asian Strain

DOTS- Directly Observed Treatment short-course

DR- Direct Repeat

DR-TB- Drug Resistant Tuberculosis

DST – Drug Susceptibility Test

DVR- Direct Variable Repeats

ETH- Ethambutol

FAFLP- Fluorescent Amplified Fragment Length Polymorphism

GENETUP- German Nepal Tuberculosis Project

HIV- Human Immunodeficiency Virus

IE- Insertion Element

INH- Isoniazid

ipl- IS\textit{6110} preferential loci

IS- Insertion Sequence
LAM- Latin American-Mediterranean

MDG- Millennium Development Goal

MDR- Multidrug Resistant

MIRU- Mycobacterial Interspersed Repetitive Units

MPTR- Major Polymorphic Tandem Repeats

MTB- *Mycobacterium tuberculosis*

MTBC- *Mycobacterium tuberculosis* Complex

NTP- National Tuberculosis Programme

PCR- Polymerase Chain Reaction

PRZ- Pyrazinamide

RFLP- Restriction Fragment Length Polymorphism

RIF- Rifampicin

TB- Tuberculosis

TDR- Totally Drug Resistant

VNTR- Variable Number Tandem Repeats

WHO- World Health Organisation
XDR- Extensively Drug Resistant
**Glossary of Key terms**

Genotype- Genetic makeup of any organism such as *Mycobacterium tuberculosis*.

MDR-TB- Multidrug-resistant TB (MDR-TB) is TB that is resistant to at least isoniazid and rifampicin, the two most powerful anti-TB drugs.

PCR- Polymerase chain reaction is an enzymatic method where one copy of segment of DNA is amplified into several million copies of DNA.

Tuberculosis- Infectious disease caused by the bacteria *Mycobacterium tuberculosis* complex

XDR-TB- Form of TB which is resistant to at least four of the core anti-TB drugs including isoniazid and rifampicin.
Outline of the thesis

This research work is divided into chapters that explain the development of an IS6110 FAFLP genomic tool box that could be used in resource poor settings.

Chapter 1 sets the scene for this thesis by reviewing the literature extensively. It gives a brief introduction about tuberculosis infection including diagnostics, treatment and TB epidemiology. Then an extensive review of the available molecular epidemiological markers is discussed including the IS6110 marker. The chapter ends with the overview of the study setting (Nepal), leading to aims and objectives of this research work.

Chapter 2 discusses the general methods that were used to achieve the objectives of this project.

Chapter 3 describes in detail how the IS6110 FAFLP PCR assay was developed and standardised.

Chapter 4 shows for the first time how the standardised method was tested by mapping the IS6110 insertion sites in the reference genome H37Rv and by showing the development of rapid detection PCR targeting specific lineages.

Chapter 5 describes characterisation of TB samples from Nepal, a resource poor setting, into different TB genetic lineages for the first time using IS6110 FAFLP PCR.
Chapter 6 describes rifampicin sensitivity in the Nepalese TB samples using RIF resistant PCR assay.

Chapter 8 finally summarises all the above chapters and concludes by putting the results into context.
My Role in the thesis

I started on this research work in 2011 as a part-time PhD student whilst still continuing my day job as a Specialist Healthcare Scientist in PHE, Colindale. I was supervised by my manager, Dr. Catherine Arnold, who was also my principal supervisor for my thesis work. My three other supervisors were Prof. Tim McHugh (UCL), Prof. Pam Sonnenberg (UCL) and Prof. Ibrahim Abubakar (UCL).

Under the able supervision of my principal supervisor and the guidance of my other supervisors, I designed and developed the method called IS6110 FAFLP PCR. I tested the procedure on the Nepal TB samples donated kindly by our collaborators in Nepal, Dr. Deanna Haggie and Ms. Saraswoti Khadge.

I initially mapped the insertion sites in the MTB genome H37Rv followed by designing new two-base TB genotype/ lineage specific primers that could rapidly identify and characterise the circulating lineages in any geographical location. At the same time, I also characterised the Nepalese TB samples into different TB lineages. At the time of this experiment, it was the first known experiment to delineate Nepalese TB samples as there was no information available publicly on the distribution of the circulating lineages in this region. During the course of the thesis, I also tested the drug resistance status of these Nepalese strains.

During the course of the thesis, I have published three manuscripts in peer reviewed journals (as discussed in chapters 4, 5 and 6). I have also presented a number of posters in various domestic and international
conferences. I have also attended the skills courses both at UCL and PHE and updated my skills during the thesis.

Publications and conference presentations

Manuscripts published from this research work


- Hameed S, Moganeradj K, Mahmood N, McHugh TD, Chaudhry MN, Arnold C (2017) Sequence analysis of the rifampicin resistance determining region (RRDR) of rpoB gene in multidrug resistance confirmed and newly diagnosed tuberculosis patients of Punjab, Pakistan. PLoS One, 12(8): e0183363. (This part of the work was carried out whilst I supervised another PhD student from Pakistan, by employing the tools used in chapter 6).

Posters presented

1. Kartyk Moganeradj, Dunstan Rajendram, Ibrahim Abubakar, Timothy D McHugh, Pam Sonnenberg, Deanna A Hagge, Murdo Macdonald, Saraswoti


Acknowledgements

I would like to thank all my four supervisors, Dr. Cath Arnold (PHE), Prof. Tim Mchugh (UCL), Prof. Pam Sonnenberg (UCL) and Prof. Ibrahim Abubakar (UCL) for their valuable time and overwhelming support, sharing their knowledge and experience in shaping my PhD thesis into a challenging and commendable piece of research work. Without their supervision and their unified guidance, I would not have completed this research and published three articles in peer-reviewed journals. My special thanks to Dr. Cath Arnold, my principal supervisor, for guiding and encouraging me throughout my PhD tenure, for pulling my spirits up whenever I was down and also helping me to achieve my goal of finishing the PhD. Her immense support by allowing me to work flexible hours when I was facing personal issues needs a special mention as I would have otherwise floundered during the middle of my research work.

I owe my thanks to my secondary supervisor, Prof. Tim McHugh, for being those extra pair of eyes in looking into my thesis and suggesting corrections. The numerous meetings I had with you definitely helped me to increase the level of scientific acumen needed for the research work.

Many thanks to my tertiary supervisor, Prof. Pam Sonnenberg, whose keen eye helped me to re-organise my thoughts and pay attention to the structure of the thesis.
I would like to thank my tertiary supervisor, Prof. Ibrahim Abubakar, whose constant encouragement motivated me whenever I was looking down the barrel during the PhD meetings.

I also thank my colleagues, both past and present, in the Genomics Services Development Unit, where I did my PhD part-time for giving me space and time to concentrate on my research work. I owe my special thanks to Dunstan, Ali, Des, Meeta, Kirstin, Julie, Raju, Danny, Sandra, Jumie, Greg and Andrew for their constant nagging to get my thesis finished. I also would like to extend my gratitude to my mentors, teachers and all my friends from my previous work places in U.K. and in India.

I take this opportunity to thank Public Health England (previously Health Protection Agency) for partly funding my PhD. I also thank University College London for accepting me as a PhD student part-time.

I would also like to thank my past manager from SGUL, Prof. Steve Goodbourn and his group, my RVC supervisor, Prof. Peter Russell and my RIVER professors especially Dr. S. Ram Kumar for supporting me at various stages of my career.

I owe my special thanks with a lot of gratitude to my parents and sister, who saw in me that one day I will definitely make them proud. I thank my aunty, Bibi and my guardians, Babu and Asha, for kick-starting my aspirations in U.K.
I thank overwhelmingly my wife, Vimala, for her support and holding hands with me firmly during this arduous but successful seven year journey and never letting me go mad.

Last by no means the least, I owe my special thanks to my angel and my beautiful daughter Yashwini who taught me how to stay strong and put up a smile, when faced with adversity. She always found it funny to see her Daddy working harder than her and studying at night.

I dedicate this thesis to my parents, my wife and my daughter who sacrificed a lot to make this happen for me.
Chapter 1 Introduction
Tuberculosis (TB) caused by the *Mycobacterium tuberculosis* complex (MTBC), is ranked as one of the top ten causes of death worldwide (WHO 2015) with an estimated 10.4 million people falling ill with TB and 1.3 million TB deaths in 2016, exacerbated by HIV co-infection and the development of drug resistance (1).

1.1 *Mycobacterium tuberculosis* complex (MTBC)

*Mycobacterium tuberculosis* belongs to the genus *Mycobacterium* in the family *Mycobacteriaceae*, order *Corynebacteriales* and phylum *Actinobacteria* (2).

TB caused by the MTBC encompasses the following members including proposed new members that affect mongooses and meerkats (3–6).

1. *M. tuberculosis*

2. *M. africanum* (phylogenetic variant of *M. tuberculosis*).

3. *M. bovis*

4. *M. caprae*

5. *M. microti*

6. *M. pinnipedii*

7. “*M. mungi*”

8. “*M. orygis*”
9. “M. suricattae”

10. “M. canetti”

1.2 TB Timeline

Before Robert Koch, two French physicians Rene Theophile Hyacinthe Laennec and Jean-Antoine Villemin reported in 1821 and 1868 respectively the first accounts of tuberculosis, even though the causative organism was unidentified at that time (7,8). However, Robert Koch’s enormous body of work from identification of the microorganism, developing the staining technique and culture methods for Mycobacterium tuberculosis, illustrating the mode of transmission to recommending isolation of the patients with active TB has to be considered as the greatest achievement in the science of TB and has now paved the way for various studies. Koch’s failed study to treat tuberculosis by using supernatants from M.tuberculosis culture which is now known as ‘tuberculin’ when injected in the skin gave rise to cell-mediated immune responses which are responsible for the clinical manifestations of the disease (9,10). Subsequently in the years that followed, various scientists and physicians were able to relate that the host immune responses especially cell-mediated immune responses played a major role in the pathogenesis of TB (9). The important discoveries and outcomes in relation to the TB timeline is given in the table 1.1 (adapted from (11)).
Table 1-1 TB timeline showcasing the important discoveries and events from the 19th century to the 21st century

<table>
<thead>
<tr>
<th>Year</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>1882</td>
<td>Rober Koch discovers the causative organism of tuberculosis, <em>M. tuberculosis</em></td>
</tr>
<tr>
<td>1912</td>
<td>Koch’s discovery</td>
</tr>
<tr>
<td>1921</td>
<td>An attenuated vaccine from <em>M. bovis</em> developed by French scientists Albert Calmette and Camille Guerin was put into use in humans</td>
</tr>
<tr>
<td>1943</td>
<td>Streptomycin discovered in 1943 by american scientists, Selman Waksman, Albert Schatz, and Elizabeth Bugie, was found to be exhibiting antibacterial properties along with para-amino salicylic acid (PAS)</td>
</tr>
<tr>
<td>1948</td>
<td>First successful use of streptomycin in trials in humans</td>
</tr>
<tr>
<td>1948</td>
<td>Combination therapy of S+PAS+H for 24 months</td>
</tr>
<tr>
<td>1952</td>
<td>Pyrazinamide discovery</td>
</tr>
<tr>
<td>1952</td>
<td>Rifampicin discovery</td>
</tr>
<tr>
<td>1954</td>
<td>Ethambutol discovery and replaces PAS in combination therapy for 18 months</td>
</tr>
<tr>
<td>1957</td>
<td>Rifampicin included in combination therapy for 9-12 months and first outbreak of drug resistant TB in US</td>
</tr>
<tr>
<td>1961</td>
<td>BCG vaccination included in UNICEF’s expanded immunisation programme</td>
</tr>
<tr>
<td>1963</td>
<td>First meeting by WHO to discuss TB-HIV coinfection</td>
</tr>
<tr>
<td>1970-1974</td>
<td>HIV impact</td>
</tr>
<tr>
<td>1974</td>
<td>WHO declares TB as major public health issue as the TB related mortality was higher than previous years</td>
</tr>
<tr>
<td>1974</td>
<td>Anti-TB drug resistance survey and a new framework for TB control including DOTS</td>
</tr>
<tr>
<td>1980</td>
<td>Short term course for 6-8 months was introduced</td>
</tr>
<tr>
<td>1980</td>
<td>First genome sequencing of Mycobacterium tuberculosis</td>
</tr>
<tr>
<td>1980</td>
<td>WHO framework to completely eliminate TB by 2050 by adopting STOP TB strategy globally</td>
</tr>
<tr>
<td>1991</td>
<td>WHO initiative for TB elimination</td>
</tr>
<tr>
<td>2001</td>
<td>STOP TB framework</td>
</tr>
<tr>
<td>2014</td>
<td>Post 2015 global TB strategy</td>
</tr>
<tr>
<td>2020</td>
<td>TB elimination</td>
</tr>
</tbody>
</table>

According to WHO in 2016, the Millennium Development Goal (MDG) by the United Nations to stop the spread of TB by 2015 has already been achieved in all six WHO regions, and in sixteen of the 22 high burden countries that collectively accounts for more than 80% of the TB cases worldwide. TB incidence has dropped to a level by 1.5% every year since 2000 and consequently, it is 18% lower than the levels of 2000. However, it is still a major global health problem of the developing world in heavily populated countries like India and China which account for more than a quarter of new TB cases in addition to sub-Saharan Africa. The number of incident TB cases varies from country to country and it is predominantly lower in high-income countries. (fig.1.1) (12). The MDGs have now been replaced by Sustainable Development Goals (SDGs) and the Stop TB strategy to the End TB strategy as shown in the figure 1-2

![Figure 1-1 Estimated rates of incidence of new TB cases world-wide, 2015](Taken from WHO's global tuberculosis report 2016).
M. tuberculosis is a rod-shaped non-spore forming aerobic acid-fast bacterium (see figure 1-3) having a unique cell wall structure that supports their survival (13).

Figure 1-3 Acid fast Ziehl Neelsen staining showing rod-shaped Mycobacterium tuberculosis bacteria

(Taken from http://www.who.int/tb/strategy/end-tb/en/)

1.3 Morphology of *Mycobacterium tuberculosis*

*M. tuberculosis* is a rod-shaped non-spore forming aerobic acid-fast bacterium (see figure 1-3) having a unique cell wall structure that supports their survival (13).
The cell wall consists of mycolic acid covalently attached to the underlying peptidoglycan bound polysaccharide arabinogalactan providing a protective barrier against host defence mechanisms (fig 1-4) thus enabling the bacterium to be resistant to the action of antibiotics. In addition, the cell wall also contains lipoarabinomannan, a carbohydrate structural antigen on the outside of the organism that is immunogenic and facilitates the survival of mycobacteria within macrophages (14,15).

Figure 1-4 Unique Cell wall structure of *Mycobacterium tuberculosis*  
(Taken from http://www.cell.com/cms/attachment/607349/4834477/gr1.jpg)

1.4 Immunopathogenesis of Tuberculosis

*M. tuberculosis* infection is caused by aerosols generated from a person infected with pulmonary or laryngeal tuberculosis. These droplets containing two to three *M. tuberculosis* organisms can stay airborne for minutes to hours after expectoration (16). The size (1-5 µm in range) and concentration of the
aerosols also have been suggested to increase the TB transmission risk in patients with intra thoracic disease (17). The number of bacilli in the droplets, the virulence of the bacilli, exposure of the bacilli to the UV light, degree of ventilation and occasions for aerosolisation all influence transmission (fig. 1-5). *M. tuberculosis* not only affects the respiratory system when it enters the lungs but can also affect the lymphatic system, pleura, bones/joints, or meninges and thus cause extra pulmonary tuberculosis (14).

**Figure 1-5 Pathophysiology of Tuberculosis infection**

A) Inhalation of bacilli, B) Formation of granulomas containing Macrophages and T-lymphocytes and C) Collapse of granuloma and release of alveoli to other alveoli or other organs (Adapted from CDCP).
1.5 Clinical signs and symptoms

The active lung tuberculosis is characterised by cough with sputum (blood at times), chest pains, loss of appetite, general weakness, weight loss, fever, and night sweats.

The various stages of disease presentation in TB are as follows (13) as shown in figure 1-5:

- In early infection or primary disease, the hosts’ immune system combats the infection and the patients present with fever, paratracheal lymphadenopathy or dyspnoea. Infection may not progress to active disease and could be subclinical.

- There are two stages in the active disease presentation. In early primary progressive stage or active disease, the immune system fails to prevent the onset of infection and the inflammation of tissues follows. As the patients have nonspecific signs like fatigue, weight loss and fever followed by the development of non-productive cough, the diagnosis could be difficult: It is because the findings on chest radiographs might be normal and the sputum smear microscopy might not contain any mycobacteria (‘negative’ result).

- In late primary progressive/ active disease or active tuberculosis, the symptoms seen are productive cough leading to purulent sputum, progressive weight loss, chills, night sweats, respiratory crackles and anaemia. This state is an infectious state harbouring more bacilli than latent infection (18). Due to the excessive loss of both fat and lean tissue, this disease is also called wasting disease. Diagnosis is carried
out by sputum cultures as sometimes the chest radiographs appear normal.

- In latent infections or latent tuberculosis infection (LTBI), mycobacteria persist in the body without any clinical signs or symptoms and the disease could be reactivated when the immune system fails to counteract the infection or if immunosuppressed. Also, calcification and fibrosis of the granulomatous lesions are noticed in radiological findings. The formation of granulomas for persons with intact cell-mediated immunity is an accumulation of activated T-lymphocytes and macrophages. This micro-environment destroys the macrophages and produces early solid necrosis at the centre of the lesion but the bacteria survive this condition by modifying its phenotypic expression as suggested by Li et al. (19), by altering its expression to changes to pH and anaerobic growth conditions. By two or three weeks, the necrotic environment resembles soft cheese called caseous necrosis characterised by low pH, low oxygen levels and limited nutrients which leads to latency. Persons with adequate immunity develop fibrosis and calcification which contains the bacilli in these lesions whereas in those with impaired immunity develop primary progressive TB (20). The necrotic material undergoes liquefaction and the fibrous wall loses structural integrity in less immunocompetent persons which can then drain into the bronchus or a nearby blood vessel. Unique histopathological features of the lungs (granulomas of multinuclear cells, giant cells and caseation) in pulmonary tuberculosis are seen in figure 1-6. If the infected person coughs, droplet infection can occur
and also if discharged into the blood vessel it leads to extra pulmonary tuberculosis (fig 1.5).

- In extra pulmonary tuberculosis, if it affects the central nervous system it leads to meningitis which left untreated results in deaths. If it affects the circulatory system causing infection of the blood stream, it leads to disseminated or miliary tuberculosis. When it affects the lymphatic system, it is called lymphatic tuberculosis and the most common presentation is cervical lymphadenopathy. In some cases, it is also shown to affect the joints, bones, pleura and the urogenital system (13).

Figure 1-6 Histopathological features of lung infected by pulmonary tuberculosis showing caseation and formation of granulomas

1.6 Diagnosis of active tuberculosis

For TB treatment to be effective, a rapid and accurate diagnosis of drug resistant TB is important in ensuring the timely clinical intervention and controlling the rise in drug resistant TB (21,22). According to the WHO Global Tuberculosis report 2014, only 64% of the estimated nine million people were reported for newly diagnosed TB cases and more than three million cases went unnoticed by the health authorities either because they were not diagnosed or not reported to National TB Programs (NTP) (23). The notification rate increased from 2013-2015 according to the recent report by WHO due to the increase in notification rate (34%) by India. In 2015, 6.1 million new TB cases were notified out of 10.4 million incident cases (12). Even though TB diagnosis still depends on both the smear microscopy and sputum culture for identification in many countries, modern techniques are rapidly shifting the diagnostic landscape of tuberculosis.

The diagnosis of TB disease generally is based on a combination of clinical symptoms, chest X-ray examination and laboratory tests. In a TB dedicated laboratory clinical setting, diagnosis of active tuberculosis involves sputum smear microscopy, identification of the TB bacillus using culture techniques, phenotypic drug susceptibility tests and molecular tests. If infection is suspected then tuberculin skin test (TST), the interferon gamma release assays (IGRA) and acid fast staining of the sputum smear are conducted. (21).
1.6.1 Chest radiographs

Chest radiographs are important in the diagnosing of active tuberculosis with prominent features including upper lobe consolidation, cavitation and pleural effusions. Additional tests like computed tomography (CT) scanning are needed along with the other tests for the diagnosis of progressive primary tuberculosis or post primary tuberculosis which occurs at the first or second year after an initial infection (24, 25).

Following chest radiographs or the initial diagnosis by the clinical symptoms, the common tests used as initial screening tests to differentiate the TB infection from disease are the tuberculin skin test, interferon-gamma release assays (IGRA) and Acid Fast Bacilli (AFB) smear staining (26).

1.6.2 Tuberculin Skin test (TST) or Mantoux test

TST is a preliminary test to identify people who are suspected to have TB infection. A sterile concoction of antigens from seven strains of *M. tuberculosis* called purified protein derivative (PPD) or tuberculin is injected intradermally to test the sensitivity of the skin to the PPD after 48 and 72 hours and the test is interpreted based on the size of induration. A positive reaction should be considered with caution as it does not exclude previous BCG vaccination, Mycobacteria other than tuberculosis (MOTT) infection and latent TB. Also due to the increased frequency of false negatives, a negative reaction does not exclude active or latent TB infection and it warrants extra tests.
1.6.3 Interferon- Gamma release assay (IGRA)

The IGRA test is used in diagnosing TB infections and has been reported to be more sensitive than TST for both active and latent pulmonary TB infections. It quantifies the T-cell response especially the interferon-gamma (IFNγ) by either directly measuring the concentration of IFNγ in the serum (Commercial kit: Quantiferon TB Gold) or by counting the number of T-cells that releases IFNγ (Commercial kit: TSPOT.TB)(26).

1.6.4 Sputum microscopy

Sputum smear microscopy is preceded by acid-fast staining (Ziehl-Neelsen or Auramine fluorochrome stains are currently used) of clinical material and is the most important test for the diagnosis and screening of active tuberculosis in poor and low-income countries (27). WHO describes the presence of at least one acid fast bacillus (AFB) in at least one sputum sample as a smear positive pulmonary TB case. The guidance for interpreting smear results by WHO (a similar but slightly different classification is followed by Centers for Disease Control and Prevention in the USA) is given in table 1-2. It is widely used and effective in countries where TB is endemic and the sensitivity of microscopy has been reported to be higher than 80% (28,29) but a recent report by WHO states that it is around 20% (30). Sensitivity of the test decreases in HIV positive patients than in HIV negative cases and thus been associated with poor prognosis including death (31,32).
Table 1-2 Acid Fast Bacilli (AFB) Smear classification
(Adapted from WHO TB microscopy handbook)

<table>
<thead>
<tr>
<th>Number of AFB seen per stated field (100x objective)</th>
<th>Smear Interpretation (Grade)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 AFB in 100 fields</td>
<td>Negative</td>
</tr>
<tr>
<td>1-9 AFB in 100 fields</td>
<td>Record exact number of bacilli</td>
</tr>
<tr>
<td>10-99 AFB in 100 fields</td>
<td>1+</td>
</tr>
<tr>
<td>1-10 AFB / field, check 50 fields</td>
<td>2+</td>
</tr>
<tr>
<td>&gt;10 AFB / field, check 20 fields</td>
<td>3+</td>
</tr>
</tbody>
</table>

1.6.5 Sputum culture

Culturing of mycobacteria follows the staining protocols either in a solid or a liquid medium after decontamination of the specimen. The recommended solid media by the European Respiratory Society, WHO, the American Thoracic Society and the International Union Against Tuberculosis (IUAT) are Lowenstein- Jensen (LJ) and Middlebrook 7H10 and 7H11 (MB) media (see fig 1-7 and 1-8). WHO has offered guidelines for the use of liquid media like Mycobacterium Growth Indicator Tube (MGIT) since they offer a more sensitive and quicker turnaround time than solid media (33,34).
Figure 1-7 LJ medium showing dry, rough, raised creamy white or yellow colour colonies, characteristic of *M. tuberculosis*  
(Taken from https://microbeonline.com/preparation-uses-lownstein-jensen-lj-medium/)

Figure 1-8 Bactec MGIT system for detection of *M. tuberculosis* drug resistance  
(Picture adapted from www.bd.com)
1.6.6 Molecular identification test and drug susceptibility tests

As drug resistance is a major problem in controlling tuberculosis especially with the rise in multidrug resistant (MDR) and extensively drug resistant (XDR) TB, drug susceptibility tests are carried out before starting treatment for tuberculosis. These tests are either phenotypic tests involving culture techniques or by adopting molecular tests. Phenotypic methods are based on testing sensitivity to first line anti-TB drugs and the most common being rifampicin, isoniazid, ethambutol and pyrazinamide. As these tests are time-consuming and may be ineffective in detecting low-level drug resistance, molecular tests are widely used. Commercial kits are available on the market based on the PCR amplification of specific genes including drug resistance genes and these molecular assays are grouped as either Nucleic Acid Amplification Tests (NAATs) or Line Probe Assays (LPA). The GeneXpert assay, a fully automated RT-PCR NAAT assay, is most suitable for patients with suspected pulmonary tuberculosis and for specific forms of extrapulmonary tuberculosis as in the case of meningitis in people with HIV and lymphadenitis, but not pleural, pericardial, or abdominal tuberculosis (34) and also determine the resistance to one of the most common TB drugs, Rifampicin (34). In countries like Tanzania which is a TB endemic country, the GeneXpert assay has been cost effective in determining the rifampicin-resistant status in the patient population (35). A second generation cartridge, the Xpert MTB/RIF Ultra (Ultra) is due to be launched on world TB day in 2017 by Cepheid and Rutgers and assessed in the TB centres globally using the current GeneXpert instruments (12).
1.6.6.1 Line Probe assays (LPA)

Line Probe Assays are based on the PCR amplification of specific fragments of the *M. tuberculosis* genome followed by hybridisation of PCR products to the oligonucleotide probes immobilised on membranes. The commercial kits that are available in the current market are INNO-LiPa Rif.Tb (Innogenetics, Belgium), GenoType MTBDR/ MTBDRplus (Hain Lifescience, Germany) and Genotype MTBDRsl (Hain Lifescience, Germany). GenoType MDRTBplus is the only kit that could detect isoniazid resistance and GenoType MTBDRsl detects resistance to fluoroquinolones and ethambutol, the second-line drugs that are used in MDR-TB and XDR-TB cases (36–40). There are several next generation technologies including GeneXpert cartridges and Hain MTBDRsl assays or rapid next generation whole genome sequencing methods, are in the advanced stages of development (41,42).

1.6.7 Diagnosis of Latent tuberculosis

Diagnosis of LTBI depends upon two factors: a. Likelihood of infection with *M. tuberculosis* and b. Likelihood of progression to TB disease (43). The treatment strategy for LTBI is shown below in table 1.3.
Table 1-3 Summary of recommendations for testing for latent tuberculosis infection (LTBI)

(taken from 43).

<table>
<thead>
<tr>
<th>Group</th>
<th>Testing Strategy</th>
<th>Considerations</th>
</tr>
</thead>
</table>
| Likely to be Infected High Risk of Progression (TST ≥ 5mM) | Adults: Acceptable: IGRA OR TST  
Consider dual testing where a positive result from either result would be considered positive  
Children ≤ 5 years of age  
Preferred: TST  
Acceptable: IGRA OR TST  
Consider dual testing where a positive result from either would be considered positive | Prevalence of BCG vaccination  
Expertise of staff and/or laboratory  
Test availability  
Patient perceptions  
Staff perceptions  
Programmatic concerns |
| Likely to be Infected Low to Intermediate Risk of Progression (TST ≥ 10mM) | Preferred: IGRA where available  
Acceptable: IGRA or TST | |
| Unlikely to be Infected (TST > 15mM) | Testing for LTBI is not recommended  
If necessary:  
Preferred: IGRA where available.  
Acceptable: Either IGRA OR TST  
For serial testing:  
Acceptable: Either IGRA OR TST  
Consider repeat or dual testing where a negative result from either would be considered negative | |

Though conventional technologies still play a major role in the diagnosis of TB, the recent availability of new rapid tests has the potential to improve TB care. For example, the introduction of GeneXpert MTB/RIF assay using the GeneXpert instruments has enabled the rapid detection of tuberculosis in hours, in countries participating in the STOP TB campaign. In addition this system, by detecting rifampicin resistance more rapidly, has helped in the treatment of MDR-TB cases much more efficiently and also has been recommended for HIV-coinfected cases where smear microscopy tests could be less sensitive as described earlier (31,44). In the year 2015 alone, 6.2 million test cartridges were procured in 122 of the 145 countries eligible to purchase them at concessional prices showing a shift in the diagnosis landscape of tuberculosis (12). There is still a requirement for newer technologies which are much more affordable than the currently available strategies (45).
The advancement in new diagnostics and drugs that has been progressing over the last decade would help in combatting tuberculosis by improving the diagnosis of the disease and also early detection of drug resistant tuberculosis (37,46). As per the review by Drobniewski et al., it is important to identify TB cases including the drug resistant strains, early and accurately (37) and to arrest TB transmission by providing appropriate therapy and curing patients.

1.7 TB treatment

A minimum of six months' treatment is recommended which is divided into two phases, the initial phase and the continuation phase (44). The standard treatment regimen according to WHO (47) for both pulmonary TB including HIV-coinfected individuals and for extra-pulmonary TB cases is as follows:

a. Initial phase: Daily medication or therapy five days/week under directly observed therapy (DOTS) with isoniazid (INH), rifampicin (RIF), pyrazinamide (PZA) and ethambutol (EMB) for 2 months. A sputum smear test is taken and the next phase is only started if the test is negative.

b. Continuation phase: Daily medication or therapy five days/ week under DOT with isoniazid and rifampicin for the next four months.

According to the review by Fonseca et al., (48) multidrug resistant (MDR) and extensively drug resistant (XDR) TB strains pose a significant threat to the control of tuberculosis as they develop resistance to first line drugs mentioned above and work still needs to be done to understand the mechanism of resistance. If there is resistance to two of the first line drugs,
RIF and INH, then these strains are termed as MDR- TB strains. The treatment course for MDR-TB lasts for more than 2 years with second line TB drugs like fluoroquinolones and injectable aminoglycosides like streptomycin. Anti-TB drug resistance occurs due to the chromosomal mutations at a rate of $10^{-3}$ to $10^{-9}$ in genes encoding drug targets (49,50). If the treatment is not completed or if the proper regimen is not followed then it results in the poor treatment of MDR-TB patients thus amplifying the resistance of the first line drugs and leading to the development of XDR-TB (34).

1.7.1 Rifampicin (RIF)

Rifampicin inhibits RNA transcription and thereby protein synthesis by binding to the \textit{rpoB} encoded subunit of RNA polymerase. Mono resistance to RIF is a rare occurrence as nearly 85-90% of RIF resistant strains are also resistant to INH. RIF resistance therefore acts as a surrogate marker for MDR-TB (51–53). A high percentage of mutations (90-95%) occurs in the 81bp rifampicin resistance-determining region (RRDR) of the \textit{rpoB} gene between codons 507-533 and a lower level of resistance (5-10%) is seen in the N-terminal or cluster II region of \textit{rpoB} / other genes at codon V146 (51,54). It has been suggested that higher frequency of mutation is seen in some strains at certain positions of the RRDR, especially at the codon 531 where serine is replaced by leucine, due to the higher relative strain fitness than other strains (55).

1.7.2 Isoniazid (INH)

INH a bactericidal drug which targets the NADH- specific enoyl acyl carrier protein (ACP) reductase (\textit{inhA}) and prevents the production of mycolic acid
which leads to the inhibition of dividing mycobacterial cells (51,56). INH resistance is commonly noticed due to mutations within the katG gene at codon 315 or low level resistance mutations are seen in the inhA regulatory region (51,56,57). However, oxidative stress does not seem to impair the mutation rates in Isoniazid resistant strains (58).

1.7.3 Pyrazinamide (PZA)

PZA, activated by pyrazinamidase (pncA), is effective against M. tuberculosis by inactivating the fatty acid synthesis pathway. In addition, it also prevents protein translation and the ribosome sparing process of translation by binding to ribosomal protein S1 (rpsA) in M. tuberculosis. PZA resistance is seen in strains containing mutations in pncA or rpsA (51,59).

1.7.4 Ethambutol (EMB)

EMB prevents the formation of cell wall assembly by inhibiting the synthesis and polymerisation of cell wall arabinan leading to the accumulation of free mycolic acid. Resistance mutations are commonly seen in three emb genes at codons 306, 406 and 497(60).

1.7.5 Second line drugs

Some of the second-line drugs used in TB treatment, which include Streptomycin (SM), Kanamycin (KAN) and Amikacin (AMI) inhibit protein synthesis and are bactericidal drugs. Fluoroquinolones are also widely used as anti-TB drugs especially Ofloxacin (OFX), Levofloxacin (LFX) as bacteriostatics and Gatifloxacin (GFX) and Moxifloxacin (MFX) as bactericidals by inhibiting DNA gyrase leading to cell apoptosis by preventing
DNA replication and repair as shown by the Rapid Evaluation of Moxifloxacin in Tuberculosis (REMox TB) phase 3 study trial by Gillespie et al (61,62).

1.7.6 Vaccination

The BCG (Bacillus Calmette-Guerin) vaccine, designed to prevent tuberculosis is derived from continuous passaging of an *M. bovis* strain and has been in the public health domain for the last eighty years (63,64) This vaccine induces a strong T-cell response by activating the Th1 cells but it offers only limited protection in adults against pulmonary tuberculosis as the immune response is not prolonged enough to prevent *M. tuberculosis* infection developing in older individuals (65,66) (64), but is effective in infant TB meningitis (67). In the last decade there has been continuous research to develop an effective vaccine that could either complement the BCG vaccine or act independently to prevent TB infection and provide a long-term protective immunity. The Modified- Vaccinia – Ankara (MVA) 85A vaccine that was trialled in children, was found to be ineffective in children in offering protection against TB (68).

1.8 Molecular markers in TB epidemiology

Various molecular tools have been employed to study the epidemiology of tuberculosis based on the availability of the instruments and the ability to genotype clinical isolates. The molecular tools are broadly divided into non-PCR based tools - Restriction Fragment Length Polymorphism (RFLP) and Polymorphic GC-rich Repetitive Sequences RFLP (PGRS-RFLP) and PCR based tools (Spoligotyping, Mycobacterial Interspersed Repetitive Unit –
Variable Number Tandem Repeat (MIRU-VNTR), Amplified Fragment Length Polymorphism (AFLP) and Whole Genome Sequencing (WGS) methods.

1.8.1 IS\textit{6110} and IS\textit{6110} RFLP

Bacterial transposons and Insertion Sequences (IS) are evolutionarily informative as they alter the genetic makeup of the host organisms during transposition (69). Insertion sequences are mobile genetic elements that code for transposases evolved to move the IS around the genome. A study in 1991 (70) showed that the \textit{M. bovis} BCG insertion element IS\textit{987} is virtually identical to the previously described IS elements IS\textit{986} (71) and IS\textit{6110} from \textit{M. tuberculosis}. The only biologically significant difference is the presence, in IS\textit{987}, of OR\textit{Fa} (Open Reading Frame) in one single ORF, whereas IS\textit{986} and IS\textit{6110} contain OR\textit{Fa}, composed of two different ORFs (70,72) (see figure 1.9).
Figure 1-9 Translated sequence alignment of ORFa in IS6110 and IS986 both from *M. tuberculosis* and IS987 from *M. bovis* BCG

(Asterisks denotes the matching amino acid residues in all three sequences) (Adapted from (70)).
Tracing the movement of these elements in the host genomes is productive in the identification of different strains of microorganisms. IS can transpose within the genomes thereby altering the position and even function of the adjacent genes, for example insertion of an insertion element (IE), IS200-like element, into the genome of Yersinia pestis inactivates the invasin gene which enables the bacteria to invade the host organism (73). IS6110 insertions have been found in some of the essential genes in drug resistant strains (74). These changes contribute to the evolution of microorganisms introducing genomic plasticity facilitating survival in different environmental conditions (75).

IS elements have been used as taxonomic markers over the years in certain bacteria like Bordetella pertussis, Salmonella typhi etc. (76). An IS-like element IS6110, with inverted (28bp with 3 mismatched bases) and direct (3bp) repeats at its ends, was identified as a repeated sequence from an M. tuberculosis cosmid library constructed in pH79 (77,78) by screening the library with labelled M. tuberculosis total DNA. This element is a 1.36kb insertion sequence found only in MTBC members and belongs to the enterobacterial Insertion Sequence 3 (IS3) family of insertion sequences. It contains an ORF encoding transposase of 1037bp length and has 28bp imperfect inverted repeats at both the ends of the sequence (78) (fig 1-10).
Figure 1-10 Pictorial representation of IS6110 sequence of *M. tuberculosis* genome showing Direct (DR in red colour) and inverted repeat (IR in green colour) sequences.
The first instance of the use of IS\textit{6110} as a probe for the identification of MTBC was reported in 1990 (78). MTBC from uncultured specimens, were identified using primers designed from the IS\textit{6110} sequence in a PCR reaction and results were successfully correlated with other identification procedures (78). This element is still being used as a tool for DNA fingerprinting to the present day as it is conserved in the MTBC (79, 80). IS\textit{6110} copy numbers vary from strain to strain; one in \textit{M. bovis} and >25 in \textit{M. tuberculosis} (81). Also, three to four nucleotides of the genomic sequence are repeated at the extremities of some IS\textit{6110} copies or the site of insertion suggesting a role of transposition of these elements in MTBC genomes (82). An oligonucleotide ligation assay (OLA) using fluorescently labelled IS\textit{6110} oligonucleotides was employed to detect MTBC (83) Since then the location of the IS\textit{6110} element has been used a powerful tool for the rapid fingerprinting of isolates of MTBC (84–86), including isolates of \textit{M. bovis} (87, 88). As there is variation in the copy numbers of IS\textit{6110} in different strains and because IS\textit{6110} does not induce \textit{in vivo} any major genomic rearrangements for approximately five to eight years (90, 91). IS\textit{6110} Restriction Fragment Length Polymorphism (RFLP) was the ‘gold standard’ typing method in strains with more than five copies before being replaced by the MIRU-VNTR technique (92–94) (see section 1.8.4). In a study reported by Warren et al in 2002, it was found that the rate of transposition occurs at a rate of 18.6\% for every 6.5 years i.e. 2.9\% per year, thereby altering the genotype of the organism (91) It has been suggested that there might be differences between different lineages, one change every 3 years in Euro-
American lineage and it might take ten times longer to observe one change in East Asian lineage (95).

Also, comparison of the number and length of fragments generated due to IS6110 insertion in isolates proved to be discriminatory (96). Sun et al., in 2004 showed that spoligotyping and Variable Number Tandem Repeats (VNTR) (see section 1.8.4) analysis can further support the resolution of IS6110-defined low copy number isolates. With these advantages put aside, IS6110 RFLP is however considered a laborious procedure involving steps from culturing the mycobacteria, extraction of DNA, restriction enzyme digestion, Southern Blotting and IS6110 probing.

1.8.1.1 Role of IS6110 in the biology of M. tuberculosis

The advantages that transposition of this element brings to the host either via gene disruption, gene excision or by enhanced gene expression suggests that they offer the selective advantage to strain fitness and may influence the genomic plasticity of these organisms. It can up-regulate downstream genes through an outward-directed promoter in its 3’ end, thus adding to the significance of this element. The ability to activate genes during infection suggests that IS6110 might have the potential to influence growth characteristics of different strains, and indicate another mechanism by which IS6110 can impact M. tuberculosis evolution (97). Gene disruption or excision leads to loss of gene expression and enhanced gene expression leads to over-expression of genes as seen in the Proline Glutamate-polymorphic GC rich Sequences (PE-PGRS gene -Rv1468c). This effect of supplying an outward-directed promoter-like sequence is also seen in ctpD
gene (a cation transporting ATPase) of a Beijing genotype useful for ctpD transcription (97,98). It is reported that the *M. tuberculosis* ctpD orthologue resists cellular redox stress by controlling the effects of cobalt and nickel ions on the bacterial cell (99). Several studies have shown that IS6110 insertion does not affect the replication machinery or the growth patterns but suggest that it definitely helps the organisms to adapt to the host and the environment. Transposition of IS elements varies from element to element within the genome (100) as an element situated within a transcriptionally active site or downstream from an external promoter will be more mobile than the ones located in the non–coding regions.

Transposition of IS6110 occurs at random, independent of IS preferential loci or hot-spots. However, sites of integration were not thought to be entirely random as different ‘hot-spot’ insertion sites have been identified previously that include the DR locus (70), the *ipl* locus (Fang and Forbes, 1997; Fang et al., 1999), *dnaA*- *dnaN* intergenic region (103), the phospholipase region (104) and the region between Rv1754c and Rv1762c (105). The property of transposition rendering a distinct signature to the genomes acts as a suitable genetic marker for studying the divergence of MTBC from its common ancestor. Mapping this element in the genomes of MTBC found that it does not favour particular nucleotide sequences (82,106) for its insertion. Insertion at the DR locus is unique as it is found in a majority of the *M. tuberculosis* complex strains especially in *M. bovis* in that it is its only site of insertion. It is also considered to be the primordial insertion site in the common ancestor before the two species diverged with further copies appearing due to the transposition of the element outside the DR region (70). In low copy strains,
lack of mobility of the IS element due to a low transcriptional activity at the site of insertion could play a role in the lack of IS6110 RFLP pattern diversity (107). It has been shown that insertion favours regions with low GC content as observed in the case of insertion into the PE/PPE gene family where PE (75% GC) is less favoured than PPE (64%) which was also complimented by another study showing a mini-transposon Tn5370 favouring lower GC content genomic regions in M. tuberculosis (98,106,108). When there are five or less copies of IS6110, this DNA fingerprinting technique is however not very reliable. So the use of polymorphic GC rich repetitive sequences (PGRSs) were highlighted along with DR spoligotyping and 16S rRNA sequences in tandem to identify strains containing few copies of IS6110, as in M. bovis or strains with no copy of IS6110 in their genomes (107,109) Dale et al. (107) suggests that there could be a low mobility rate of transposition in strains with low copy numbers.

1.8.2 PGRS- RFLP

A specific 3.4kb fragment of the PGRS cloned in plasmid pTBN12 was used as a probe for differentiation of unrelated strains of TB and identification of related strains using identical banding patterns (110). Interpretation of the banding patterns is more difficult as it does not contain as many variations as other markers like IS6110 and also requires a large amount of DNA as well as being laborious and time consuming and so this technique is no longer used (111).
1.8.3 Spoligotyping

All the members of the MTBC have the Direct Repeat (DR) locus in their genomes. The DR locus is a member of the CRISPR family (Clustered regularly interspaced short palindromic repeats) and is susceptible to polymorphic changes. This property was exploited for a genetic strain typing technique called spacer-oligo typing or spoligotyping (112,113). There are 43 unique non-repetitive spacer sequences (35-45bp in length) interspersed with direct repeat sequences (36bp in length) known as direct variable repeat (DVR) sequences. Due to the loss of single or consecutive DVR sequences caused either by homologous recombination between neighbouring or distant direct repeat sequences or by the loss of sequences caused by the excision of the IS6110 element, this method is useful for discriminating between strains. As the DVR sequences are numbered 1-43 and are well conserved between strains, spoligotyping is quite a quick and useful method (see figure 1-11 below) to interpret into either octal or hexadecimal codes (102,114).
Figure 1-11. Principle of spoligotyping and the processing of signals

(A) *M. tuberculosis* genome with well-conserved 36-bp direct repeats (DRs) which are interspersed by 35-43 bp of unique spacer sequences. Genetic diversity depends on the deletion of these spacer regions. The spacer regions are amplified by primers, and the presence of at least one spacer fragment shows a PCR positive reaction. On the membrane, 43 probes targeting each spacer are spotted, and a unique pattern of spoligotyping is visualized after hybridization with PCR product. (B) Signals of reference strain H37Rv. (C) A typical signal pattern of Beijing family *M. tuberculosis* strain. (D) To analyze signal patterns, the signals are converted to binary code of ‘on (1) and off (0)’. (E) The 43-digit binary code is converted to a 15-digit octal (i.e., base 8, having the digits 0-7) designation by a two-step process. First, the 43-digit binary code is divided into 14 sets of three digits (spacers 1 through 42) plus one additional digit (spacer 43). (F) Each 3-digit binary set is converted to its octal equivalent, with the final additional digit remaining as 1 or 0. The translation of binary numbers to octal numbers is done as follows: 000 = 0; 001 = 1; 010 = 2; 011 = 3; 100 = 4; 101 = 5; 110 = 6; and 111 = 7 (Taken from (115)).
Spoligotyping is a simple PCR based reverse hybridisation blotting technique where the DR locus is first amplified followed by the hybridisation with 43 synthetic spacer oligonucleotide probes covalently bound to the nylon membranes is visualised using a chemiluminescence system (112,115) for the presence of a ‘dark’ band (presence of a spacer) or ‘no’ band (absence of a spacer). There are at least nine spoligotype-defined families with specific hybridisation signatures (Haarlem, Beijing, LAM, CAS, EAI, S,T, X and AFRI families as seen in figure 1.12 which can be further divided into 36 subfamilies of the MTBC (113,116). This technique has a lower sensitivity than IS6110 RFLP but is useful in delineating strains with fewer than five IS6110 copies (117,118) and is the gold standard typing method for M. bovis strains lacking spacers 39-43.

Currently there is a database which can be accessed online (http://www.pasteur-guadeloupe.fr.8081) containing both octal and binary (as seen above in figure 1.12) spoligotype descriptions for strains of the MTBC species isolated globally known as SpolDB4 (113) (see figure 1-12). It contains 1939 shared-types (STs) representing a total of 39,295 clinical isolates originating from 122 countries. Due to the low resolution of the marker when used alone, a publicly available database known as SITVITWEB that utilises SpolDB4 data along with MIRU-VNTRs (see section 1.8.4 below) is in place for high-resolution epidemiological studies (119).
First column ST n°: Shared-type (ST) number of prototype pattern for the lineage/sub lineage. Second column: lineage/sub lineage name. Third column: Binary spoligo display with black-white squares for respectively hybridizing-non hybridizing spacers. Fourth column: Octal code (in red: defining octal rule). Fifth column: total absolute number of isolates of the subclass when variant ST spoligos are included (using SpolNet). Sixth column: same but expressed as percentage of total clustered isolates. * Total number and Frequency for these types are already included in their mother clade if known. Undesignated types are counted within the T1 ill-defined lineage. ** in red: octal rule defining the genotype (Taken from (113)).

<table>
<thead>
<tr>
<th>ST n°</th>
<th>lineage/sublineage</th>
<th>Binary spoligotype</th>
<th>Octal**</th>
<th>Total</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>161</td>
<td>AFB1</td>
<td></td>
<td>ST1</td>
<td>420</td>
<td>0.20</td>
</tr>
<tr>
<td>331</td>
<td>AFB2</td>
<td></td>
<td>ST2</td>
<td>279</td>
<td>0.13</td>
</tr>
<tr>
<td>438</td>
<td>AFB3</td>
<td></td>
<td>ST3</td>
<td>132</td>
<td>0.07</td>
</tr>
<tr>
<td>503</td>
<td>BOV1</td>
<td></td>
<td>ST4</td>
<td>30</td>
<td>0.01</td>
</tr>
<tr>
<td>582</td>
<td>BOV2</td>
<td></td>
<td>ST5</td>
<td>68</td>
<td>0.03</td>
</tr>
<tr>
<td>36</td>
<td>CA8-1</td>
<td></td>
<td></td>
<td>1076</td>
<td>0.19</td>
</tr>
<tr>
<td>21</td>
<td>CA8-2</td>
<td></td>
<td></td>
<td>1158</td>
<td>0.20</td>
</tr>
<tr>
<td>26</td>
<td>CA8-3</td>
<td></td>
<td></td>
<td>1250</td>
<td>0.22</td>
</tr>
<tr>
<td>288</td>
<td>CA8-4</td>
<td></td>
<td></td>
<td>29</td>
<td>0.01</td>
</tr>
<tr>
<td>452</td>
<td>CA8-5</td>
<td></td>
<td></td>
<td>130</td>
<td>0.02</td>
</tr>
<tr>
<td>193</td>
<td>MAN1</td>
<td></td>
<td></td>
<td>75</td>
<td>0.01</td>
</tr>
<tr>
<td>64</td>
<td>MAN2</td>
<td></td>
<td></td>
<td>160</td>
<td>0.03</td>
</tr>
<tr>
<td>1224</td>
<td>MAN3</td>
<td></td>
<td></td>
<td>28</td>
<td>0.00</td>
</tr>
<tr>
<td>1139</td>
<td>MN1</td>
<td></td>
<td></td>
<td>28</td>
<td>0.00</td>
</tr>
<tr>
<td>583</td>
<td>PHN1</td>
<td></td>
<td></td>
<td>38</td>
<td>0.01</td>
</tr>
<tr>
<td>637</td>
<td>PHN2</td>
<td></td>
<td></td>
<td>93</td>
<td>0.17</td>
</tr>
<tr>
<td>236</td>
<td>EAL-S</td>
<td></td>
<td>ST1</td>
<td>1014</td>
<td>2.02</td>
</tr>
<tr>
<td>20</td>
<td>LAM1</td>
<td></td>
<td>ST2</td>
<td>484</td>
<td>0.95</td>
</tr>
<tr>
<td>37</td>
<td>LAM2</td>
<td></td>
<td>ST3</td>
<td>411</td>
<td>0.81</td>
</tr>
<tr>
<td>57</td>
<td>LAM3</td>
<td></td>
<td>ST4</td>
<td>593</td>
<td>1.18</td>
</tr>
<tr>
<td>88</td>
<td>LAM4</td>
<td></td>
<td>ST5</td>
<td>149</td>
<td>0.29</td>
</tr>
<tr>
<td>93</td>
<td>LAM5</td>
<td></td>
<td>ST6</td>
<td>264</td>
<td>0.52</td>
</tr>
<tr>
<td>64</td>
<td>LAM6</td>
<td></td>
<td>ST1</td>
<td>172</td>
<td>0.28</td>
</tr>
<tr>
<td>239</td>
<td>LAM7</td>
<td></td>
<td>ST2</td>
<td>160</td>
<td>0.29</td>
</tr>
<tr>
<td>42</td>
<td>LAM8</td>
<td></td>
<td>ST3</td>
<td>22</td>
<td>0.04</td>
</tr>
<tr>
<td>81</td>
<td>LAM9</td>
<td></td>
<td>ST4</td>
<td>1590</td>
<td>2.97</td>
</tr>
<tr>
<td>299</td>
<td>LAM12</td>
<td></td>
<td>ST5</td>
<td>432</td>
<td>0.75</td>
</tr>
<tr>
<td>157</td>
<td>LAM12-Med1</td>
<td></td>
<td>ST6</td>
<td>240</td>
<td>0.45</td>
</tr>
<tr>
<td>53</td>
<td>T1</td>
<td></td>
<td>ST1</td>
<td>6412</td>
<td>12.56</td>
</tr>
<tr>
<td>299</td>
<td>T1-HLR1</td>
<td></td>
<td></td>
<td>63</td>
<td>0.01</td>
</tr>
<tr>
<td>34</td>
<td>S</td>
<td></td>
<td>ST2</td>
<td>648</td>
<td>1.28</td>
</tr>
<tr>
<td>55</td>
<td>T1</td>
<td></td>
<td>ST3</td>
<td>642</td>
<td>1.24</td>
</tr>
<tr>
<td>135</td>
<td>T2-Lupande</td>
<td></td>
<td>ST4</td>
<td>48</td>
<td>0.01</td>
</tr>
<tr>
<td>37</td>
<td>T3</td>
<td></td>
<td>ST5</td>
<td>277</td>
<td>0.53</td>
</tr>
<tr>
<td>149</td>
<td>T4-Eth</td>
<td></td>
<td>ST6</td>
<td>221</td>
<td>0.42</td>
</tr>
<tr>
<td>25</td>
<td>T4</td>
<td></td>
<td>ST1</td>
<td>168</td>
<td>0.32</td>
</tr>
<tr>
<td>44</td>
<td>T5</td>
<td></td>
<td>ST2</td>
<td>299</td>
<td>0.56</td>
</tr>
<tr>
<td>56</td>
<td>T5-Madrasi</td>
<td></td>
<td>ST3</td>
<td>103</td>
<td>0.20</td>
</tr>
<tr>
<td>254</td>
<td>T11-War</td>
<td></td>
<td>ST4</td>
<td>169</td>
<td>0.32</td>
</tr>
<tr>
<td>1737</td>
<td>T12</td>
<td></td>
<td>ST5</td>
<td>11</td>
<td>0.00</td>
</tr>
<tr>
<td>407</td>
<td>T12-KoRov</td>
<td></td>
<td>ST6</td>
<td>147</td>
<td>0.27</td>
</tr>
<tr>
<td>627</td>
<td>T12-OSA</td>
<td></td>
<td>ST1</td>
<td>19</td>
<td>0.03</td>
</tr>
<tr>
<td>159</td>
<td>X1</td>
<td></td>
<td>ST2</td>
<td>938</td>
<td>1.87</td>
</tr>
<tr>
<td>586</td>
<td>X2</td>
<td></td>
<td>ST3</td>
<td>1007</td>
<td>1.95</td>
</tr>
<tr>
<td>34</td>
<td>X2</td>
<td></td>
<td>ST4</td>
<td>503</td>
<td>0.99</td>
</tr>
<tr>
<td>92</td>
<td>X3</td>
<td></td>
<td>ST5</td>
<td>143</td>
<td>0.27</td>
</tr>
<tr>
<td>147</td>
<td>X3</td>
<td></td>
<td>ST6</td>
<td>244</td>
<td>0.47</td>
</tr>
<tr>
<td>81</td>
<td>X3-variantI</td>
<td></td>
<td></td>
<td>156</td>
<td>0.30</td>
</tr>
<tr>
<td>4</td>
<td>LAM30</td>
<td></td>
<td></td>
<td>115</td>
<td>0.22</td>
</tr>
<tr>
<td>72</td>
<td>X3-variantII</td>
<td></td>
<td></td>
<td>114</td>
<td>0.22</td>
</tr>
<tr>
<td>219</td>
<td>X3-variantIII</td>
<td></td>
<td></td>
<td>104</td>
<td>0.20</td>
</tr>
<tr>
<td>48</td>
<td>X4-variantI</td>
<td></td>
<td></td>
<td>102</td>
<td>0.20</td>
</tr>
<tr>
<td>46</td>
<td>X4-variantII</td>
<td></td>
<td></td>
<td>95</td>
<td>0.20</td>
</tr>
</tbody>
</table>

**Figure 1-12 Pictorial representation of SpolDB4 classification analysis of different TB lineages**

First column ST n°: Shared-type (ST) number of prototype pattern for the lineage/sub lineage. Second column: lineage/sub lineage name. Third column: Binary spoligo display with black-white squares for respectively hybridizing-non hybridizing spacers. Fourth column: Octal code (in red: defining octal rule). Fifth column: total absolute number of isolates of the subclass when variant ST spoligos are included (using SpolNet). Sixth column: same but expressed as percentage of total clustered isolates. * Total number and Frequency for these types are already included in their mother clade if known. Undesignated types are counted within the T1 ill-defined lineage. ** in red: octal rule defining the genotype (Taken from (113)).
1.8.4 MIRU-VNTR

Mycobacterial interspersed repetitive units (MIRUs) in *M. tuberculosis* are grouped under minisatellites which are short highly repetitive DNA sequences and are 40-100 bp long. They are suggested to have a role in chromosome structure and rearrangement, tandem duplications, differential translation, transcriptional termination and antigenic variation (120). MIRUs are highly polymorphic at tandem repeat loci, and are useful in DNA typing studies in both prokaryotes and eukaryotes. They can also be referred to as variable number tandem repeats (VNTRs) and exact tandem repeats (ETRs). ETRs are 53-79bp long tandem repeats and another repeated region called Major Polymorphic Tandem Repeats (MPTRs) were first described in 1998. Some MIRU loci are positioned in the polycistronic operons and variability in copy number can affect the expression of the flanking genes (121).

Mycobacterial Interspersed Repetitive Units (MIRU) first described by Supply et al., in 2000 and the Exact Tandem Repeats (ETR) both comprise MIRU-VNTR, a typing technique useful in epidemiology, population genetics and phylogenetic studies (122). The technique involves the PCR amplification of the entire tandem repeat loci using primers in the DNA sequences flanking the repeats. The number of repeats or alleles is calculated from the amplicon sized by electrophoresis as the length of the repeats and the position of the primers are known. Twelve MIRU loci (2, 4, 10, 16, 20, 23, 24, 26, 27, 31, 39, 40) and five ETR loci have been used previously for this genetic fingerprinting technique. This method was widely adopted in the UK. There is some overlap of the nomenclature of these repeat schemes, for example MIRU 4 and MIRU 31 are the same loci as ETR-D and ETR-E respectively.
Supply et al., li 2006 proposed a 24 loci typing technique comprising of 12 MIRU loci, 3 ETR loci, six ‘Mtubs’ and three ‘QUB’loci. This was the gold standard typing method used in the UK until recently but now WGS has become the method of choice in the UK. A fifteen loci subset system with the highest evolutionary rates was proposed as the best system for first-line analysis of outbreak samples and the 24 loci subset for the phylogenetic studies which need higher resolution. The twelve loci subset, analysing the most variable regions, are also used for initial investigations along with data from other techniques like spoligotyping and IS6110 fingerprinting (123). The following table 1-4 shows the widely used MIRU-VNTR schemes worldwide.

Table 1-4 Composition of the different MIRU-VNTR sets widely used

(Taken from http://www.miru-vntrplus.org/MIRU/miruinfo.faces)

<table>
<thead>
<tr>
<th>Loci</th>
<th>Alias 1</th>
<th>Alias 2</th>
<th>24</th>
<th>15</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>154</td>
<td>MIRU02</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>424</td>
<td>Mtub04</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>577</td>
<td>ETR-C</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>580</td>
<td>MIRU04</td>
<td>ETR-D</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>802</td>
<td>MIRU40</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>960</td>
<td>MIRU10</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>1644</td>
<td>MIRU16</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>1955</td>
<td>Mtub21</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2059</td>
<td>MIRU20</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>2163b</td>
<td>QUB11b</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2165</td>
<td>ETR-A</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2347</td>
<td>Mtub29</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2401</td>
<td>Mtub30</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2461</td>
<td>ETR-B</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2531</td>
<td>MIRU23</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>2687</td>
<td>MIRU24</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>2996</td>
<td>MIRU26</td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>3007</td>
<td>MIRU27</td>
<td>QUB5</td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>3171</td>
<td>Mtub34</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3192</td>
<td>MIRU31</td>
<td>ETR-E</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>3690</td>
<td>Mtub39</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4052</td>
<td>QUB26</td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>4156</td>
<td>QUB4156</td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>4348</td>
<td>MIRU39</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
</tr>
</tbody>
</table>
1.8.5 AFLP and FAFLP

1.8.5.1 AFLP

Amplified Fragment Length Polymorphism (AFLP) was designed as an identification and typing method developed in the nineties by the biotechnology company called Keygene (124,125). It is a selective restriction fragment amplification technique based on the ligation of adapters to genomic restriction fragments and PCR amplification with adapter specific primers. A small amount of genomic DNA is digested with two restriction enzymes and then ligated with double-stranded oligonucleotide adapters designed specifically with a point mutation so that the initial restriction is not restored after ligation. With the aid of stringent PCR conditions and adapter-specific primers at the 3' ends, one of the three nucleotides is extended from the unknown restricted genome fragment. The primers initially were radioactive labelled for easy identification of the fragments in polyacrylamide gels which were later replaced by fluorescently labelled primers for visualisation in automated platforms (125–127).

Since then, AFLP has been used as a diagnostic tool in plants and animals in the field of genetic mapping, phylogenetic studies, microbial typing and for diagnostic purposes (124,128,129).

1.8.5.2 FAFLP

Fluorescent Amplified Fragment Length Polymorphism (FAFLP) exploits the use of fluorescently labelled primers and the resulting genomic restriction fragments are visualised using an automated platform. Since the restricted fragments with base substitutions are visualised using a platform such as the
ABI Genetic Analyser (Life Technologies, UK), this technique can give a snapshot of the insertion variation on a genomic level. It has been shown that a single fragment difference (± 1 bp) in the FAFLP profile signifies a new strain (130). FAFLP should, like all other techniques, be considered in tandem with epidemiological data if available. FAFLP has been successfully applied to several organisms and some of them have been shown below in the form of a table (see table 1-5). It has also been used in the identification of plant species with 98% accuracy (131).
Table 1-5 List of some common microorganisms studied using FAFLP analysis

<table>
<thead>
<tr>
<th>Microorganism</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Campylobacter jejuni</em></td>
<td>(132)</td>
</tr>
<tr>
<td><em>Escherichia coli</em> K12 MG1655</td>
<td>(127)</td>
</tr>
<tr>
<td><em>Escherichia coli</em> O157</td>
<td>(127,133)</td>
</tr>
<tr>
<td><em>Listeria monocytogenes</em></td>
<td>(134)</td>
</tr>
<tr>
<td><em>Mycobacterium tuberculosis</em></td>
<td>(130,135)</td>
</tr>
<tr>
<td><em>Neisseria gonorrhoeae</em></td>
<td>(136)</td>
</tr>
<tr>
<td><em>Neisseria meningitides</em></td>
<td>(137)</td>
</tr>
<tr>
<td><em>Streptococcus pyogenes</em> serotype M1</td>
<td>(126)</td>
</tr>
<tr>
<td>Bacterial/Fungal Name</td>
<td>Reference</td>
</tr>
<tr>
<td>---------------------------------------------</td>
<td>-----------</td>
</tr>
<tr>
<td><em>Staphylococcus aureus</em></td>
<td>(138)</td>
</tr>
<tr>
<td><em>Bacillus thuringiensis</em></td>
<td>(139)</td>
</tr>
<tr>
<td><em>Candida tropicalis</em></td>
<td>(140)</td>
</tr>
<tr>
<td><em>Acanthamoeba Sp.</em></td>
<td>(141)</td>
</tr>
<tr>
<td><em>Salmonella enteritidis</em></td>
<td>(142)</td>
</tr>
<tr>
<td><em>Leptospira borgpetersenii serovar Arborea</em></td>
<td>(143)</td>
</tr>
<tr>
<td><em>Pseudomonas aeruginosa</em></td>
<td>(144)</td>
</tr>
<tr>
<td><em>Legionella pneumophila</em> serogroup 1</td>
<td>(145)</td>
</tr>
</tbody>
</table>
1.8.5.3 FAFLP for TB

1.8.5.3.1 IS6110 FAFLP
IS6110 RFLP is a time consuming and low-throughput technique. To overcome these drawbacks and utilise the superior resolution property of IS6110, IS6110 FAFLP was developed. FAFLP derived sequence data is congruent with the IS6110 typing of *Mycobacterium tuberculosis* but it should be used along with other techniques like spoligotyping and MIRU-VNTR to supplement epidemiological data (Goulding, Stanley, et al. 2000; Thorne et al. 2007). In a study by Thorne et al, in 2007, it was noticed that the FAFLP results of 57 (97%) of the 59 clustered isolates were congruent with the RFLP results. Use of coloured fluorescent primers can differentiate between recent transmission of strains and epidemiologically unrelated but genetically related strains (147). IS6110 FAFLP has been used recently to identify common fragments in MTBC that likely evolved from common ancestors and were thereby able to differentiate the strains phylogenetically into different TB lineages (135). This study also indicated that IS6110 FAFLP contains a strong phylogenetic signal in modern TB lineages by assigning common fragments to their respective spoligotypes/Principal Genetic Groups (PGGs).

1.8.5.4 DNA Sequencing
From Sanger sequencing to Next generation sequencing technologies, there are different sequencing methodologies that can be used to generate data for molecular epidemiological studies of tuberculosis (148). Due to the high cost and low quality of the sequences both in the first 15-40bp and after 700bp of sequenced product, Sanger sequencing is being replaced by Next Generation Sequencing (NGS), also known as Whole Genome Sequencing.
(WGS). Over the last ten years, platforms including, Illumina, Roche, ABI SOLID, ION PGM, Heliscope, Pacific Biosciences Single Molecule Real Time (SMRT) and the Nanopore MinION (149) have been used. WGS is likely to become the gold standard approach in the future but specialist software to analyse the data and a skilled workforce are still required to interpret the results.

1.9 Global TB lineages

New species of mycobacteria have evolved over the years by many adaptive changes to different niches (6). This has led to the reduced function or inactivation of certain genes. A classic example is *M. leprae* which has transformed to become an obligate intracellular pathogen by losing many genes involved in metabolic and respiratory pathways (150). Genes can also be acquired by some species like *M. abscessus* which is one of the organisms responsible for causing infection in those suffering from cystic fibrosis. This organism was thought to originally be a soil saprophyte, and has acquired essential genes to survive phagocytosis enabling intracellularisation in host organisms. Recently it has been shown that *M. abscessus* has acquired resistance to arsenic and mercury which are typically found in resistance plasmids of environmental organisms like *M. marinum* through horizontal gene flux (151).

All the members of the MTBC share identical 16S rRNA sequences and more than 99% similarity at the genomic level and yet are phenotypically different causing different pathologies in different host species. *M. canetti* is unique in this group in that it does not share homogeneity with the MTBC in some
house-keeping genes, Insertion Sequence 1081 copy number, colony morphology and lipid content of the cell. Brosch et al., in 2002 showed that *M. canetti* might have diverged directly from the common ancestor of MTBC some 30,000 to 40,000 years ago as it has Regions of Difference (RD), H37Rv related deletions- Rvd and TB specific deletions-TbD1 that are absent in the other six members (114) (fig1-6). Thus genetic polymorphism can occur in MTBC either due to a single nucleotide mutation in specific gene coding regions, or by the presence or absence of a gene or region, thereby altering the gene content of the microorganisms (fig. 1.13). Linkage Disequilibrium (LD) studies now support the theory that *M. tuberculosis* has evolved clonally (152,153). LD refers to the non-random association of alleles at two different loci, which are not independent of one another (154).
Figure 1-13 Novel evolutionary scenario of MTBC proposed by Brosch et al in 2002 based on the regions of difference (RD) and sequence polymorphisms within five genes.

Blue arrows indicate strains that are characterised as group one organisms, the green arrow is group two and the red arrow group three, as defined by \( \text{katG}^{463} \) and \( \text{gyrA}^{95} \) codon sequence polymorphisms as described by Sreevatsan et al., in 1997 (153) as shown in figure 1.4 (Taken from (114)).
Several clonal species of mycobacteria are not identifiable when they are isolated on solid media as they are thought to have descended from a common ancestor and MTBC fall under this category as previously discussed.

1.9.1 Major Genetic Groups of MTBC and its global distribution

Major Genetic Groups (MGG) classification is based on single nucleotide mutations or polymorphism (SNP). Generally, there are two types of SNPs synonymous and non-synonymous. The former are important to study evolution as they are not prone to divergence whereas the latter refers to translational changes that are not evolutionarily informative (155). Sreevatsan et al., in 1997 identified two SNPs; one in the catalase peroxidase (katG codon 463) and one gyrase A (gyrA codon 95) encoding genes and classified MTBC into three Major Genetic Groups (MGGs) or Principal Genetic Groups (PGGs). Here they showed that there are three MGGs- MGG1, MGG2 and MGG3 where MGG3 is derived from MGG2 and MGG2 from MGG1. An M. tuberculosis precursor or common ancestor would have given rise to MGG1 (153) (fig 1-14).
Several other studies by Gutacker et al., 2006, Baker et al., 2004, Rad et al., 2003 have all classified the MTBC into different lineages either based on SNPs or the presence or absence of TB specific deletion regions (TbD1) (155–157). Brosch et al. in 2002 showed that the MTBC diverged from the most recent common ancestor over thousands of years by losing regions of deletion (RD) or due to large sequence polymorphisms (LSPs) in housekeeping genes. It is important to note that the TbD1 deletion is common to all ‘modern’ TB strains which include the Haarlem, Beijing and CAS spoligotype strains and that this deletion is irreversible because of the lack of DNA recombination between strains of the MTBC. The TbD1 deletion is therefore absent from ancient strains such as the East African-Indian (EAI) strains. The division of modern and ancient families (where DNA markers characterised, not the genomes themselves, are thought to be similar to those found in
modern or ancient strains) is based on spoligotyping, identifying the deletions or polymorphism in the DR locus (DVR24) in M. tuberculosis genomes (112). Different spoligotypes or families derived by adopting this technique, based on detection of the presence or absence of 43 unique sequences in the Direct Repeat region, are given below in the table 1-6.

Table 1-6 Table showing the distribution of M. tuberculosis complex (MTBC) strains with respect to the Principal Genetic Groups (PGGs) and the TB spoligotypes families / lineages

<table>
<thead>
<tr>
<th>Principal Genetic Groups</th>
<th>Mycobacteria</th>
<th>TB Spoligotype family / lineage</th>
</tr>
</thead>
<tbody>
<tr>
<td>PGG1</td>
<td>M. microti, M. africanum, M. tuberculosis and M. bovis</td>
<td>Central Asian Strain (CAS) Beijing</td>
</tr>
<tr>
<td>PGG2</td>
<td>M. tuberculosis</td>
<td>S X T-Uganda Haarlem Latin American Mediterranean (LAM)</td>
</tr>
<tr>
<td>PGG3</td>
<td>M. tuberculosis</td>
<td>T</td>
</tr>
</tbody>
</table>
Due to the inclusion of various sub-species of MTBC (4) including the classification of TB genetic lineages according to the geographical migration (158), the evolutionary scenario originally proposed by Sreevatsan et al. in 1997 and Brosch et al. in 2002 has been combined as shown in fig 1.15 below.
Figure 1-15 Schematic representation of the proposed phylogenetic relationships between members of MTBC.

The phylogeny is based on presence or absence of Region of Difference (RD) (114), SNPs (153) and Large Sequence Polymorphism (LSP) (159). (Taken from (4)).
1.9.2 Global distribution of TB genetic lineages including Co-Evolution of TB and Humans

It has been suggested that *M.tuberculosis* might have co-evolved with the human host and have followed the ‘out-of-and-back-to-Africa’ evolutionary scenario (159–161). In 2006, Gagneux and his colleagues addressed the co-evolutionary scenario of both human and TB populations for the first time and also proposed six phylogeographical lineages (fig 1-16) (158).
Figure 1-16 The phylogeographical distribution of *M. tuberculosis* and its lineages

(a) 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 1 of 80 countries represented in the global strain collection. The colours of the dots relate to the six main lineages defined in (a) and indicate the dominant lineage(s) in the respective countries (Taken from (158))
1.9.3 Molecular Clocks in the evolution of *M. tuberculosis*

Using different molecular markers (SNPs, RDs, TbD1 and spoligotypes), a timeline of evolutionary events can be generated that strongly suggests that the currently circulating MTBC isolates might have originated from a common ancestor as shown in figure 1-17 below (162) as discussed in section 1.11 and shown in fig 1-13 by Brosch et al (114).

![Figure 1-17 Timeline of evolutionary events in *Mycobacterium tuberculosis*](image)

Over recent years, data have accumulated suggesting that the different MTBC lineages might have adapted to different human populations and thus showing predominance in certain regions like CAS in Northern India (163) and Beijing in Tibet and China (164). Not only does MTBC exhibit a global
biogeographic population structure, but the associations between the particular MTBC lineages and human populations are maintained in cosmopolitan settings where human populations and their associated MTBC strains experience at least some degree of intermingling (161, 165, 166). Sequencing six housekeeping genes (*katG, gyrB, gyrA, rpoB, hsp65* and *sodA*) and the complete 16SrRNA gene in smooth tubercle bacilli suggests that *M. tuberculosis* might have descended 3 million years ago (167). Also, ancient DNA studies (160) suggest that MTBC predates the arrival of the human population (fig 1-18), thus showing the adaptability of this ancient microorganism.

![Evolution of Mycobacterium tuberculosis complex along with human population](image)

Figure 1-18 Evolution of *Mycobacterium tuberculosis* complex along with human population

(Taken from(160)).
1.9.4 Importance of genetic polymorphism

Different markers have been used to study the phylogenetic relationships between the common ancestor and the currently circulating TB genetic lineages. As IS\textit{6110} is one of the earliest seen polymorphic changes in terms of evolution of TB, they could theoretically play an important role in the evolution of different TB genetic lineages. IS\textit{6110} insertions occur mainly in multi-gene families such as the PPE gene family and they could be beneficial for the organism to evade immunity (74) and cause disease (168). In addition, several studies have suggested that drug resistance, formation of LSPs and high frequency of SNPs were noticed in regions flanked by IS\textit{6110} (169). Thus, IS\textit{6110} transposition could be essential in the evolution of the \textit{M.tuberculosis} genome.

1.10 Molecular Epidemiology of TB in resource poor settings

Among the estimated 9.4 million new TB cases reported in 2008, 55% and 30% of the cases were reported in resource poor countries in Asia and Africa respectively (170). Depending on the availability of instruments and reagents, the techniques used to study the molecular epidemiology of TB are different in different countries. Epidemiological studies are however limited when compared to developed countries due to financial constraints in these areas despite the fact that tuberculosis is endemic.

1.11 Currently used epidemiological tools, its suitability and constraints

The tools used in a resource-limited country mainly depend on the country's ability to afford equipment required for different technologies, their
maintenance and the availability of the skilled force to perform the techniques. Resource poor countries may need improved access to diagnosis and care (171). Without an effective surveillance strategy, TB control measures could be difficult to be monitored. So, routine surveillance methods will contribute in assessing drug resistance strains and monitoring. Within their limits these settings continue to contribute to the understanding of the molecular epidemiology of TB using the tools available to them including the most commonly used techniques IS6110 RFLP, MIRU-VNTR and spoligotyping.

1.12 Nepal TB epidemiology

Though the global TB burden has been falling globally in the last few years due to the introduction of the STOP TB programme by the WHO (1), the incidence of HIV-associated TB is high in Myanmar, Thailand, India and Indonesia. Nepal used to be one of the five countries listed in SE Asia region with a high burden of TB. But the new global TB report from WHO (172) has excluded Nepal from the list of 30 high burden countries due to the effective STOP TB policy since 2006. But tuberculosis still remains a public health problem and ranked as the sixth leading cause of death among the top 20 causes of death in Nepal according to the National Tuberculosis programme survey carried out in 2014 (173).

1.12.1 TB in Nepal

Nepal has 147,181 square kilometres of land mass area with a population of 29 million people and is interlocked between the Indian subcontinent on three sides and China’s Tibet to the North. Nepal lies on the major trade route
between India and China (see figure 1-19) and has an 1200-km long controlled border with China in the north and an 1700-km long open border with India in the east, west and the south (174). It has been noted that migration across these international borders take place for reasons like work, study, tourism, religious purposes like pilgrimages and cultural exchanges like marriages. Kathmandu, the capital city of Nepal is highly populated with a high number of TB cases according to the Nepal's National Tuberculosis programme.

Figure 1-19 A map of Nepal showing the topography of Nepal interlocked between India (in the south, east and west of Nepal) and China (in the north)

(Taken from http://www.echoway.org/graphic/oupartir/cartespays/nepal.gif)

In 2014, the WHO estimated that 4600 (2100-7500) people died from TB in Nepal. Even though MDR-TB levels are low (2.2%) in new cases (high in
retreatment cases-15.4%), Drug-resistant TB (DR-TB) is still a major public health concern. The epidemiological indicators of the TB burden in Nepal according to the report published in 2014 by National Tuberculosis Programme is summarised below in the table 1-7.

**Table 1-7 Estimated TB burden in Nepal in 2014**
(Adapted from NTP, Nepal 2014 report)

<table>
<thead>
<tr>
<th>Epidemiology Indicators</th>
<th>Estimated patient cases (Range)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prevalence</td>
<td>59 (27-100) x 1000</td>
</tr>
<tr>
<td>Prevalence rate</td>
<td>211 (99-365)</td>
</tr>
<tr>
<td>Incidence</td>
<td>43 (39-49) x 1000</td>
</tr>
<tr>
<td>Incidence rate</td>
<td>156 (139-178)</td>
</tr>
<tr>
<td>Cases notified to NTP</td>
<td>37025</td>
</tr>
<tr>
<td>Annual case notification rate</td>
<td>136 / 100,000</td>
</tr>
<tr>
<td>HIV negative / TB positive cases</td>
<td>4.6 (2.1-7.5) x1000</td>
</tr>
<tr>
<td>HIV positive / TB positive cases</td>
<td>0.32 (0.17- 0.51) x1000</td>
</tr>
</tbody>
</table>

**1.12.2 Current diagnostic procedures in Nepal**

Sputum smear microscopy is used as the main method to detect TB and monitor treatment responses in patients. Culturing the bacteria and
phenotypic Drug Susceptibility Tests (DSTs) are the two principal procedures carried out to monitor new and relapsed cases within the 581 microscopy centres, two solid culture laboratory including capacity for first line drug (FLD) susceptibility testing (DST), first line probe assay (LPA), one liquid culture and one second line drug (SLD) DST facility and 26 GeneXpert centres spread all over this mountainous country. This is in addition to two other culture and DST facilities (NTC and GENETUP) functioning in Kathmandu valley.

1.12.3 Constraints in diagnosis of TB in Nepal

According to the NTCP’s annual report, the challenges faced by the authorities to control TB in Nepal are mainly due to the increase in MDR and XDR cases, low case notification by different districts, insufficient supervision, monitoring and evaluation of TB cases and lack of skilled labour (173). In a recent survey in 2014, it has been found that nearly 9.3% of new patients are resistant to at least one drug. Also, resistance to fluoroquinolones (26.4%) has led to a heavy burden of pre-XDR and XDR-TB among MDR-TB cases (8% of the cases were found to be XDR among MDR patients). It has been speculated that it might be due to the fact that fluoroquinolones could be purchased by the patients over the counter easily and also its unregulated use. Case Notification Rate (CNR) in 2015 was 123/100,000 a drop of 11% compared to the previous survey in 2010 (173). Among different regions, there has been a drop in the CNR in the Eastern and the central regions of Nepal. So, an efficient, cost-effective and rapid epidemiological tool that can be useful in these settings was needed and thus
this research project aimed to develop a genomic tool box that can be easily used in these settings along with other techniques.
1.13 Aims and Objectives

The aim of my PhD thesis is to develop a genomic tool box to aid epidemiological investigations using a novel IS6110 FAPLP sampling method for application in resource poor settings.

1.14 Hypothesis

Characterisation of IS6110 insertion sites in TB genomes of strains from poorly studied regions will give insight into the geographical distribution of disease and their relationships with different lineages thereby showing that IS6110 is not only a typing tool but also an effective evolutionary marker.

The objectives of this thesis are as follows:

- **Chapter 1**: To review the TB literature in detail focusing on the suitability of the technique and the sample population.
- **Chapter 2**: To describe IS6110 FAPLP methodology in addition to the reagents needed for the technique.
- **Chapter 3**: To optimise and develop the IS6110 FAPLP methodology using H37Rv and Nepal strains.
- **Chapter 4**: To map the IS6110 insertion sites in the *M. tuberculosis* H37Rv genome and to rapidly detect TB genetic lineages using IS6110 FAPLP derived PCR.
- **Chapter 5**: To classify the TB samples from Nepal into their different TB genetic lineages.
- **Chapter 6**: To understand the Rifampicin resistance status in the TB samples from Nepal.
Chapter 7: To contextualise that IS6110 FAPLP as a genomic mapping tool and discuss its suitability as a method to understand the phylogenetic relationships between TB genetic lineages and further indicate that IS6110 is a valuable tool to study the epidemiology and evolution of TB.
Chapter 2  General Materials and Methods
2.1 Bacterial samples

2.1.1 Sample collection centres

Sputum samples from consecutive new TB patients were collected over one year between 2007 and 2008 and cultured alongside routine diagnostic testing from two Nepalese tuberculosis reference centres located in the Kathmandu valley: the National Tuberculosis Centre (NTC) and the German Nepal Tuberculosis Project (GENETUP) (fig 2.1).

Figure 2-1 Geographical location of the two Nepal TB treatment centres, GENETUP and NTC, where samples were collected
Bacterial genomic DNA from the strains isolated from the above samples were extracted at the National Mycobacterial Research Laboratory (NMRL) in Nepal using Cetyltrimethyl Ammonium Bromide (CTAB)(175).

2.1.2 Ethics

Nepal Health Research Council (NHRC), the Ethics Approval Committee approved the research projects ongoing at MRL in general between 2006 and 2008 when these samples were collected. During this time, the specific procedures including consent were approved by the Head of MRL, MRL Lab Manager, the Anandaban Medical Director, faculty of Tribhuvan University and Directors of both Tuberculosis centres where samples were collected. As sputum samples were classified as non-invasive samples during collection, written consents were not emphasised and only verbal consents were taken but not documented from the patients who attended the hospitals primarily for TB treatment. The anonymized DNA samples were then transported to Public Health England (PHE), Colindale in 2008 and this particular study started in 2011 and was solely conducted in Public Health England, Colindale.

2.1.3 IS6110 Fluorescent Amplified Fragment Length Polymorphism (FAFLP)

This method published by Thorne et al. (146) was optimised and developed as discussed in chapter 3. The method comprised of four major steps: Digestion of genomic DNA, Ligation of adapters, PCR and Capillary Electrophoresis (Figure 2-2). All reagents used in this study were risk assessed locally as the Genomic Services Development Unit (GSDU) 2D01
lab where this research was carried out already uses the FAFLP procedure for the day-to-day service work.

Figure 2-2 Schematic representation of the IS6110 FAFLP methodology.
2.1.4 Genomic DNA Restriction Endonuclease digestion

Reagents:

1. *MseI* (New England Biolabs (NEB), UK)

2. *TaqI* (NEB, UK)

3. 1x *MseI* buffer II (NEB)

4. Bovine Serum Albumin (BSA) , (NEB)

5. Ribonuclease A (RNase A) (Sigma Aldrich, UK)

6. 1M proline (Sigma Aldrich, UK)

7. 2mM betaine (Sigma Aldrich, UK)


2.1.4.1 Protocol:

Five µl of 200-500 ng of bacterial genomic DNA quantified using Qubit (Thermofisher, UK) was digested with 5 U of *MseI* (New England Biolabs, UK) in a total reaction volume of 20µl containing 1x *MseI* buffer, 0.1mg/ml of bovine serum albumin, 0.5mg/ml of RNase A, 1M of L-proline and 2mM of betaine at 37°C for 2 h to release the DNA ready for a final digestion with 10U of *TaqI* (New England Biolabs, UK) at 65°C for 3 h and the reaction was inactivated for endonucleases at 80°C for 10 min.
2.1.5 Ligation of digested DNA to double stranded adapters

2.1.5.1 Reagents:

1. T₄ DNA Ligase (NEB, UK)

2. 2 x T₄ DNA Ligase buffer (NEB, UK)

3. Double stranded Taq I adaptor (Eurofins, UK)

2.1.5.2 Protocol:

Double stranded oligonucleotide adaptors specific to TaqI site (5’-TACTCAGGACTGGC) were ligated to the double- digested DNA in a total reaction volume of 25µl containing 40U of T₄ DNA ligase (New England Biolabs, UK), 0.2µM of double stranded TaqI adaptor and 2x T₄ ligase buffer (NEB, UK). The reaction mixture was then incubated at 12°C overnight followed by heating at 65°C for 10 minutes to inactivate the ligase and later stored at -20°C.

2.1.6 PCR amplification of fragments

2.1.6.1 Reagents:

1. 10 x PCR reaction buffer (Thermofisher, UK)

2. 50mM MgCl₂ (Thermofisher, UK)

3. 10mM dNTPs (Thermofisher, UK)

4. 5µM TaqI primer (Eurofins, UK)
2.1.6.2 Preparation of 10µM of 4-dye labelled TaqI primer

10µM of each selective TaqI primer labelled with 6-FAM (blue), PET (red), VIC (green) and NED (Black) fluorescent dyes at the 5’ end and a selective base either A or C or G or T nucleotide respectively according to the dye listed above were made from 100µM stocks with molecular biology grade water. 10µl of each 10µM primer i.e. 40µl in total were then mixed with 60ul of water to make up 10µM TaqI primer stock.

5. 5uM IS6110 specific primer (Eurofins, UK)

6. Molecular Biology Grade water (Fisher Scientific, UK)

7. Recombinant Taq DNA Polymerase (Thermofisher, UK)

2.1.6.3 Protocol:

Four labelled adapter-specific forward primers, each with a single unique selective nucleotide (A / C / T/G) at the 3’ end which extends into the unknown genomic DNA sequence of the fragment, together with an IS6110 specific reverse primer, were used to amplify the fragments. In a total volume of 20 µl, 1µl of the adaptor ligated DNA was added to the reaction containing 1X reaction buffer (Invitrogen, UK), 1.5 mM MgCl₂ (Invitrogen, UK), 0.2 mM dNTPs (Invitrogen, UK), 1 µM of labelled selective TaqI forward primer (5’-CGATGAGTCCTGACCGA/C/T/G- (see table 2-1), 1 µM of IS6110 reverse primer (5’- CTGACATGACCCCATCCTTT) and 1U of recombinant Taq polymerase (Invitrogen, UK). The following PCR conditions were carried out
in a Veriti thermocycler (Applied Biosystems, UK): 94°C for 15 min followed by 35 cycles of 94°C for 20 s, 66°C for 30 s and 72°C for 2 min with the 66°C annealing temperature reducing by 1°C every cycle for nine cycles and the last 25 cycles at 56°C. Finally, an extension of 72°C for 60 min was carried out before further manipulations.

Table 2-1 List of *Taq*I fluorescently labelled selective primers

<table>
<thead>
<tr>
<th>Fluorescent labelled selective primer</th>
<th>Primer Sequence (selective base (135,147))</th>
</tr>
</thead>
<tbody>
<tr>
<td>1) <em>TaqI</em>- A (labelled with 6-FAM dye)</td>
<td>5'-CGATGAGTCCTGACCGA(A)-3'</td>
</tr>
<tr>
<td>2) <em>TaqI</em>- C (labelled with PET dye)</td>
<td>5'-CGATGAGTCCTGACCGA(C)-3'</td>
</tr>
<tr>
<td>3) <em>TaqI</em>- T (labelled with NED dye)</td>
<td>5'-CGATGAGTCCTGACCGA(T)-3'</td>
</tr>
<tr>
<td>4) <em>TaqI</em>- G (labelled with VIC dye)</td>
<td>5'-CGATGAGTCCTGACCGA(G)-3'</td>
</tr>
</tbody>
</table>

2.1.7 Capillary Electrophoresis of fragments using ABI3730xl Genetic Analyser

2.1.7.1 Reagents:

1. Hi_Di formamide (Applied Biosystems, UK)

2. Genescan Liz600 sizing standard (Applied Biosystems, UK)
2.1.7.2 Protocol:

One microlitre of the IS6110 FAFLP PCR reaction from the PCR step was mixed with 10µl of Hi-Di formamide (Applied Biosystems, UK) and 0.5 µl of the Genescan Liz600 sizing standard (Applied Biosystems, UK). The reaction was denatured at 95° C for 5 min before running the reaction on the ABI genetic analyser 3730xl. The run module settings are given in the table 2.2 below:

<table>
<thead>
<tr>
<th>Run parameters</th>
<th>Actual setting</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oven temperature</td>
<td>63C</td>
<td>18-70°C</td>
</tr>
<tr>
<td>Buffer temperature</td>
<td>35°C</td>
<td>30-35°C</td>
</tr>
<tr>
<td>Prerun voltage</td>
<td>15.0 kV</td>
<td>0-15 kV</td>
</tr>
<tr>
<td>Prerun Time</td>
<td>180 sec</td>
<td>1-1800 sec</td>
</tr>
<tr>
<td>Injection voltage</td>
<td>1.6 kV</td>
<td>0-15 kV</td>
</tr>
<tr>
<td>Injection time</td>
<td>15 sec</td>
<td>1-90 sec</td>
</tr>
<tr>
<td>First readout time</td>
<td>200 ms</td>
<td>100-16000 ms</td>
</tr>
<tr>
<td>Second readout time</td>
<td>200 ms</td>
<td>100-16000 ms</td>
</tr>
<tr>
<td>Run Voltage</td>
<td>15 kV</td>
<td>0-15 kV</td>
</tr>
<tr>
<td>Voltage no. of steps</td>
<td>10</td>
<td>0-100 steps</td>
</tr>
<tr>
<td>Parameter</td>
<td>Value 1</td>
<td>Value 2</td>
</tr>
<tr>
<td>-------------------------</td>
<td>-----------</td>
<td>-------------</td>
</tr>
<tr>
<td>Voltage step interval</td>
<td>20 sec</td>
<td>0-180 sec</td>
</tr>
<tr>
<td>Voltage tolerance</td>
<td>0.6 kV</td>
<td>0-6.0 kV</td>
</tr>
<tr>
<td>Current stability</td>
<td>30 µA</td>
<td>0-2000 µA</td>
</tr>
<tr>
<td>Ramp Delay</td>
<td>1 sec</td>
<td>1-1800 sec</td>
</tr>
<tr>
<td>Data Delay</td>
<td>500 sec</td>
<td>1-1800 sec</td>
</tr>
<tr>
<td>Run Time</td>
<td>2700 sec</td>
<td>300-14000 sec</td>
</tr>
</tbody>
</table>
Chapter 3 Optimisation and Development of Insertion Element IS6110 Fluorescent Amplified Fragment Length Polymorphism (FAFLP) PCR
3.1 Introduction

Fluorescent Amplified Fragment Length Polymorphism (FAFLP) has been widely used as a strain typing tool for several bacteria like *Listeria monocytogenes* (134), *Legionella pneumophila* (145,176) and *Salmonella enteritidis* (142). Using IS6110 as a biomarker, the global phylogeny of MTBC has been studied using FAFLP (135).

In the era of whole genome sequencing, where repeat elements are difficult to sequence with short read technologies (177), and in the absence of an alternative approach to their characterisation in poorer resource settings, a simple but effective technique would enable the facilitation of outbreak samples by not only strain typing the bacterial strains but also to help in the classification of TB genetic lineages in any geographical location. A rapid and simple method of insertion site mapping using IS6110 fluorescent amplified fragment length polymorphism (FAFLP) PCR was developed using fluorescent labelled primers (147). Borrell et al. first used four different fluorescent labelled primers with the selective base at the 3' end of the *TaqI* primer for increased resolution to differentiate between epidemiologically unrelated TB strains. Thorne et al. used fluorescent labelled primers to further distinguish the fragments (135,146). To suit the demands of a resource poor setting the technique was modified further by utilising the resolution of the fragments derived by this technique, and to develop a lineage specific PCR as described in chapter 4.

3.1.1 Aims

The main aim of this chapter was to modify, optimise and standardise the IS6110 FAFLP PCR methodology so that it could be used to map and classify *M. tuberculosis* samples into different lineages.
3.1.2 Objectives

- To test the IS6110 FAFLP methodology using DNA from different strains of *M. tuberculosis* and different PCR reagents/conditions.

3.1.3 IS6110 FAFLP Methodology

The technique was optimised so that the results obtained could be easily identified by the presence or absence of the fragments/peaks and by their respective colours as shown in the figure 3-1. Each fragment indicates the point of insertion of IS6110 in the bacterial genome and also the number of fragments indicates the number of IS6110 insertion sites i.e. copy number of IS6110 in the genome. The PCR amplified fragments were run on ABI 3730XL (Thermofisher, U.K.) and visualised using software called Peak Scanner (Thermofisher, U.K.).

![Figure 3-1 Pictorial representation of the IS6110 FAFLP PCR methodology](image)

(Coloured fragments are amplified using uniquely labelled adapter specific primers. For example, blue coloured fragments are amplified if the primers are labelled with 6-FAM fluorescein 'blue' dye, red coloured fragments with PET dye, green coloured fragments with VIC dye and black coloured fragments with NED dye.)
3.2 Materials and Methods

The process to optimise the IS6110 FAFLP PCR methodology consisted of testing the DNA samples of adapter ligated H37Rv DNA from previous study by Kremer et al. (178), H37Rv and Nepal TB strains with different polymerases and optimising the PCR conditions. Approximately 200-500ng of bacterial genomic DNA was digested with 5U of MseI (New England Biolabs, UK) in a total reaction volume of 20µl containing 1x MseI buffer, 0.1mg/ml of BSA, 0.5mg/ml of RNase A, 1M of L-Proline and 2mM of Betaine at 37°C for 2h to release the DNA ready for a final digestion with 10U of TaqI (New England Biolabs, UK) at 65°C for 3 h and the reaction was inactivated for endonucleases at 80°C for 10 min. Double stranded oligonucleotide adaptors specific to TaqI site (5’- TACTCAGGACTGGC) were then ligated to the double- digested DNA in a total reaction volume of 25µl containing 40U of T4 DNA ligase (New England Biolabs, UK), 0.2µM of double stranded TaqI adaptor and 2x T4 ligase buffer (NEB, UK). The reaction mixture was then incubated at 12°C for 17 h followed by heating at 65°C for 10 minutes to inactivate the ligase and later stored at -20°C. In a total volume of 25µl, 2.5µl of the DNA sample was mixed with 10x Pfx reaction buffer, 50mM MgSO4, 10mM dNTP, 1µM of TaqI primer, 1µM of IS6110 specific primer and 1U Pfx polymerase/ 1µl of Hotstar Taq plus mastermix or 10µl of My Taq Premix or 1U of recombinant Taq polymerase. The following PCR conditions were carried out in a Veriti thermocycler (Thermofisher, U.K.): 94°C for 15 min followed by 35 cycles of 94°C for 20 s, 66°C for 30 s and 72°C for 2 min with the 66°C annealing temperature reducing by 1°C every cycle for nine cycles and the last 25 cycles at 56°C. Finally, an extension of 60°C for 30 min was carried out before further manipulation. When there were no peaks or there was an anomaly in the
electropherogram such as amplification artefacts and double peaks, then the PCR extension was increased from 60°C for 30 min to 72°C for 30min and a final adjustment was carried out from 72°C for 30min to 72°C for 60min. The criteria for the peak characteristics seen in an electropherogram using Peak Scanner software were as follows:

a. The fragment should be a single peak (Double peaks should be avoided). Two peaks over 0.5 base pair difference between them will be considered as two separate fragments.

b. The fragment could be identified by the colour of the fluorescently labelled dye, either blue for FAM or red for PET or green for VIC or black for NED according to the dye incorporated.

c. The peak should be a true peak shape and size (artefacts were discarded). A difference of ± 0.5bp in fragment sizes was considered as non-identical fragments or peaks.

d. The relative fluorescent units (rfu) on the X-axis had been assigned a cut-off of 500rfu.

Some of the well-characterised artefacts are as follows:

a. Stutter peaks: Stutter artefacts are observed as multiple peaks preceding the true allele peak. The number of peaks and their intensities are proportional to the length of the repeat and the number of repeats in the PCR product (Applied Biosystems Fact sheet).
b. Plus A additions or Double peaks: Plus A additions, caused by incomplete A nucleotide addition, also increase the complexity of the peak pattern, which makes recognition of true allele peaks more difficult (Applied Biosystems Fact sheet).

c. Bleed through or Pull-up peaks: A Bleed-through or pull-up is a typical problem in multichannel capillary electrophoresis, which is caused by fluorescent dye cross-talk. A high signal in one fluorescent channel thereby gives rise to a pull-up peak in another channel (Applied Maths Inc., Belgium).

The step by step procedure to achieve the standardisation of IS6110 FAFLP PCR assay is given in the form of flowchart below in the figure 3-2.
Figure 3-2 Step-by-step procedure for optimising and developing the IS6110 FAFLP methodology
3.3 Results

The IS6110 FAFLP PCR methodology was standardised by testing different DNA polymerases from different PCR kits and by changing the PCR extension step as shown in figure 3-3 using adapter ligated H37Rv DNA from a previous study by Kremer et al. (178), H37Rv and Nepal TB DNA samples. The sequential procedure undertaken to achieve this is described below in the form of a flowchart.
Figure 3-3 Step-by-step procedure of standardising IS6110 FAFLP methodology including the accepted methodology
3.3.1 Amplification of DNA fragments

Amplification of the DNA fragments was first checked for the adapter ligated DNA by the presence or absence of peaks (fragments) in the PeakScanner analysis software. Absence of peaks (figure 3-4) except for the presence of LIZ600 indicated that either poor quality DNA or absence of DNA.

![PeakScanner screenshot](image)

Figure 3-4 Example of a Peakscanner screenshot showing the absence of amplification of fragments except for Liz600 marker in orange colour.

3.3.2 Validation of the PCR reagents

Different Taq polymerases (*PfX* polymerase, Hotstar *Taq* plus mastermix, *My Taq* Premix and recombinant *Taq* polymerase) were subjected on both H37Rv and Nepal TB DNA and the results were as follows.
3.3.2.1 Hotstar Taq Plus mastermix (Qiagen, U.K.)

The fragments were not amplified except for the presence of the visualisation marker Liz600 using this mastermix and so the next reagent was tested.

3.3.2.2 MyTaq PCR premix (Bioline, U.K.)

Using this mastermix, strong amplification occurred but the interpretation was difficult as seen in figure 3-5 due to the strong signal intensity of more than 28000 relative fluorescent units (rfu).

Figure 3-5 Example of a Peakscanner screenshot showing strong signals of more than 28000 rfu when MyTaq mastermix was used
3.3.2.3 Platinum \textit{Pfx} proof-reading polymerase (Invitrogen, U.K.)

Here, the signal intensity of the samples under analysis was optimum but the interpretation of the data became difficult as there was more than one coloured fragment at any single position as seen in figures 3-6 and 3-7.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{peakscanner.png}
\caption{Example of a Peakscanner screenshot showing amplification of fragments with correct signal intensity in a sample when Platinum \textit{Pfx} Polymerase proof-reading enzyme mastermix was used}
\end{figure}
Figure 3-7 Second example of the Peakscanner screenshot showing amplification of fragments from the sample above in the previous figure but a mixture of fragments noticed at a single position leading to difficult interpretation of the data when Platinum Pfx Polymerase proof-reading enzyme mastermix was used

3.3.2.4 Recombinant Taq Polymerase (Invitrogen, U.K.)

The majority of the samples analysed using this enzyme were amplified with optimum signals apart from a few samples which showed stronger signal intensity (figure 3-8) and double peaks (figure 3-9). The electropherogram for the reference strain, H37Rv, was very clean and the signal intensity was normal (see figure 3-10). The negative control, water, did not show any amplification (see figure 3-11).
Figure 3-8 Example of a Peakscanner screenshot showing strong raw signal intensity in H37Rv sample using recombinant Taq Polymerase reagents
Figure 3-9 Example of a Peakscanner screenshot showing double peaks (green dye - VIC) in H37Rv sample using recombinant Taq Polymerase reagents
Figure 3-10 Example of a Peakscanner screenshot showing the presence of amplified fragments with single peaks in an H37Rv sample when using recombinant Taq Polymerase reagents.
3.3.3 H37Rv and Nepal TB DNA study samples

Changing the PCR conditions significantly increased the efficiency of the IS6110 FAFLP assay and, together with enzyme optimisation development, the assay was standardised. So, using this technique, H37RV (see figure 3-9) and the 176 Nepal samples were characterised (see appendix 5- raw FAFLP data stored in CD format). It is not possible to include the electropherograms for all the 176 samples due to space constraints but the example below, figure 3-12, shows the amplification of fragments for two samples displaying well-defined fragments without artefacts.
3.4 Discussion

Whilst testing the samples from already adapter-ligated DNA (178), it was apparent that using poor quality DNA was unsuccessful as there was no amplification of products as this DNA was from the 1999 study. This was overcome by testing fresh DNA samples from H37Rv and Nepal TB strains and the FAFLP procedure was carried out on the samples including restriction enzyme digestion, ligation and PCR. Unaltered PCR conditions and reagents led to the generation of double peaks (figure 3-8), which created difficulty in interpreting the data. It was apparent that the PCR conditions required optimisation, including change of PCR reagents including PCR enzymes. The PCR final extension step was increased from 60°C to 72°C for 15 minutes but otherwise all other conditions remained the same. Of four PCR reagents tested, the fragments were not amplified using Hotstar Taq Plus mastermix. With
regards to MyTaq mastermix, the signal intensity of 28000rfu was high and led to the
difficulty in interpretation of data. The signal intensity for the third PCR reagent, Pfx
Polymerase gave optimum signal intensity and also the background noise was low
compared to former two reagents mentioned above but there were more than one
fragment on one given location which made the analysis difficult. The fourth reagent
tested, recombinant Taq polymerase mastermix, gave good signal intensity but also
produced double bands which indicated that the PCR conditions needed further
optimisation. A final extension of 72°C for 60 minutes was carried out instead of 15
minutes as in previous protocols and gave well-defined, easily interpretable
fragments and thus the standardisation was complete. This technique was used in
chapter 4 for mapping the IS6110 insertion sites in H37RV genome and in chapter 6
for characterisation of Nepalese TB strains into different TB genetic lineages.

3.5 Summary

In this chapter, the IS6110 FAFLP PCR procedure was standardised and tested
using the adapter ligated DNA products from a global TB collection (178), H37Rv
and followed by samples from Nepal. I selected the latter for my thesis to develop
the genomic tool box for resource poor settings as Nepalese samples were very
poorly characterised when this study was started in 2011.
Chapter 4  Mapping of Insertion sites (IS6110) in the *M. tuberculosis* H37Rv reference genome and rapid definition of genetic lineages

(This chapter was published in the Journal of Microbiological Methods) (179)
4.1 Introduction

Insertion elements play a significant role in the evolution of genomes of various organisms (69) and can transpose within the genomes thereby altering the position and even the function of the adjacent genes. These transpositional changes have contributed to the evolution of different microorganisms thereby exhibiting a genome plasticity function where the genome is altered to survive in different environmental niches (75,180).

Tracing the movement of insertion sites can be useful in two ways, firstly by acting as genetic markers to define strains of different microorganisms and thus determine the transmission relationship between two strains (181) and secondly by mapping the exact position of IEs whose transposition could alter the phenotype of the strain. Mapping of IS6110 sites in the Zaragoza *M. tuberculosis* strain that caused a sudden outbreak in Spain in the years between 2001 and 2004, has offered clues of the adaptability and virulence of *M. tuberculosis* (182). The Zaragoza strain belongs to PGG3 and is responsible for widespread outbreaks compared to PGG1 and PGG2 (refer section 1.9.1 for the descriptions about PGGs) in Spain. On further analysis of the mapping results, Millan-Lou et al. have found that there are twelve copies of IS6110 in this strain and that the IS6110 insertion in the DR region is found 3bp away from the DR sequence which is different to the compared H37Rv strain. Also, five of the twelve insertion points were at positions already reported by other scientific groups over the years. Most importantly, one of the insertion locations, in the Rv2823c gene was unique and specific to the Zaragoza strain which then led to the development of multiplex PCR assay for the rapid detection of this strain of *M. tuberculosis* (182).
Genomic mapping identified that there are differences between two strains of Beijing family, W and 210. These two strains have seventeen IS6110 insertion sites, twelve are identical in terms of insertion locations in the genome but five are different. Two unique sites were found in W strain and also one IS6110 copy was found upstream of the ctpD gene wherein IS6110 was providing a promoter element for the transcription of this gene (98). Alonso et al., found specific IS6110 insertion sites in the Beijing genotype that helped in the mapping of this family of TB and also showed that the Beijing TB lineage displayed a higher frequency of IS6110 insertions than other lineages (183). TB genetic lineages, as described in chapter 1 section 1.9, are grouped mainly into two types, either by SNP classification (153) into 3 Principal genetic groups (PGG1, PGG2 and PGG3) or by LSP classification (184) into seven genetic lineages. These genetic families or lineages are important as it has been suggested that there is phenotypic diversity existing between M. tuberculosis clinical isolates (165). Also, it has been reported that Beijing/W lineages, with their unique phenotypic traits, have an increased ability to cause disease (185,186). Also, this strain has been shown to be predominantly distributed in Eastern Asia and due to human migration this genotype migrated from Beijing City to the other parts of the world over time. However, a high number of insertions in certain strains does not necessarily suggest pathogenicity but that the molecular clocks of IS6110 movement is faster in some strains than others, as it has been shown that low copy numbers are also highly pathogenic (107).

As there is a lot of ambiguity in defining a specific TB genetic lineage using the available epidemiological markers, and because the techniques described above are complex in terms of technicality, a rapid definition assay will be helpful in providing
information about the TB genetic lineage and possibly then allowing improved management of the disease caused by a particular genetic lineage.

4.1.1 Aims and Objectives

The aim of this chapter is to locate or map the precise genomic IS\textit{6110} insertion sites in the \textit{M. tuberculosis} H37Rv strain using IS\textit{6110} FAFLP PCR methodology and define the TB genetic lineages, H37Rv (PGG3 group) and Beijing (PGG1 group). Rapidly defining TB lineages would help in identification of TB genetic lineages circulating in a given area of the outbreak and thereby indirectly help in the informed management of the disease including control and surveillance.

The following specific objectives are used to achieve the aim mentioned above:

- To carry out IS\textit{6110} FAFLP PCR methodology on H37RV genomic DNA
- To decrease the number of fragments generated by FAFLP to enable sequencing of individual fragments so that it is easy to interpret the insertion points of IS\textit{6110} in the genome.
- To identify the sequence of individual fragments generated and map the insertion site IS\textit{6110} position in the reference sequence of H37Rv.
- To develop and perform lineage specific PCR to define lineages as PGG1-Beijing and PGG3-H37Rv.
4.2 Materials and Methods

H37Rv TB DNA was subjected to IS6110 FAFLP methodology as described in chapter 3 (chapter 3, section 3.1.3) and the fragments generated by capillary electrophoresis under the run conditions described in chapter 2 (chapter 2, table 2.2), were analysed using PeakScanner software (Thermofisher, UK). To further reduce the number of fragments, new two base selective primers were designed with an addition of another selective base at the 3’end of the existing TaqI primers (see table 4-1 below) and subjected to the previously used PCR conditions using IS6110 reverse primer. One microlitre of the PCR product is used to sequence in both directions using the same primers. The sequence data is then subjected to BLAST analysis to reveal the identity of the fragment thereby allowing the fragment of interest to be mapped at the genome level (see Tables 4-2 and 4-3)

Table 4-1 List of unlabelled two base selective primers

<table>
<thead>
<tr>
<th>Two base selective TaqI primer</th>
<th>Primer Sequence (Two selective bases at 3’ end)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1) TaqI- AA</td>
<td>5’- CGATGAGTCCTGACCAGA(AA)-3’</td>
</tr>
<tr>
<td>2) TaqI- AC</td>
<td>5’- CGATGAGTCCTGACCAGA(AC)-3’</td>
</tr>
<tr>
<td>3) TaqI- AT</td>
<td>5’- CGATGAGTCCTGACCAGA(AT)-3’</td>
</tr>
<tr>
<td>4) TaqI- AG</td>
<td>5’- CGATGAGTCCTGACCAGA(AG)-3’</td>
</tr>
<tr>
<td>5) TaqI- CA</td>
<td>5’- CGATGAGTCCTGACCAGA(CA)-3’</td>
</tr>
<tr>
<td>6) TaqI- CC</td>
<td>5’- CGATGAGTCCTGACCAGA(CC)-3’</td>
</tr>
</tbody>
</table>
7) *TaqI*- CT  
| 5'- CGATGAGTCCTGACCGA(CT)-3' |

8) *TaqI*- CG  
| 5'- CGATGAGTCCTGACCGA(CG)-3' |

9) *TaqI*- TA  
| 5'- CGATGAGTCCTGACCGA(TA)-3' |

10) *TaqI*- TC  
| 5'- CGATGAGTCCTGACCGA(TC)-3' |

11) *TaqI*- TT  
| 5'- CGATGAGTCCTGACCGA(TT)-3' |

12) *TaqI*- TG  
| 5'- CGATGAGTCCTGACCGA(TG)-3' |

13) *TaqI*- GA  
| 5'- CGATGAGTCCTGACCGA(GA)-3' |

14) *TaqI*- GC  
| 5'- CGATGAGTCCTGACCGA(GC)-3' |

15) *TaqI*- GT  
| 5'- CGATGAGTCCTGACCGA(GT)-3' |

16) *TaqI*- GG  
| 5'- CGATGAGTCCTGACCGA(GG)-3' |

4.2.1 Rapid definition of TB genetic lineages-PGG3 (H37Rv) and PGG1 (Beijing)

Samples: H37Rv and samples belonging to other lineages already assigned using FAFLP (chapter 5) (N10, N25, N34, N46, N62, N63 N70), were selected (appendix 1) to test the proof of principle. Based on FAFLP, the following lineages were assigned to these samples above (Table 4-2).
<table>
<thead>
<tr>
<th>Sample ID</th>
<th>FAFLP derived lineages (PGG)</th>
</tr>
</thead>
<tbody>
<tr>
<td>N10</td>
<td>CAS – PGG1</td>
</tr>
<tr>
<td>N25</td>
<td>Unassigned</td>
</tr>
<tr>
<td>N34</td>
<td>T (H37Rv-like)- PGG3</td>
</tr>
<tr>
<td>N46</td>
<td>Unassigned</td>
</tr>
<tr>
<td>N62</td>
<td>Unassigned</td>
</tr>
<tr>
<td>N63</td>
<td>Unassigned</td>
</tr>
<tr>
<td>N70</td>
<td>Beijing (PGG1)</td>
</tr>
</tbody>
</table>

Following sequence determination (and therefore specific genomic insertion sites) of lineage specific fragments found only in defined lineage/spoligotype groups by FAFLP, primers were designed to amplify common fragments from the PGG1 (5’-gggctcccccagatca-3’) and PGG3 (5’-gtgtgccgcgaggtg-3’) groups and comprised a 200bp product and a 296bp product on a gel respectively, based on the hypothesis that a PCR product of a known size will be generated if the IS element is inserted at the exact same genomic region. In a total volume of 50µl, 1µl of the DNA extracted was added to the reaction containing 1X PCR reaction buffer, 1.5 mM MgCl₂, 0.2 mM dNTPs (Invitrogen, UK), 5 µM of the PGG1 or PGG3 group primer, 5 µM of IS6110 reverse primer as described above and 1U of recombinant Taq polymerase (Thermofisher, UK). The following PCR conditions were carried out in a Veriti thermocycler (Thermofisher, UK): 96°C for 15 min followed by 30 cycles of 96°C for
1 min, 60°C for 1 min and 72°C for 30 seconds with a final extension of 72°C for 1 min was carried out before cleaning up the PCR products using AMPure XP magnetic beads. The products were then run on a gel to check the presence of the fragment, 296bp product for the PGG3 group and 200bp product for the PGG1 lineage.

4.3 Results

4.3.1 Prediction of fragments using SeqBuilder (DNASTAR Lasergene version 8, USA) software (in silico analysis)

The genomic sequence of H37Rv (accession no: NC_000962.2) was digested in silico with TaqI enzyme followed by the identification of 59 bases of IS6110 sequence (CTGACATGACCCCATCTTTCCAAGAAGACTGGAGTCTCCGGACATGCGGGGC GGTTCAG) to indicate the size of the expected product by PCR using primers with a two base extension at the 3’ end to reduce the number of fragments generated as shown in Tables 4-3 and 4-4 and Appendix 1. The table shows the presence of sixteen insertion sites corresponding to the presence of sixteen IS6110 copies in H37Rv as was reported previously by Philipp et al in 1996 (187).

4.3.2 Identification of fragments using Peak Scanner (Thermofisher Scientific, UK) software (in vitro analysis)

Sixteen fragments were mapped as seen in columns 1 and 2 in tables 4-3 and 4-4 using their fluorescent dye colours.
4.3.3 Mapping the insertion site IS6110 position in the genome of *M. tuberculosis*.

All but two, 71.5 Y and 83.8 Y as seen in tables 4-3 and 4-4 in column 1, of the sixteen fragments generated could be identified by this method. Two fragments could not because the flanking genomic data contained in the sequence was too short to analyse as the restriction site was too close to the end of the transposon sequence. Although the two base selective primers are specific and generate a largely specific PCR product for sequencing, a degree of mis-priming does occur. These non-specific extension products are generated due to mis-priming occurring in the PCR. The data generated for the H37Rv genome fragmented using the 4-dye FAFLP PCR in Tables 4-3 and 4-4 and Appendix 1 correlate with the sequencing data and also the *in silico* data generated using DNASTAR Lasergene Seqbuilder as described above.

Table 4-3 Mapping of H37Rv genome using 4-dye FAFLP PCR, DNA sequencing including BLAST results and *in silico* analysis with sequence orientation 5’-3’ according to Figure 3-1 Method Schematic

<table>
<thead>
<tr>
<th>FAFLP fragment colour and size in base pairs*</th>
<th>Sequence product size</th>
<th>Insertion site BLAST result from sequence data</th>
<th><em>In silico</em> predicted insertion site</th>
</tr>
</thead>
<tbody>
<tr>
<td>99.1 G</td>
<td>(102)</td>
<td>Transcriptional regulator for LacI family protein</td>
<td>Transcriptional regulator for LacI family protein</td>
</tr>
<tr>
<td>FAFLP fragment colour and size in base pairs*</td>
<td>Sequence product size/in silico predicted size in base pairs</td>
<td>Insertion site BLAST result from sequence data</td>
<td>In silico predicted gene insertion site</td>
</tr>
<tr>
<td>---------------------------------------------</td>
<td>----------------------------------------------------------</td>
<td>-------------------------------------------------</td>
<td>--------------------------------------</td>
</tr>
<tr>
<td>81.0B</td>
<td>Too short to analyse</td>
<td>Cutinase1</td>
<td></td>
</tr>
<tr>
<td>173.6R</td>
<td>Conserved membrane protein</td>
<td>Conserved membrane protein</td>
<td></td>
</tr>
<tr>
<td>736.6R</td>
<td>IS6110</td>
<td>IS6110</td>
<td></td>
</tr>
</tbody>
</table>

* G=Green, B=Blue, R=Red, Y=Yellow

Table 4-4 Mapping of H37Rv genome using 4-dye FAFLP PCR, DNA sequencing including BLAST results and in silico analysis with sequence orientation 3’-5’ according to Figure 3-1 Method Schematic.
<table>
<thead>
<tr>
<th>ID</th>
<th>Score</th>
<th>Description</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>191.7R</td>
<td>196</td>
<td>Conserved hypothetical protein</td>
<td>Conserved hypothetical protein</td>
</tr>
<tr>
<td>71.5Y</td>
<td>58</td>
<td>Too short to analyse</td>
<td>Cutinase1</td>
</tr>
<tr>
<td>160.3Y</td>
<td>164</td>
<td>LytB related protein</td>
<td>LytB related protein</td>
</tr>
<tr>
<td>83.8Y</td>
<td>58</td>
<td>Too short to analyse</td>
<td></td>
</tr>
<tr>
<td>Not seen in FAFLP data</td>
<td>289</td>
<td>Putative transposase for IS986</td>
<td>Putative transposase for IS986?</td>
</tr>
<tr>
<td>Not seen in FAFLP data</td>
<td>214</td>
<td>Integrase</td>
<td>Integrase?</td>
</tr>
<tr>
<td>329.2B</td>
<td>338</td>
<td>Glycerolphosphodiesterase</td>
<td>Glycerolphosphodiesterase?</td>
</tr>
<tr>
<td>124.9G</td>
<td>127</td>
<td>Transposase</td>
<td>Transposase</td>
</tr>
<tr>
<td>360.0R</td>
<td>360</td>
<td>Oxidoreductase</td>
<td>Oxidoreductase</td>
</tr>
</tbody>
</table>

* G=Green, B= Blue, R=Red, Y=Yellow
4.3.4 Rapid definition of TB genetic lineages- PGG3 and PGG1.

For the seven samples tested as proof of principle, five belonged to the PGG3 lineage: H37Rv, N25, N34, N46, N63, one belonged to the PGG1 lineage, N70 according to FAFLP data, one belonged to the CAS lineage, N10 and one (N62) did not give any result (see figure. 4-1). N70 reacting to PGG3 primer shows there might be fragments shared by different PGG groups and needs to be looked into detail.

A.

![Image of gel electrophoresis with marker bands and sample wells labeled from 10 to 70, with a 296bp band indicated.]
Figure 4-1 Agarose Gel Electrophoresis showing *M.tuberculosis* strains, N10, N25, N34, N46, N62, N63, N70, H37Rv, amplified using PGG3 (A) and PGG1 (B) specific primers, amplifying 296 bp and 200 bp products respectively.
**4.4 Discussion**

Although a large number of studies have been published based on the fragment patterns generated by IS6110 RFLP, little or no data are available on the distribution of IS6110 genomic insertion sites in *Mycobacterium tuberculosis* complex strains. This will have an impact on the molecular epidemiological studies of *M. tuberculosis*. Knowledge of specific sites may lead the way to develop rapid techniques to identify specific lineages which could be validated in future using WGS technology.

This technique exploits the use of selective bases at the 3’ end of primer sequences to reduce the fragment numbers generated during FAFLP PCR amplification. Also, use of differentially labelled fluorescent primers aids the identification of fragments. Using selective primers reduces the number of fragments amplified, which can then be sequenced and the position of the insertion site in the genome identified by BLAST analysis of the sequence. The BLAST results show that, except for the three fragments (81B, 71.5Y and 83.8Y) that were too short to analyse, the predicted and actual insertion sequence data correlate with each other and the FAFLP fragment size and colour. As every fragment generated always contained 75 bp of transposon/adaptor sequence, any fragments below 100bp were too small to identify flanking genomic sequence by this method. Sequencing data usually starts around 25 bp past the sequencing primer which effectively cuts off 100bp in total. Apart from the fragments too small to analyse (this could be remedied by choosing an alternative enzyme for FAFLP), there were three fragments that were not predicted in silico. These were an IS986 transposase (corresponding to 289bp fragment in table 4-3), integrase (214bp in table 4-3) and glycerol phosphodiesterase (329.2B in table 4-3), involved in the glycerol metabolic process. That another transposon was
identified strongly suggests that a further transposition has occurred following the many cultural iterations of H37Rv since whole genome sequencing as seen in the case of IS986 transposase. If this was the case then these fragments would not be predicted from the genome sequence available. The predicted insertion sequences seen offer interesting avenues for further study, including characterisation of the surface genes disrupted by IS6110 including PPE and those affecting metabolism. The characterisation of insertion sites of this and other transposons is important for clonal organisms such as those belonging to the MTBC.

In 2011 Alonso et al., showed the importance of mapping insertion sites by demonstrating that the IS6110 in the Beijing strains of MTBC can up-regulate downstream genes via an outward-directed promoter in its 3’end (Alonso et al. 2011). Further mapping of the specific insertion sites will generate information on the nature of the gene disruption, whether the insertion has a detrimental effect, for example disrupts the proposed reading frame, or has a potentially beneficial effect through up regulation. The next steps for this work after it has been shown to map successfully the IS6110 insertion sites in H37Rv is to apply this knowledge to design a rapid tool for the definition of lineages.

The principle of rapid detection of lineages was successful using the lineage specific fragments targeting the common fragments as described above (see figure 4.1). This assay needs a simple PCR thermocycler, reagents and facilities to run and visualise the gel and the ability to interpret the results. Direct detection from sputum would be a useful next step but sensitivity is likely to be low. Detection of lineages by the LSP or the SNP methods are time consuming as they need good quality DNA after bacterial DNA extraction and a bioinformatics approach to deduce the TB families.
They also need expensive instruments like the sequencers and extra manpower to accomplish the definition of lineages. Although an alternative approach would now be made if starting the project, the FAFLP approach was the only approach available at the time and as such it still forms the basis of genomic tool box development for the rapid assignation of TB genetic lineages based on IS\textit{6110}.

4.5 Summary

The data presented shows that the IS\textit{6110} FAFLP PCR technique, though relatively simple, is robust and effective in mapping IS\textit{6110} insertion in H37Rv that could be applied to any bacterial species with similar repeated elements. Also in this chapter, rapid definition of TB genetic lineages using newly designed primers were carried out using PCR without the use of digestion and ligation, for the rapid detection of both H37Rv-like (T-group) and Beijing lineages to initiate the development of a suite of PCRs based on detection of lineage specific common IS\textit{6110} insertion points as a typing tool for resource poor settings.
Chapter 5 Classification of Nepalese TB clinical isolates into different TB genetic lineages

(This chapter was published in the open access journal of Clinical Microbiology and Infectious Diseases (188).)
5.1 Introduction

Nepal is geographically interlocked between China and India, which together account for approximately a third of annual global new cases (11% and 24%, respectively) (WHO 2014) and is ranked 43rd in the world in terms of age-adjusted death rate, which is 27.80 / 100000 of the population. During the year ending of 2014, there were 5506 deaths in Nepal due to tuberculosis that accounts to 3.5% of total deaths in Nepal. TB cases and deaths are more prevalent in men than women. In 2006 Nepal’s National Tuberculosis Programme (NTP) developed the STOP TB strategy to fight against tuberculosis, in line with WHO. The NTP is fully integrated with the national healthcare policy of the government of Nepal. The Millennium Development Goal (MDG), another initiative by WHO to tackle TB, was found to be effective in halting and reversing global TB prevalence. The prevalence rate (all per 100,000 population) has gone down from 348 (162-602) in 1990 to 211 (99-365) in 2015. The death rate (all per 100,000 population) has also reduced from 52 (32-70) in 1990 to 19 (13-25) in 2015. Due to all these efforts, Nepal is no longer in the list of high burden TB countries. TB programs were able to save 32,973 lives in 2014 but despite this progress, there were still 978 deaths. Overall treatment success rates of drug susceptible TB was 91.5% with 0.92 % failure rates, 2.17% failed to follow-up and 2.7% death rates. The proportion of MDR-TB was 2.2% among new TB cases and 15.4 % among retreatment cases based on a survey carried out in 2011-2012. New surveillance studies have not been carried out recently. There were a total of 22 deaths among MDR cases and 3 deaths in XDR reported in 2014/15. For MDR the drug resistance pattern showed higher levels of resistance in fluoroquinolones and 8% of those MDR patients further developed XDR. Sputum microscopy is still the gold standard method for the diagnosis of TB in Nepal. Currently there are 581
microscopy centres providing this smear microscopy service throughout the country. As a high proportion of rifampicin resistant strains are resistant to Isoniazid, detection of rifampicin resistance can be used as a marker for MDR-TB as described in chapter 1. Between 2011 and 2012, GeneXpert MTB/RIF assays were introduced in 3 centres and now there are 26 centres nationally (173). The STOP TB program has reduced the number infected from 29 million people, as estimated by WHO in 2014, to 15 million people according to the recent report by NTC. Despite the efforts to control the spread of TB, there is still a long way to go to achieve the goals of STOP TB program as 40,000 people are infected every year, with 20,000 new primary cases and 5000-7000 deaths each year from tuberculosis in Nepal.

It is important to understand the molecular diversity of the *M. tuberculosis* population in Nepal as it has been reported recently that there is a high similarity between TB strains in Nepal and in Northern India, with which Nepal shares an open border policy of human migration (189). This might indicate that the TB lineages that were found in India, especially the CAS lineage, have successfully established in Nepal due to the human migration from India into Nepal (190), enabling the TB lineages to be successful in any geographical location (189).

Very limited data are available on the characterisation of *Mycobacterium tuberculosis* strains and genotypes circulating in Nepal. A recent study of 261 Nepalese isolates found drug resistance in 12.8% of *M. tuberculosis* strains that were from new untreated cases, with the most frequent lineages reported as CAS/Delhi (40.6%), East Asian (including Beijing) (32.2%), Euro-American (15.7%) and Indo-oceanic (11.5%) (191). To gain further insight into the characteristics and diversity of
mycobacteria in Nepal, this study aimed to categorise isolates for the first time using IS6110 FAFLP PCR and to assign them to different genetic lineages.

5.1.1 Aims and Objectives

The main aim of this chapter is to classify the Nepal samples into different TB lineages to understand the genetic distribution of TB lineages in Nepal.

The following objectives were used to achieve this aim:

- To perform IS6110 FAFLP PCR on Nepal samples
- To characterise the samples into different TB lineages

5.1 Materials and Methods

5.1.1 Strains

(Refer General methods section 2.1)

Sputum samples from 176 consecutive new TB patients were collected between 2007 and 2008 and cultured alongside routine diagnostic testing from two Nepalese tuberculosis reference centres located in the Kathmandu valley: the National Tuberculosis Centre (NTC) and the German Nepal Tuberculosis Project (GENETUP). The patient population represented local and referred cases from across Nepal. Bacterial genomic DNA from isolated strains was extracted by the Cetyltrimethylammonium Bromide (CTAB) method.
5.1.2 IS\textit{6110} FAFLP PCR, Fragment Sizing & Analysis

(Refer to the General methods section 2.2 for IS\textit{6110} FAFLP methodology and fragment sizing.). The four-dye FAFLP data collected from the different profiles were recorded and compared with a well characterised reference collection of \textit{M. tuberculosis} isolates (178) using BioNumerics software v6.1 (Applied Maths Inc., Belgium). These data were then used to build a dendrogram using the Dice coefficient of similarities to compare the similarity matrix and Unweighted Pair Group Method with Arithmetic Mean (UPGMA) derived cluster analysis with cophenetic correlation for the branch quality. Dice similarity coefficient and UPGMA clustering method was chosen because of the high correlation in terms of their similarity obtained (in this case, genetic similarity) between the samples to form a group or cluster compared to the simple matching coefficient.

5.2 Results

5.2.1 Analysis of Data using BioNumerics software v6.1

Of the 176 DNA extracts from isolates analysed, the majority of the samples (97, 55.4%) belonged to either the spoligotype-defined CAS lineage (64, 36.6%) or the Beijing lineage (33, 18.8%) grouping under PGG1 and the rest of the samples group under either PGG2 (1.7% S, 3.97% X, 7.95% Haarlem and 2.27% LAM, 2.27% T-Uganda) or PGG3 (2.27% of T) (Table 5-1). Forty three samples (24.4%) were “unassigned”. Common fragments specific to different TB lineages seen were exactly the same as the earlier published report by Thorne et al, (2011) except for an additional fragment, 78.4 G, for the CAS lineage (135). These fragments specify different insertion points where IS\textit{6110} has inserted or transposed. A dendrogram
was generated using only the IS6110 FAFLP data confirming again the above mentioned lineages in relation to the PGGs (Fig. 5-1).
Table 5-1 Common fragments identified using IS6110 FAFLP PCR in TB genetic lineages of the 176 bacterial DNA isolates in Nepal

<table>
<thead>
<tr>
<th>PGG-spoligotype/sub-lineage</th>
<th>Common fragment sizes</th>
<th>No. Of Nepal strains (Total=176)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PGG1-CAS</td>
<td>78.4G, 92.0B, 117.9R, 206.2G, 275.1R</td>
<td>64</td>
</tr>
<tr>
<td>PGG1-Beijing</td>
<td>101.7B, 102.5Y, 139.1R, 180.7Y, 254.8G, 332.4R, 353.5B</td>
<td>33</td>
</tr>
<tr>
<td>PGG2 –Haarlem</td>
<td>87.0Y, 89.4G, 148.7B, 300.2R (H/X), 445.7Y</td>
<td>14</td>
</tr>
<tr>
<td>PGG2-LAM</td>
<td>71.5Y, 105.2R, 116.1R</td>
<td>4</td>
</tr>
<tr>
<td>PGG2-S</td>
<td>88.0G, 112.9R, 217.2R, 445.3G</td>
<td>3</td>
</tr>
<tr>
<td>Group</td>
<td>Spoligotypes</td>
<td>Count</td>
</tr>
<tr>
<td>----------------</td>
<td>-------------------------------</td>
<td>-------</td>
</tr>
<tr>
<td>PGG2-X</td>
<td>83.8Y, 300.2R (H/X)</td>
<td>7</td>
</tr>
<tr>
<td>PGG2-T Uganda</td>
<td>88.9Y, 119.5G, 122.9Y, 228.4Y, 266.8R</td>
<td>4</td>
</tr>
<tr>
<td>PGG3- T</td>
<td>81.3R, 192.4R, 360R</td>
<td>4</td>
</tr>
<tr>
<td>Ungrouped</td>
<td></td>
<td>43</td>
</tr>
</tbody>
</table>

(PGG represent Principal Genetic groups according to Sreevatsan et al, spoligotypes follow spolDB4 classification and spoligotypes derived sub-lineages are grouped following Gagneux’s classification (153,158)).
Figure 5-1 UPGMA derived dendrogram showing the predominant genetic lineages/spoligotypes of 176 Nepalese Mycobacterium tuberculosis isolates.

Coloured branches represent Nepalese samples following Gagneux’s global phylogeography of MTBC (158) and black branches represent the in-house global Mycobacterium tuberculosis collection. PGGs are numbered following Sreevatsan’s classification (153). PGG1 outliers share one IS6110 copy with the PGG1 group. Those which contained one IS6110 copy but could not be assigned to any group (unassigned group in the figure) are shown within the green box.
5.3 Discussion

Thorne et al. showed that IS\textit{6110} FAFLP PCR can be used to delineate the phylogeny of MTBC as shared common fragments can identify the different lineages in a geographical location by comparison with a reference database collection (135). As very limited lineage information is available from strains in Nepal (up until 2012 there was none), the IS\textit{6110} method (published recently) was applied to map the IS\textit{6110} sites in H37Rv (179) and also carried out rpoB sequencing to further characterize strains from this important region.

Fifty five percent of the 176 Nepalese strains analysed belong to the CAS (36.6\%) and Beijing (18.8\%) modern genetic spoligotypes (PGG1). The remaining 24.4 \% of the samples belong to the PGG2 and PGG3 groups (Haarlem, LAM, S, X, T-Uganda and T). However, a limitation of this technique is its difficulty to characterise the samples with less than 4-5 copies of IS\textit{6110} as seen in the unassigned group (24.4\%) in figure 5-1, which can be overcome by the use of other typing techniques like MIRU-VNTR (146). The geographical position of Nepal is likely to have influenced this distribution of lineages, with a mixture of predominantly Beijing lineage from the North of the Himalayas and the CAS lineage from the south (190).

The fairly high percentage of mainly European lineages (Haarlem, LAM and T, 12.5\%) indicates that there has also been mixing of the different lineages over an extended time and that European travellers/migrants to South East Asia and Nepal may have transmitted European strains to the local population, probably due to the rise in globalisation leading to increased human migration between countries as reviewed by Soto (192). The IS\textit{6110} FAFLP data from our study supports the
hypothesis that the geographic location of Nepal is the key for the circulation of PGG1 TB lineages, CAS and Beijing, which were predominant in India and China respectively. This shows that the human migration from India and China into Nepal has helped *M. tuberculosis* lineages, especially Beijing and CAS, establish in this country. A large number of people flow between these regions due to various reasons, for example cheaper medical treatment facilities, work, study, trade, pilgrimage and cultural visits. It has been shown that there is a high prevalence of CAS lineages from North India circulating in Nepal due to the migration of Indian population from this region into Nepal for the purposes mentioned above (163,189,190). Likewise, the other predominant lineage, Beijing, found widely distributed in China and Tibet, is seen in Nepal (164,193,194). This simple and informative PCR-based molecular epidemiological technique might prove useful for the study of outbreaks of the disease and importantly also to detect cross-contamination between different strains or isolates in resource poor settings, aiding cluster investigation and possibly informing outbreak management.

### 5.4 Summary

To summarise, in this chapter I utilised the IS6110 FAFLP methodology successfully on DNA from TB samples from a resource poor setting, Nepal, and characterised the samples into different TB lineages or genotypes. This method also showed that the majority of the Nepal samples belonged to the PGG1 groups particularly, CAS and the Beijing.
Chapter 6  Rifampicin Resistance status in Nepalese TB isolates from clinical samples

(This chapter was published in the open access journal of Clinical Microbiology and Infectious Diseases (188). Also, this methodology was employed by another PhD student from Lahore University (Pakistan) supervised by me in a recent paper (195) whilst determining the rifampicin status from clinical samples in Pakistan).
6.1 Introduction

6.1.1 Drug Resistant TB in Nepal (adapted from NTC, 2015)

A recent Drug Resistance Survey in Nepal conducted between 2011 and 2012, has shown that DR-TB levels are increasing with nearly 9.3% of new patient resistant to at least one drug. The issue of increasing proportion of resistance to fluoroquinolones (26.4%) is a major public health concern in Nepal. Among MDR cases, 8% of the cases were found to be XDR due to the fluoroquinolones resistance. So to prevent the increased death rate due to XDR-TB in Nepal, the government is performing DST for second line drugs of all MDR-TB cases at the start of treatment (173). Early case finding is important to decrease human suffering, duration of disease and / or control of DR-TB, decrease financial burden and improve treatment outcome (23). As part of the STOP TB strategy in 2010, Nepal is successfully managing to control TB due to the presence of 581 microscopic centres, 2 solid culture laboratories including capacity for first line (FLD) drug susceptibility testing (DST), first line probe assay (LPA), one liquid culture and one second line drug (SLD) DST facility and 26 GeneXpert centres spread all over this mountainous country. This is in addition to 2 other culture and DST facilities (NTC and GENETUP) functioning in Kathmandu valley.

During the course of this research work the study of 261 Nepalese isolates by Malla et al., found that around 8.04% of the MDR-TB cases were new untreated cases. (191). This study reports that the two major TB genotypes circulating in Nepal are CAS and Beijing. According to the NTP’s annual report published in 2015, the national rate of MDR is low in new cases (2.2%)
and high in retreatment cases (15.4%). Since rifampicin resistance has been widely accepted as a ‘surrogate’ marker for MDR (53,196), it is useful to investigate the MDR-TB status in this set of samples and compare them to the results of Malla et al., as they are new untreated TB cases and, moreover, the phenotypic drug susceptibility test results were unavailable for these samples. It is also been reported that the predominant strains circulating in India and China, CAS and Beijing, are mostly MDR-TB and the same strains are also found in Nepal (164,189,190,193,194). It will be valuable to investigate if this finding is supported with this set of samples using the Rifampicin Resistance detection assay targeting the RRDR region of the rpoB gene.

6.1.2 Aims and Objectives

The aim of this chapter is to determine the status of rifampicin resistance as an indicator of Drug Resistance in the 176 bacterial DNA samples from Nepal.

6.2 Materials and Methods

6.2.1 rpoB Analysis

The 81bp Rifampicin Resistant Determining Region (RRDR) of the rpoB gene in the M. tuberculosis genome of all strains was sequenced using published primers (Arnold et al. 2005) and analysed in BIOEDIT software using ClustalW alignment parameters (see appendix 3). The PCR was carried out in a total volume of 50µl where 1ng of the DNA was added to the reaction containing 1xPCR reaction buffer, 1.5mM MgCl₂, 0.2mM dNTPs (Thermofisher, UK), 20µM each of both rpoB-RRDRforward (5’-
CGATCACACCGAGACGTTGA) and reverse primers (5’-GCCACGCTCACTGACAGACC) and 5U recombinant Taq polymerase (Thermofisher, UK). The following PCR conditions were carried out using a Veriti thermocycler (Applied Biosystems, UK): 94°C for 2 min followed by 35 cycles of 94°C for 30 sec, 60°C for 30 sec and 72°C for 1 min. Finally, an extension of 72°C for 10 min was performed before cleaning the products using AmpureXP magnetic beads (Beckman Coulter, UK) and sequenced using the forward primer, rpoB-RRDR forward.

6.3 Results

Of 176 DNA extracts analysed for rpoB mutations, seven samples (3.9%) had a single non-synonymous base change which would likely confer resistance (see table 6-1 and appendix 3). Six of these seven samples also showed a second base mutation in a codon triplet whereas sample N70 showed a first base mutation. There were no silent mutations observed in the RRDR region of any of these samples.

Table 6-1 List of mutations seen in rpoB Rifampicin Resistance- Determining Region (RRDR) of rifampicin resistant M. tuberculosis isolates from Nepal.

<table>
<thead>
<tr>
<th>Sample</th>
<th>Mutated locus</th>
<th>Nucleotide modification</th>
<th>Amino acid modification</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. N70</td>
<td>516</td>
<td>GAC &gt; TAC</td>
<td>(Asp&gt;Tyr)</td>
</tr>
<tr>
<td>2. N10</td>
<td>522</td>
<td>TCG &gt; TTG</td>
<td>(Ser&gt;Leu)</td>
</tr>
</tbody>
</table>
6.4 Discussion

Inferred Rifampicin resistance status for all the Nepal samples was determined using rpoB analysis. According to a recent study (191), fifty strains had any drug resistance and sixteen (6.1%) out of 261 isolates were MDR. Among the fifty any drug resistant strains, 29 cases were previously treated and twenty-one were new untreated cases (8.04% of 261 total strains and 12.8% of 164 new untreated cases). In this study, MDR-TB was tested by using rifampicin as the resistance marker and 7 isolates out of 176 were found (3.9%) from new untreated cases possessing drug resistance genotypes. However, their reported MDR percentage was based on the total number of isolates, of which 37.2% of isolates were from previously treated cases, which may enhance a probability of drug resistance development compared to untreated cases. Our results indicate that the prevalence of Rifampicin (RIF) resistant TB (surrogate marker for MDR) was higher than the nationally reported 2.2% MDR in new untreated cases. Further, our results are concordant with a recently conducted study by Creswell et al., where they have shown that the genotypic rifampicin resistance in newly

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>3. N25</td>
<td>526</td>
<td>CAC &gt; CTC</td>
<td>(His&gt;Leu)</td>
</tr>
<tr>
<td>4. N63</td>
<td>526</td>
<td>CAC &gt; CGC</td>
<td>(His&gt;Arg)</td>
</tr>
<tr>
<td>5. N34</td>
<td>531</td>
<td>TCG &gt; TGG</td>
<td>(Ser&gt;Trp)</td>
</tr>
<tr>
<td>6. N46</td>
<td>531</td>
<td>TCG &gt; TTG</td>
<td>(Ser&gt;Leu)</td>
</tr>
<tr>
<td>7. N62</td>
<td>531</td>
<td>TCG &gt; TTG</td>
<td>(Ser&gt;Leu)</td>
</tr>
</tbody>
</table>
diagnosed TB patients to be 3.3% in Nepal (198). In this study, the most common mutation site in the RRDR is at codon 531 concordant with the global data and parallels the findings of earlier studies (190,197). It also demonstrated the successful use of a surrogate rifampicin marker in analysis of MDR in *Mycobacterium tuberculosis* strains isolated from newly diagnosed primary TB patients originating from different regions of Nepal.

The major circulating genotypes in Nepal according to our study have been shown to be CAS and Beijing in chapter 5 table 5-1 which was concordant with Malla’s study (191) and Sharma’s study (189). Sharma et al. reported that the major circulating genotype in northern India was CAS and they were predominantly MDR (189). As Nepal shares an open border with India in the south, this had contributed to the spread of the CAS lineage into Nepal and thereby the MDR-TB (189,190). Again the same scenario is true for the Beijing lineages from China contributing to the MDR burden in Nepal (164,189,190,193). Despite the unavailability of the drug susceptibility tests, this rpoB assay was able to infer the rifampicin resistance status for this set of samples.

The combination of the rapid lineage specific PCR assay as discussed in chapter 4 along with the RIF resistant PCR assay will be more useful in resource limited setting as there is a reduced requirement for highly specialised equipment/infrastructure and staff for these molecular tests.
6.5 Summary

In this chapter by utilising the rpoB PCR assay targeting the 81bp RRDR region, seven samples (3.9%) of the 176 Nepal samples tested were found to be rifampicin resistant and is concordant with the other studies as described above.
Chapter 7  Final Discussion
7.1 Introduction

Control of tuberculosis in high burden countries is of paramount importance for a global TB prevention strategy to be successful. India and China, along with Indonesia, account for 45% of the new TB cases according to the global tuberculosis report by WHO (12). Despite recent advances in drug development and rapid diagnostic assays, the majority of TB diagnostic centres are still dependent on clinical diagnosis and interpretation of bacterial cultures to formulate the drug regimen and treat TB infection (12).

7.2 Discussion of findings in this study

The hypothesis of this research work is to substantiate the analysis of the position and copy number of IS\textit{6110}, a bacterial transposon, as a genomic tool to characterise the TB genetic lineages from Nepal, one of the low and middle income countries. TB strain typing can support molecular epidemiological investigations in controlling onward transmission and outbreaks, by identifying patients who are linked in the same chain of TB transmission (PHE, 2014).

7.2.1 Optimisation and Development of Insertion Element IS\textit{6110} Fluorescent Amplified Fragment Length Polymorphism (FAFLP) PCR

The main aim of this thesis was to develop a molecular tool box using IS\textit{6110} FAFLP PCR that can characterise \textit{Mycobacterium tuberculosis} complex strains genotypically and be used to understand the distribution and evolutionary relationships between different TB genetic lineages in resource poor settings in any geographical location. Various practical issues/artefacts
were addressed by using a recombinant Taq polymerase and by increasing the extension temperature from 60°C to 72°C and the extension time from 15 minutes to 1 hour, as shown in chapter 3. The increase in extension temperature reduced mis-priming of the polymerase, producing cleaner, more discrete products for more effective downstream analysis.

7.2.2 Mapping of Insertion sites (IS6110) in the M. tuberculosis H37Rv reference genome and rapid definition of genetic lineages (published work- (179))

To test the developed assay the reference TB genome H37Rv, was mapped in silico and the predicted fragment sizes compared with laboratory results. In this study laboratory results showed that sixteen IS6110 insertion sites were found in H37Rv (chapter 4 figure) using IS6110 FAFLP PCR, which correlates with the study by Philipp et al. in 1996 (187), who showed that H37Rv has sixteen IS6110 insertion sites by pulse field electrophoresis. Here, all sixteen fragments were identified using a fluorescently labelled dye as shown in table 4-2 in chapter 4, together with the position of in silico predicted insertion sites.

The IS6110 insertions have predilection to insert into mainly the surface proteins PPE (560.4B in table 4-2), cutinase1 (71.5Y in table 4-2) and those affecting glycerol metabolism such as IS986 and integrase (table 4-2) which was recently shown in Roychowdhury’s study, where it is shown that the majority of the transpositions of IS6110 occur in the surface proteins especially the PE/ PPE family of proteins (74). IS6110 has been shown by Alonso et al. to upregulate downstream genes with the help of the promoter
at its 3'end (199). One of the important properties of IS6110 is that it carries a promoter element upstream of the coding region, which has the potential to up-regulate expression of downstream genes (74,199,200). In a recent study, 178 unique genes belonging to mainly PPEs / transposases and oxidoreductases, were found to be carrying the IS6110 promoter element upstream of the coding regions (74). This study also found one instance of IS6110 insertion upstream of the trpD gene (ATP-binding cassette transporter), a metabolic enzyme essential for the bacterial survival in activated macrophages whilst colonising the lungs during infection (74,201). Further analysis revealed insertions in MDR-TB isolates, 75bp upstream of the phoP region (a transcriptional regulation factor), suggested to be necessary for M. tuberculosis virulence by Soto et al. in that the IS6110 insertion upregulates the over expression of phoP gene in M. bovis B strains (200). IS6110 insertions have been found upstream of early secretory antigenic target protein 6 (ESAT-6) that has been linked with mycobacterial virulence (202). It has been suggested that ESAT-6 interacts with the host protein Beta 2 microglobulin (β2M), a serum protein found associated with Major Histocompatibility Complex-1 (MHC-1), thereby inhibiting the expression of β2M- MHC-1 and down regulating the antigen presentation function of class I MHCs (202,203). The identification and confirmation of IS6110 insertions in silico into different genes like PPE, IS986, cutinase1, integrase, oxidoreductase and glycerolphosphodiesterase as shown in table 4-2 in chapter 4, indicates that IS6110 transposition might play a role in the pathogenesis of TB as suggested by other groups and the work presented in this thesis supports that theory.
In addition to the insertions at different genes, common IS6110 insertion points or fragments occur between lineages but some of them are unique to specific lineages as shown in chapter 5 and reported earlier by Thorne et al. (135). In this study, it has been shown that there are common IS6110 insertion points / fragments between isolates in one lineage, but not shared with other lineages (see table 5-1), for example Lineage 3, PGG1-CAS (78.4 G, 92.0B, 117.9R, 206.2G, 275.1R) does not share the same fragments as Lineage 4, PGG2-Haarlem (87.0Y, 89.4G, 148.7B, 300.2R, 445.7Y) or PGG3-T (81.3R, 192.4R, 360.0R,) This work was published in 2016 (188) and is concordant with the earlier study by Thorne et al. In the major study by Roychowdhury et al., a computational pipeline was developed to analyse the insertion sites of 1377 whole genome sequenced M. tuberculosis isolates, representing the 7 major global lineages, from all publically available datasets. Using density distribution studies, they found that different lineages have different copy numbers and insertion points and these insertions are unique to particular lineages (74). This property was observed in this study prior to Roychowdhury et al., as shown in chapter 5, where common fragments occur between lineages, together with some unique fragments within lineages. Part of the work in this thesis was cited by Roychowdhury et al. as being the first to identify this. In the East Asian- (Lineage 2, PGG1-Beijing) and Indian-East African (Lineage 3, PGG1- CAS) lineages, there is a high level of conservation in insertion sites as well as noticeable differences. Lineage 2, PGG1-Beijing isolates have extra IS6110 insertions in regions like Rv0001-Rv0002, Rv1371 and idsB, involved in biosynthesis of membrane ether-linked lipids (199). In Lineage 3 isolates, there are unique intergenic
regions mapping to regions such as Rv0395, Rv1504c and Rv3845-3846 that are not present in L2 (74).

Another point of discussion is whether the insertion sites of IS6110 occurs at insertional hotspots, i.e. independent insertion at the exact same point in different strains, with little or no phylogenetic signal, or inherited and relatively stable over time with significant phylogenetic signal. In this study, it was found that some Beijing strains have high copy numbers of IS6110 whereas the others have low copy numbers as shown in appendix 2. It might be possible that the rate of transposition is high in some of the strains compared to the others as reported earlier by Dale et al. (107). This was not studied in this thesis but could be investigated in future works. It also has been shown by Dale et al., that these insertions of IS6110 do not occur at hot spots but the transposition of IS6110 is slow in low copy number strains than high copy strains (107). Roychowdhury et al supported the previous work of this thesis in that the IS6110 hotspots observed are unique insertions subsequently inherited by daughter strains showing vertical transmission within genetic lineages, and confirmed by WGS analysis.

Following mapping of insertion sites, the IS6110 FAFLP assay paved the way to initiate the development of a rapid PCR assay that could identify the TB genetic lineages directly from DNA via identification of lineage specific IS6110 insertion points, without the need for performing FAFLP. This tool could rapidly assign lineages to the strain collection in any geographical setting with a basic PCR set up. The IS6110 FAFLP method, though relatively simple and robust, would still be challenging to implement in
settings with very poor resources, as the procedure has multiple steps, needs
good quality DNA to be extracted from isolates, uses relatively expensive
molecular biology reagents and equipment and also requires a cold chain
infrastructure. The development of lineage specific PCR was important so
that identification of genetic lineages directly from sputum samples would be
possible (not tested in this project). This assay will not directly affect the
treatment regimen of a patient but it will give an informed decision about the
TB lineage prevalent and could help to direct the control measures for that
particular lineage. It has been noticed that in certain regions like India,
Pakistan, China, certain TB lineages (CAS-Delhi in India and Pakistan,
Beijing in China) are more prevalent than other lineages (163,164) and does
not necessarily correlate with the increased drug resistance patterns seen in
these lineages (204) but could be to do with either a high density of human
population in these areas (205) or the ineffective or mismanagement of the
antibiotic usage in these countries (23).

In chapter 4, isolates representative of PGG3 and PGG1 lineages were
selected and tested. This PCR method with specific primers targeting unique
IS6110 fragments was able to identify PGG3 and PGG1 genetic lineages, by
detecting the unique 296bp product of PGG3 (360 R fragment inserted into
oxidoreductase as shown in chapter 4, table 4-2) and a 200bp product of
PGG1. Thus, the proof of principle was established and it would be useful to
design PCRs to common fragments in the remaining lineages.
7.2.3 Classification of Nepalese TB clinical isolates into different TB genetic lineages (published work - (188))

The main function of this genomic tool is to characterise the TB strains studied into different TB genetic lineages from any geographical location. This technique was applied on readily available TB DNA samples from Nepal as shown in chapter 5 as there was limited or no information present on the genetic diversity of the circulating TB lineages in Nepal when this study started in 2011. However, a study published by Malla et al., in 2012, described the initial characterisation of these lineages and the importance of understanding their distribution in Nepal (191). Nepal’s position geographically, sharing its borders between India and China, also proved relevant as one third of the total human TB cases are from these two countries (23). This emphasised that Nepal’s location would have acted as a mixing pot for the strains originating from infections in India and China, reflected in our results in chapter 6 that 55% of the strains were either Beijing or CAS belonging to PGG1 group (188).

The findings of this study corroborates Malla’s study (191) in that the majority of strains circulating in this geographical location in Nepal are from Lineages 2 and 3, PGG1- CAS and Beijing. These strains are predominantly found in India (CAS) and China (Beijing). Migration of the human population carrying these strains from these countries would have settled in Nepal and mixed with the local population thereby passing on these successful TB lineages as discussed in chapter 5. In chapter 5, IS6110 FAFLP derived phylogeny also clearly demonstrates that in the majority of Nepalese TB patients, although usually infected with strains belonging to PGG1, can also be infected with a
range of genetic lineages. This scenario clearly shows that various TB genetic lineages have entered Nepal at some point of time and the most successful lineages among them are the Beijing and the CAS lineages. This again focuses the discussion on human migration between the open borders of India in the south and China in the north of Nepal. In Northern India, the CAS- Delhi lineages are commonly seen and in China, Beijing lineages are predominant and this is evident in this study too (163,164,190). In chapter 5, the distribution of Beijing and CAS lineages are shown as 18.8 % and 36.6% respectively as seen in table 5-1. The perfect scenario is being offered by Nepal for the adaptation of *M. tuberculosis* strains by means of human migration for the purposes of tourism, religious travel like the pilgrimage to the famous religious shrines, work, and also cheaper treatment facilities (190).

In chapter 5, it was shown by IS6110 FAFLP that there are a high number of strains belonging to the CAS lineages compared to the Beijing lineages which was also the case in Malla’s study (191) where CAS lineages were more predominant than Beijing lineages. In Northern India, the predominant lineage seen is CAS and Beijing in China as shown in the figure 7-1 which shows that the geographical location of Nepal has played an important role in the TB lineages’ distribution in Nepal (189), by being a perfect geographical location for the intermingling of different lineages from these two high burden countries.
Figure 7-1 A. Distribution of different TB lineages in Nepal from this study and B. the distribution of different lineages seen in India, China and other countries near the Indian subcontinent (this figure has been adapted from (189)).
7.2.4 Rifampicin Resistance status in Nepalese TB isolates from clinical samples (published work - (188,195)).

Finally, as shown in chapter 6 by the rpoB PCR assay, seven of the 176 isolates were found to be likely resistant to rifampicin and all the seven mutations identified were in the RRDR region of rpoB. They were found to be non-synonymous thereby likely affecting the sensitivity of the strain. All 176 isolates in this study were primary untreated cases and this incidence of resistance seemed to be higher (3.9%) than the average national average of new, untreated cases (2.2%) (173,188). However the result was concordant with the recent study by Creswell et al., that showed the resistance as 3.3% (198) by analysing the resistance patterns using GeneXpert MTB/RIF kits. Again the increase in drug resistant TB in Nepal is probably due to the increase in drug resistant strains in India and China. The most frequently mutated codon is 531 followed by 526, 522 and 516 (see table 6-1), as observed by Poudel et al in his study (190). In Northern India, it has been reported that the majority of rifampicin mutation occur at codon 531 situated in the RRDR region (53,206) which reinforces the fact that the movement of human population helps the transmission of TB (159). One drawback of this technique is that it tests resistance to rifampicin by analysis of mutation in the RRDR only, even though sequencing of this region of rpoB has been suggested as a surrogate marker for MDR (53,196,207,208), it will not pick up 100% of mutations associated with resistance to this drug. This strategy will not be as useful for successful treatment of the patients if they are multi-resistant to other TB drugs and ideally the inclusion of a test to identify katG 315, the marker most commonly associated with isoniazid resistance, and
other front line drugs, could also be included in a PCR screen for optimum assay sensitivity directly from samples.

Nepal, the main setting of this study is no longer classified as a high burden TB country (172). However it is considered as a resource limited setting due to the poor infrastructure in remote villages and the lack of technologically advanced diagnostic assays and skilled labour for analysing the results, thus preventing routine molecular analysis. The priority of these settings is to identify tuberculosis infection by collecting samples for bacterial isolation and to inform treatment. The period of time taken to identify TB and start patient specific therapy is where these settings fail to keep up pace with a resourced setting. An assay such as IS6110 FAFLP PCR and, in particular the rapid PCR lineage detection method, although not a TB diagnostic test, could be useful to characterise TB isolates into different TB lineages directly from the clinical sample and thereby aid in contact tracing and indirectly in the control of outbreak. The resources needed are thermal cyclers, PCR reagents and trained staff for analysis and interpretation of results, which are already available in these settings. As contamination of cultures is potentially a major issue in outbreak investigation, the highly specific and sensitive nature of the assay could aid epidemiological investigations if implemented. The rapid PCR assay described in chapter 5 identifies lineage related strains and, with an expanded panel of PCRs for epidemiological investigations, control measures can be implemented within 2-3 days. If more information is needed on drug sensitivity then additional tests could also be added to this assay.
Novel diagnostic methods and assays are always needed for the effective and timely treatment of tuberculosis in resource limited settings. As the expenses incurred by the patients and their families accounted to 53% of annual household income per capita (209), the treatment of TB patients becomes extremely challenging in low income/high burden TB countries.

Different technologies evolved during the timeframe of this thesis work. WGS can now be performed relatively easily in high resource settings. Within a short period of time, strains could be identified and characterised using NGS technologies. In terms of rapid diagnosis however no other technology apart from single molecule WGS has shown promise in this area. Oxford Nanopore’s MinION sequencing could sequence TB genomic DNA directly from samples following extraction and depletion of human DNA. Currently however there is still an issue with the cost and availability of this technique in resource limited settings (148).

The advantages and disadvantages of the IS6110 FAFLP genomic tool are summarised in the table 7-1.
Table 7-1 Brief summary of the advantages and disadvantages of IS6110 FAFLP

<table>
<thead>
<tr>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>High resolution marker and precise as 0.5bp difference can be a different</td>
<td>Difficult to characterise strains with less than 5 IS6110 copies as shown in table 5-1 and figure 5-1 in chapter 5 where nearly 24% of the</td>
</tr>
<tr>
<td>lineage or strain (116.R is CAS and 117.1 R is Haarlem- table 5-1 and</td>
<td>Nepalese samples were ungrouped and warrants another high resolution marker like MIRU-VNTR.</td>
</tr>
<tr>
<td>figure 5-1)</td>
<td></td>
</tr>
<tr>
<td>Simple, reproducible and robust technique than RFLP.</td>
<td>Although more rapid than RFLP, it can take up to several days to perform the technique, get the results and analyse them but the rapid</td>
</tr>
<tr>
<td></td>
<td>lineage specific PCR described in chapter 4 overcomes this and the requirement for multiple enzymatic steps on extracted DNA from cultures.</td>
</tr>
<tr>
<td>Variations of the assay can suit both a high throughput modern lab and a</td>
<td>Basic equipment like PCR machines, heat blocks, gel electrophoresis and centrifuges are still needed, even if rapid tests are employed.</td>
</tr>
<tr>
<td>resource poor setting depending on the equipment available</td>
<td></td>
</tr>
<tr>
<td>Continuous screening of the samples using the technique would aid</td>
<td>Additional sequencing or other mutational studies like the rpoB PCR assay used in chapter 6 needed as this study does not detect any</td>
</tr>
<tr>
<td>contact tracing for immediate epidemiological intervention</td>
<td>mutational changes like SNP.</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>
7.3 Future work

A full toolbox of lineage specific PCRs could be developed and multiplex panels created. WGS may offer more information regarding the biological impact of IS$6_{110}$ insertion. Some insertion sites have remained in lineages for thousands of years and it is likely that they may confer a biological advantage. Knowing that they are present in specific locations could inform disease progression or pathogenicity in certain demographic groups. Characterisation of these and development of rapid PCRs will be useful for intervention. In terms of Nepal’s story, it would be interesting to study the pattern of resistance in Nepal’s TB strains and compare the resistance between the bordering countries. This will help in understanding the measures taken by the organism in developing resistance in a new niche. Also, it would be important to study the effect of different IS$6_{110}$ copy numbers in different TB lineages and their pathogenicity and if there is a relation between the two concepts.

7.4 Summary

The findings of all chapters have been summarised and tabulated below in table 7-2.
<table>
<thead>
<tr>
<th>Chapter Headings</th>
<th>Aims</th>
<th>Approach</th>
<th>Main Findings</th>
<th>Page Numbers</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Introduction</td>
<td>To set the scene for the thesis work</td>
<td>Extensive review of the literature</td>
<td>-Several epidemiological and genetic markers available to type TB.</td>
<td>1-62</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>-Among others, IS6110 found to contain phylogenetic signal and could therefore be used to develop a genomic tool box for resource poor settings.</td>
<td></td>
</tr>
<tr>
<td>2. General Materials and Methods</td>
<td>To document the standard methods used throughout the thesis.</td>
<td>Systematic approach from conception of idea, followed by development and implementation of the methods to achieve the IS6110 FAFLP PCR method using specific primers found to be the principal tool used in the project.</td>
<td></td>
<td>63-72</td>
</tr>
<tr>
<td>3. Optimisation and Development of IS6110 FAFLP PCR</td>
<td>To develop and standardise IS6110 FAFLP PCR that is both sensitive and specific.</td>
<td>Test the samples using different conditions and reagents to achieve the desired result.</td>
<td>The method was standardised and was ready to be validated.</td>
<td></td>
</tr>
<tr>
<td>-------------------------------------------------</td>
<td>---------------------------------------------------------------------------------</td>
<td>---------------------------------------------------------------------------------</td>
<td>---------------------------------------------------------------------------------</td>
<td></td>
</tr>
</tbody>
</table>
| 4. Mapping of Insertion sites (IS6110) in the M. tuberculosis H37Rv reference genome and rapid definition of genetic lineage | - To prove that IS6110 FAFLP PCR is accurate and could map insertion sites in the completely sequenced H37Rv strain.  
- To achieve the aim of developing a rapid lineage specific PCR | - H37Rv was subjected to IS6110 FAFLP and then insertion sites were mapped using two base selective primers.  
- Primers designed for the PGG1 and PGG3 strains were used in PCR reactions and tested for specificity | - Sixteen insertion sites were mapped onto the H37Rv genome, thereby proving the accuracy of the technique.  
- Proof of principle was established by detecting 100bp and 296bp products specific for PGG1 and PGG3 groups respectively |
| 5. Classification of Nepalese TB clinical | To characterise the TB strains from Nepal, a Bacterial DNA from all 176 TB isolates was subjected to | -All 176 isolates were delineated into different TB |

(73-91) | (92-107) (published) (179) | 108-118 (published) |
| isolates into different TB genetic lineages | resource poor setting, using IS\textit{6110} FAFLP PCR. | IS\textit{6110} FAFLP PCR and then characterised using BioNumerics software. | genetic lineages. -55\% of the strains belonged to PGG1 group, either Beijing or CAS strains. | (188) |
| 6. Rifampicin Resistance status in Nepalese TB isolates from clinical samples | To infer the drug-resistance status in the Nepalese TB samples. | Using \textit{rpoB} resistant PCR, all 176 samples were checked for inferred drug-resistance by observing the mutations in the RRDR region of the \textit{rpoB} gene. | -Seven isolates were found to have mutations at the normally reported mutation sites. -3.9\% of the strains were found to be MDR-TB -Prevalence of MDR-TB is high - Results concordant with previously published results | 119-125 (published) (188,195) |
| 7. Final Discussion | To contextualise the results obtained and make judgement of the aims of the thesis. | Summarise the results obtained and discuss the relevance of the results with respect to the aim and objectives of the thesis. | -Summarised the results of all the chapters.  
-Discussion of different results sections in detail.  
-Showed the efficacy of the IS6110 FAFLP PCR assay as a genomic tool in a resource poor setting. | 126-146 |
7.5 Conclusions

The aim of this thesis was to develop a genomic mapping tool using IS6110 FAFLP that could be used in resource poor settings as an assay for the characterisation of *M. tuberculosis* strains. This aim was achieved by first optimising the assay (chapter 3), testing it on H37Rv (chapter 4) by mapping the IS6110 insertion sites followed by the delineation of Nepalese TB strains into different TB lineages as shown in chapter 5. The IS6110 FAFLP PCR data led to the development of the rapid lineage specific PCR assay (chapter 4) that can identify the circulating TB genotypes in a specific geographic area. Though a rapid PCR assay was not developed for all the lineages or genotypes, the proof of principle was established by the identification of PGG1 and PGG3 strains. This work has successfully developed a simple, rapid and robust method to characterise different strains of *M. tuberculosis* complex that could aid and inform epidemiological investigation and intervention that could be used in resource poor settings.
References


7. Laënnec RTH. De l’auscultation médiate ou Traité du Diagnostic des


21. O’Connor JA, O’Reilly B, Corcoran GD, O’Mahony J, Lucey B.


34. WHO. Global Tuberculosis Report. 2012;


48. Fonseca JD, Knight GM, McHugh TD. The complex evolution of


54. Al-Mutairi NM, Ahmad S, Mokaddas E. Performance comparison of four methods for detecting multidrug-resistant *Mycobacterium*


67. Rodrigues LC, Diwan VK, Wheeler JG. Protective effect of BCG against tuberculous meningitis and miliary tuberculosis: a meta-


109. van Soolingen D, de Haas PE, Hermans PW, Groenen PM, van Embden JD. Comparison of various repetitive DNA elements as


fluorescent amplified-fragment length polymorphism analysis of
*Escherichia coli*: high-resolution typing method with phylogenetic

128. Meudt HM, Clarke AC. Almost Forgotten or Latest Practice? AFLP

129. Bensch S, Akesson M. Ten years of AFLP in ecology and evolution:

130. Goulding JN, Stanley J, Saunders N, Arnold C. Genome-sequence-
based fluorescent amplified-fragment length polymorphism analysis of

131. Randall MJ, Karst J, Pec GJ, Davis CS, Hall JC, Cahill JF. A molecular
identification protocol for roots of boreal forest tree species. Appl Plant

132. Desai M, Logan JM, Frost JA, Stanley J. Genome sequence-based
fluorescent amplified fragment length polymorphism of *Campylobacter
jejuni*, its relationship to serotyping, and its implications for

133. Smith D, Willshaw G, Stanley J, Arnold C. Genotyping of verocytotoxin-
producing *Escherichia coli* O157: comparison of isolates of a prevalent
phage type by fluorescent amplified-fragment length polymorphism and


146. Thorne N, Evans JT, Smith EG, Hawkey PM, Gharbia S, Arnold C. An 
IS6110-targeting fluorescent amplified fragment length polymorphism 
alternative to IS6110 restriction fragment length polymorphism analysis 
for Mycobacterium tuberculosis DNA fingerprinting. Clin Microbiol 
Infect. 2007b;13:964–70.

Comparison of four-colour IS6110-fAFLP with the classic IS6110-RFLP 
on the ability to detect recent transmission in the city of Barcelona, 

148. Desikan S, Narayanan S. Genetic Markers, Genotyping Methods & 
Next Generation Sequencing in Mycobacterium Tuberculosis. Indian J 

149. Mardis ER. Next-generation DNA sequencing methods. Annu Rev 

150. Singh P, Cole ST. Mycobacterium leprae: genes, pseudogenes and 

Non mycobacterial virulence genes in the genome of the emerging 

152. Supply P, Warren RM, Bañuls A-L, Lesjean S, Van Der Spuy GD,


171. Sulis G, Roggi A, Matteelli A, Raviglione MC. Tuberculosis:


Appendices

Appendix 1. Sequences (genes of interest, see Table 4-1) using the DNASTAR Lasergene8 software where the IS\textit{6110} insertion flanks the genome

1. IS\textit{6110} Sequence orientation \textit{per se} in the whole genome:

<table>
<thead>
<tr>
<th>Sequence Orientation</th>
<th>Primers</th>
</tr>
</thead>
<tbody>
<tr>
<td>a) 1543249 to 1543307 (58bp) to next \textit{TaqI} site</td>
<td>CCGCGCCGGCTGGCAACCGTTCCCGCT (27bp) ~102bp</td>
</tr>
<tr>
<td>b) 1989000 to 1989058 (58bp) to next \textit{TaqI} site</td>
<td>AGAGGACTTCAT (12bp) ~87bp</td>
</tr>
</tbody>
</table>
| c) 2431414 to 2431472 (58bp) to next \textit{TaqI} site | GGGCTTCCCCGAGACTGCAGTTCCCAAACGATGACGCCAAACAAAAAGCGGGA
                | CCGCGATGGCTGCCCCTGCGCTTGTTGCGTTGGGCTTTTACTCGT
                | (100bp) ~ 175bp                                                      |
| d) 3121821-3121879 (58bp) to next \textit{TaqI} site    | GTTTTGGGTCTGACGACTTCGCGCAGAGCACGTCTACCCACGGAGCGGTGAG
                | GTTGGTTTCTCCGCTCCCTTCTCGGGGTTTGGGTGCTGACGACCGAGAGCTG
                | GACCACATCAGCGATGCTGAGCTAGGTTTCCGTCCCTCCTCTCGGGGTTTGGGT
                | GTCTGACGACTTGTCTCAATCTGTCGCCGCTCTCGGCTGACGCTCAAAGTTTCC
                | GTCGCCCTCTCTGGGATTTGGGTGCTGACGACCGGATCAGCGCCAAGCCA
                | GTTAGCGGAAATCCAGTTTTCGTGCCCTCTCTCGGGGTTTGGGTGCTGACGACCTC
                | CCGGACCACCTGCAGCTCGCCGGCGTCCATCGCAGGTTTCCGTCCCTCCTCGGG
                | GTTTTGGGTCTGACGACTTCGCGCAGAGCACGTCTACCCACGGAGCGGTGAG
                | GGGTTTCCCGTCCCCCTCTCGGGGTTTGGGTGCTGACGACTTGGTCTGACGACCG
                | CTGACGGAAGTCTCGCGAGTATTCAAGGTTTCCGTCCCTCCTCGGGGTTTGGGT
                | TCCCGGAATTCACTCGCGTTATCTCAAGGTTTCCGTCCTCCTCGGGGTTTGGGT
2. Reverse Complement of the whole genome and the IS6110 Sequence orientation per se:

a) 520697 to 520755 (58bp) to next TaqI site

TCCGTATCGTGACCGAGCCGACCGTCCGTCGACACGTGGGGAGAGGGAATGGCAATGATGGT

(697bp) ~772bp

b) 701094 to 701152 (58bp) to next TaqI site

TGTCCATAACAGGGGCGTCCGTTGACCTGTTGGAGACCGCGCAGATGCACTGTGG

(136bp~201bp)

c) 1775899 to 1775957 (58bp) to next TaqI site

AAACTTCGGTAGCCAAAACCTGGGTAGCGGCAACATCGGCAGCACCAACGTGGGCAGCGGCAACATCGGCAGACACGAACTTCGGTAACGGAAACAACGGCAACTTCAACTTTGGTAGCGGCAATACCGGCAGTAACAACATCGGCTTCGGAAACACCGGCAGCGGGAATTTCGGTTTCTGGAAACACGGGCAACAACATCGGTATCGGGCTCACCGGCGATGGTCAGATCGGCATCGGCGGACTGAACTCGGGCAGCGGAAACATCGGTTTCGGGAACTCCGGCACCGGCAACGTAGGCTTCGGGAAACACAGGCTTTTT (486bp) ~561bp

d) 2415375 to 2415433 (58bp) to next TaqI site

GTTGCCGACCTTGTCTTCTGGACCTCTGGTGCGGGCTGACTGCCGCTCTA
CACTACGCGGCGGCGTTACGCGCTGCCACCTCACTCATGCGCGGCCGGCGCGG
TGCCGACC GG TAGCGT (121bp) ~196bp

3. IS6110 Sequence orientation per se in the whole genome
   (terminal G replaced)

   a) 2785912 to 2785969 (58bp) to next TaqI site

   Too short to be seen in vitro

   b) 3796355 to 3796412 (58bp) to next TaqI site

   AATCAAGTCCCCGCGTCCGTTGCGAATCGTGTTGTCATTGCGCGCGAACCTG
   TTTGGGAAGGCCGAATCGCACCCTCTCGGCTCGAT (89bp) ~164bp

4. Reverse Complement of the whole genome and the IS6110
   Sequence orientation per se (terminal G replaced)

   a) 858763 to 858820 (58 bases) to next TaqI site

   AATCCGTACAAACCGCGACGCGCGCGGAAACAGGCTGGACGGCGGAAATAGAGGCT
   GATCTGGACGTATGAGACCGACCGCAGCGGATGGGAGAGATTGCGACTGGTTAGCT
   ATAACCCAGAGGTCAGGTTTCTGGAACCGCCCCGGTGAGTCCCAGGTTCG
   AGCCT (214bp) ~289bp

   b) 1439367 to 1439424 (58bp) to next TaqI site

   GAAGATGAGCTCGCAAGCGTGGACGTGAGCGTCGCCCGTTGGACAAGTTGC
   CGCCGGGGCCATCCCCGCCACCGGCGGTCAGCGCTCAACCGTAAACC
   ATCCCCGTAGCAGCTTCGCTCCTCTGCTCGGGCTCGGTGT (139bp) ~214bp

   c) 1861462 to 1861519 (58 bases) to next TaqI site

   GGTGGCGCTGTCGTGGATCCTGCCTTGGGAGAGCTTTGCGGTGTTGCGGCTGGT
   CCGCTGGGCCAGACTGACGACCAGTGATTGTCCGGCATCGCGCGGGTACGC
GCGATTTCGGAGAACACCGTTCTGGCGATTACCAACGCAGTCGCAGCTGGTGCTGGATGGCATGTGGCTGACCGTCCAGGTCAGCAGCGATGGCGTGCCGGTGCTGTATCGTCCGTCCGATCTGGCAACGTTGACCGACGGCGCCGGCCCGGGTGA

**d) 2046062 to 2046119 (58 bases) to next TaqI site**

CGCGCGGCTACAAATAAAGCTTTGTATCACTTAGCTAGTGGGATCGTGTTGCT

(52bp) ~127bp

**e) 3522455 to 3522512 (58 bases) to next TaqI site**

CCTCCTTCATGAGGGGTGTGGAGGTGGCCGGTGCGCGGACACATCTC

AGTGGGGGCGAAAGCCAGCTCTCATATCAAGTCAGCGCCGCGGTCCCTTCTCA

CCTAGTGCGCGCAATACTGCTCAACGCAAGACCCCAACCCCAGGAGGACCGCGACGCA

AAGTGAGCGCAGCCGAACGATCCGGAAACAAACCAACCGCGCAGACCGCTCT

GTACACCGGACACCCACCCCTCCGCGGACACACACACACCGACACTGAGACATA

AAGGTCCCTTTCTGCGACGGCGTGT (285bp) ~360bp
Appendix 2: Nepal strains

N1-N176

Each strain has a list of fragments that are generated from differing insertion points of the \textit{IS6110} element. Each fragment is listed as a size in base pairs and can be one of four colours, \textit{R} = red, \textit{B} = blue, \textit{G} = green and \textit{Y} = yellow. For example 80.4Y is an 80.0 base pair fragment which is yellow.

If two strains share an identical \textit{IS6110} insertion site then they will each possess a fragment of the EXACT same size and colour. For example, below \textit{N36} and \textit{N34} share most of the same fragments, including: 81.0B, 100.0G, 124.9G, 140.4R, 197.2B, 207.1Y, 264.2G, 360.0R, 401.5B and 560.4B. The majority of these fragments are also shared by H37Rv and H37Ra but not all (i.e. they are not identical but very similar).

Unassigned

N102

84.4R, 106.6G, 312.8G

N107
84.4R, 106.6G, 312.8G

N104

84.4R, 106.6G, 312.8G

N25

178.1R, 237.8 G, 300.8G

N62

74.9Y, 78.2R

PGG3 T group (H37Rv-like)

N108

88.9Y, 106.2G, 179.1B, 183.6R, 254.3Y, 301.5G, 312.2G, 360.0R, 443.1R
N36
80.4Y, 81.0B, 83.8Y, 100.0G, 124.9G, 140.4R, 193.2R, 197.2B, 207.1Y, 264.2G, 360.0R, 401.5B, 560.4B

N34
81.0B, 100.0G, 124.9G, 139.1R, 140.4R, 192.4R, 197.2B, 207.1Y, 264.2G, 360.0R, 401.5B, 560.4B

N163

CAS group

N103
78.4G, 88.2R, 92.0B, 93.5B, 114.3Y, 127.7R, 148.4R, 207.1G, 249.6G, 275.1R, 394.7R, 434.3R

N101
74.4G, 92.0B, 115.6B, 116.8G, 123.9B, 127.7R, 136.0Y, 157.8G, 275.1R, 316.4R, 432.9R
N116


N109


N111

84.4R, 106.6G, 121.6B, 142.0B, 142.5Y, 145.1R, 147.6B, 159.5Y, 186.4B, 194.2R, 240.9Y, 285.6B, 312.2G, 326.4Y

N113

77.3R, 83.5G, 87.0Y, 96.7B, 97.8G, 100.8R, 105.5R, 112.4G, 124.0R, 127.1R, 133.7R, 169.7Y, 176.8Y, 220.5B, 320.5R, 376.5R, 432.9R

N114

84.9B, 102.1B, 102.5Y, 112.4G, 127.1R, 139.5R, 158.5Y, 181.2Y, 248.3R, 255.3G, 334.9R, 354.3B, 461.1G
N12
74.4G, 92.0B, 111.4G, 117.9R, 127.7R, 131.2Y, 170.7R, 206.2G, 232.3R, 275.1R, 393.0R, 432.9R, 560.0G

N50
78.4G, 92.0B, 117.9R, 127.7R, 131.2Y, 178.1R, 206.2G, 275.1R, 321.6R, 393.0R, 432.9R, 560.0G

N73
77.3G, 86.1G, 92.0B, 117.9R, 127.7R, 206.2G, 275.1R, 393.0R, 432.9R, 493.1G,

N36
78.4G, 83.3R, 92.0B, 127.7R, 188.9Y, 192.8R, 198.3B, 206.2G, 275.1R, 324.7G, 393.0R, 432.9R,

N37
78.4G, 92.0B, 117.9R, 127.7R, 206.2G, 275.1R, 393.0R, 432.9R,

N67
189
78.4G, 92.0B, 117.9R, 127.7R, 142.6R, 150.8R, 206.2G, 275.1R, 316.4R, 318.7Y, 393.0R, 432.9R, 493.1G,

N68

77.3G, 92.0B, 117.9R, 127.7R, 142.6R, 150.8R, 206.2G, 275.1R, 316.4R, 318.7Y, 393.0R, 432.9R,

N7

77.3G, 92.0B, 106.6G, 117.9R, 127.7R, 131.1G, 150.8R, 206.2G, 247.3R, 275.1R, 316.4R, 393.0R, 432.9R,

N18


N17

78.4G, 92.0B, 115.6B, 116.8G, 127.7R, 136.0Y, 157.8G, 275.1R, 316.4R, 393.0R, 432.9R,

N10

74.4G, 92.0B, 127.7R, 149.4Y, 150.8R, 206.2G, 230.1Y, 247.3R, 275.1R, 316.4R, 379.7R, 393.0R, 432.9R,
N72
78.4G, 92.0B, 127.7R, 149.4Y, 150.8R, 206.2G, 230.1Y, 247.3R, 275.1R, 316.4R, 379.7R, 393.0R, 432.9R,

N19
78.4G, 92.0B, 96.2R, 114.5B, 127.7R, 136.0Y, 148.0Y, 150.8R, 171.9R, 206.2G, 275.1R, 316.4R, 393.0R, 432.9R,

N6
77.3G, 88.2R, 92.0B, 93.5B, 114.3Y, 117.9R, 127.7R, 148.4R, 206.2G, 275.1R, 328.4G, 393.0R, 432.9R, 455.9G,

N3
77.3G, 88.2R, 92.0B, 93.5B, 114.3Y, 117.9R, 127.7R, 148.4R, 206.2G, 275.1R, 393.0R, 455.9G,

N91
78.4G, 88.2R, 92.0B, 93.5B, 114.3Y, 117.9R, 127.7R, 138.9G, 206.2G, 299.2Y,

N5
78.4G, 92.7B, 101.5Y, 107.1R, 116.7R, 117.9R, 183.8G, 206.2G, 251.8R, 275.1R, 321.6R, 393.0R, 432.9R,

N56

92.0B, 117.9R, 160.3Y, 206.2G, 318.0B, 352.9G, 393.0R, 432.9R,

N65

77.3G, 92.0B, 101.2B, 129.8G, 132.0R, 305.9R, 352.9G, 393.0R, 432.9R,

N29

78.4G, 81.0B, 92.0B, 100.5Y, 117.9R, 127.7R, 194.2R, 354.2Y,

N69

78.4G, 81.0B, 92.0B, 100.5Y, 117.9R, 127.7R, 194.2R, 266.1G, 275.1R, 310.6B, 316.5B, 353.7Y, 393.0R, 494.3B,

N48

94.8B, 117.9R, 275.1R, 296.9Y, 297.4R, 302.4G, 393.0R, 432.9R, 437.5B,
N58

94.8B, 117.9R, 275.1R, 296.9Y, 297.4R, 302.4G, 393.0R, 432.9R, 437.5B,

N54

78.4G, 81.7Y, 91.0R, 92.0B, 107.5Y, 112.4G, 127.7R, 142.6R, 150.8R, 216.3R, 318.7Y, 393.0R,

N81

78.4G, 81.7Y, 91.0R, 92.0B, 107.5Y, 114.3Y, 127.7R, 142.6R, 150.8R, 318.7Y, 345.5R,

N28

78.4G, 92.0B, 129.8G, 136.1G, 206.2G, 275.1R, 283.0R, 303.3G, 354.0G, 393.0R, 432.9R, 565.5R,

N93

78.4G, 92.0B, 122.8B, 136.1G, 176.8Y, 185.3G, 187.0Y, 275.1R, 393.0R, 432.9R, 563.6R

N26

193

N32

78.4G, 79.0Y, 92.0B, 127.7R, 206.2G, 239.3G, 256.2B, 273.1Y, 275.1R, 321.6R, 393.0R, 432.9R, 868Y

N45

87.3Y, 92.0B, 93.0G, 94.5R, 118.0G, 127.7R, 156.0Y, 206.5R, 279.5Y, 393.0R, 432.9R

N75

92.0B, 115.6B, 116.8G, 127.7R, 135.9R, 151.4R, 275.1R, 317.0R, 370.5G, 393.0R, 432.9R, 461.1G

N61

94.8B, 129.8G, 160.3Y, 206.2G, 275.1R, 318.0B, 353.5G, 393.0R, 432.9R

N78

91.5G, 100.3R, 120.0R, 132.2Y, 152.4R, 173.2B, 206.2G, 275.1R, 317.0R, 351.9R, 393.0R, 432.9R
N51
92.0B, 93.0G, 206.5G, 593.2G

N124

N135

N138

N139
N140


N141


N143


N144


Haarlem

N39
87.0Y, 148.7B, 149.6G, 151.4R, 224.3G, 265.6B, 300.2R, 311.0G, 445.7Y,

N87

87.0Y, 116.1R, 148.7B, 151.4R, 224.3G, 265.6B, 300.2R, 311.0G, 445.7Y, 787.0B

N110

87.0Y, 116.1R, 148.7B, 151.4R, 167.7G, 177.5R, 224.3G, 265.6B, 301.1R, 312.2G, 446.6Y, 787.0B

N106

87.3Y, 89.2B, 124.0R, 149.1B, 194.2R, 224.3G, 266.1B, 301.1R, 312.8G, 446.6Y

N137, N152, N153, N157, n160

X group

N27
311.0G, 300.2R
N77
80.4Y, 83.7Y, 301.1R 311.0G
N42
80.4Y, 83.7Y, 148.4R, 301.1R, 410.3R
N129

LAM group
N53
71.5Y, 96.7B, 105.2R, 116.1R, 121.2R, 161.2B, 164.6Y, 338.6Y, 441.9Y
N64
88.6G, 96.7B, 102.4G, 105.2R, 116.1R, 136.9Y, 227.5R, 431.9R, 444.1Y,

N9

71.5Y, 96.7B, 103.0B, 105.2R, 191.3G, 198.6R, 256.5R, 333.9G, 431.9R, 463.2R, 606.6B

N127

Beijing

N115

84.5Y, 95.7Y, 101.7B, 102.5Y, 112.4G, 115.6B, 127.1R, 139.1R, 180.7Y, 195.1Y, 199.5B, 254.8G, 334.9R, 354.3B, 389.8Y, 461.1G

N24
84.5Y, 95.7Y, 101.7B, 102.5Y, 103.0B, 112.4G, 127.1R, 131.2Y, 139.1R, 180.7Y, 234.8G, 254.8G, 265.7G, 309.9R, 332.4R, 353.5B, 389.8Y, 433.6B,

N8

80.0R, 84.5Y, 95.7Y, 101.2B, 102.5Y, 112.4G, 127.1R, 139.1R, 180.3R, 254.8G, 265.1G, 309.9R, 332.4R, 353.5B, 389.8Y, 432.7B, 458.8G,

N23

84.5Y, 95.7Y, 101.7B, 102.5Y, 112.4G, 127.1R, 135.5G, 139.1R, 140.7R, 180.7Y, 254.8G, 332.4R, 353.5B, 389.8Y, 433.6B, 460.3G,

N70

84.5Y, 93.5B, 95.7Y, 101.7B, 102.5Y, 112.4G, 127.1R, 139.1R, 147.7Y, 180.7Y, 201.1G, 254.8G, 266.1G, 309.9R, 332.4R, 353.5B, 389.8Y, 433.6B, 460.3G,

N40
84.5Y, 95.7Y, 101.7B, 102.5Y, 112.4G, 115.5B, 127.1R, 139.1R, 180.7Y, 254.8G, 265.7G, 309.9R, 332.4R, 353.5B, 389.8Y, 433.6B, 460.3G,

N33

84.5Y, 95.7Y, 101.7B, 102.5Y, 112.4G, 127.1R, 139.1R, 180.7Y, 309.9R, 332.4R, 353.5B

N82

84.5Y, 95.7Y, 101.7B, 102.5Y, 112.4G, 127.1R, 135.5G, 139.1R, 140.7R, 180.7Y, 254.8G, 332.4R, 333.5R, 353.5B, 389.8Y, 433.6B, 460.3G,

N35

84.5Y, 95.7Y, 101.7B, 102.5Y, 112.4G, 127.1R, 139.1R, 180.7Y, 246.5B, 254.8G, 266.1G, 309.9R, 332.4R, 353.5B, 389.8Y, 433.6B, 460.3G

N14
84.5Y, 95.7Y, 101.7B, 102.5Y, 107.5Y, 112.4G, 125.2Y, 127.1R, 139.1R, 180.7Y, 254.8G, 332.4R, 353.5B, 390.5Y, 434.5B, 460.3G, 
N74

84.5Y, 95.7Y, 101.7B, 112.4G, 127.1R, 139.1R, 180.7Y, 254.8G, 266.1G, 309.9R, 332.4R, 353.5B, 389.8Y, 433.6B, 460.3G, 
N76

95.7Y, 101.7B, 112.4G, 114.9B, 127.1R, 139.1R, 181.2Y, 195.1Y, 253.2B, 254.8G, 266.1G, 332.4R, 309.9R, 353.5B, 389.8Y, 433.6B, 460.3G, 
N90

95.7Y, 101.7B, 112.4G, 127.1R, 139.1R, 181.2Y, 253.2B, 254.8G, 266.1G, 309.9R, 332.4R, 353.5B, 389.8Y, 433.6B, 460.3G, 
N47

95.7Y, 101.7B, 112.4G, 127.1R, 129.8G, 182.8G, 254.8G, 353.5B, 433.6B, 460.3G, 
N55
95.7Y, 101.7B, 112.4G, 254.8G, 266.1G, 460.3G,

N59

95.7Y, 101.7B, 112.4G, 127.1R, 223.7G, 246.5B, 254.8G, 266.1G, 353.5B, 460.3G,

N120, N125, N136, N130, N131, N132, N133, N134, N149, N151, N154, N175

*M. africanum*

N31

87.1B, 88.6G, 134.7Y, 140.4R, 152.4R, 183.6R, 195.1R, 195.8R, 239.3Y, 287.2R, 302.5R, 598.0R

N164

PGG1 Outliers
N13
87.4R, 90.2B, 95.6B, 96.4G, 187.2R, 219.8R, 311.0G, 374.6R, 445.3G,

N20
95.6B, 311.4G

N21
137.1R, 150.8G, 224.6R, 310.5G, 435.4Y

N38
107.3G, 117.8Y, 137.1R, 150.8G, 166.3R

N22
190.3R, 224.6R, 239.9B, 253.4G, 263.2G, 309.9G, 404.3G,

N4
204
171.9B, 196.1G, 225.8Y, 253.4G, 262.7G, 272.7R, 308.7G, 356.4Y, 403.5G, 505.2Y,

N121, N122, N123, N142, N150, N159, N162, N166, N169, N170, N171, N172
Appendix 3: rpoB mutations seen in the RRDR region using Clustal W alignment in BioEdit sequence alignment software
|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
Appendix 4: Publications from this PhD work
Appendix 5: IS6110 FAFLP raw data of 176 Nepal samples _Stored in CD and attached at the back of thesis
Insertion Element IS6110 based characterisation of Nepalese tuberculosis strains into different genetic lineages

Kartyk Mogeranadri¹, Dunstan Rajendram¹, Saraswoti Khadge¹, Pam Sonnenberg¹, Ibrahim AbuBakar⁴⁺, Timothy D. McHugh¹, Murdo Macdonald⁴, Deanna A. Hagge⁴ and Catherine Arnold¹⁺

¹Genomic Services and Development Unit, Public Health England, London, UK
²Mycobacterial Research Laboratories, Anandaban Hospital, Kathmandu, Nepal
³Research Department of Infection & Population Health, University College London, London, UK
⁴Respiratory Diseases Department, TB section, Public Health England, London, UK
⁵Centre for Clinical Microbiology, University College London, London, UK

Abstract

Nepal is geographically located between India and China, a region containing significant Tuberculosis (TB) and Multi-Drug Resistance (MDR-TB) burdens. However, limited information is available on the phylogenetic diversity of Mycobacterium tuberculosis (Mtb) in Nepal. To gain further insight into the diversity of Mtb in Nepal, consecutive clinical samples from 176 newly diagnosed pulmonary tuberculosis patients were collected from two hospitals in Nepal. Insertion Site IS6110 Fluorescent Amplified Fragment Length Polymorphism (FAFLP) PCR and rpoB sequence analysis were carried out on genomic DNA extracts of cultured strains to assign them to accepted genetic lineages and identify MDR-TB. In this study, the IS6110 based characterisation showed a prevalence of 36.36% Central Asian Strain (CAS), 18.75% Beijing, 7.95% Haarlem, 3.97% X, 2.2% each of Latin American Mediterranean (LAM), T-Uganda and T, 1.7% S and 24.4% were unassigned. Further, 3.9% of total M. tuberculosis isolates were of rifampicin resistant genotypes thus indicating that the prevalence of MDR could be higher than the country wide prevalence of MDR among new TB cases (2.2%) as reported by the national drug resistance survey carried out in 2011/2012.

Introduction

TB is ranked as the sixth leading cause of death among the top 20 causes of death in Nepal. According to National Tuberculosis Control Programme (NTCP) in Nepal, in 2014 37,025 TB cases were registered and among them 15,947 (43%) cases were new sputum smear positive TB cases. It was estimated by WHO [1] that 4.6 (2.1-7.5) thousand people in Nepal died from TB in 2014. Even though short course TB drug treatment regimen could cure around 89% of cases, TB mortality was still unacceptably high in Nepal. Since 2006, the STOP TB strategy has been adopted by NTCP. However Drug resistant TB (DR-TB) still threatens national TB control and is a major public health concern. The proportion of MDR-TB cases in new cases was 2.2% and retreatment cases were 15.4%. Even though the Millennium Development Goal (MDG) to halt and reverse TB incidence has been achieved in all six WHO regions, work remains to be done to prevent the deaths from this dreadful disease [1].

The identification of the number and position of Insertion Sequence IS6110 elements in the Mtb genome has been widely used as a genomic tool for the rapid fingerprinting of isolates of Mycobacterium tuberculosis complex (MTBC) [2]. IS6110 based Restriction Fragment Length Polymorphism (RFLP) is considered as the ‘gold standard’ typing method for strains with more than five copies [3-5]. As IS6110 transposition is among the first genetic changes to occur in strains from a transmission chain [6], this marker has also been used for outbreak analysis [5].

Modification of the conventional IS6110 typing method, using differentially labelled primers has allowed characterisation of Mtb isolates into the key genetic lineages more rapidly than traditional methods [7]. This approach can be facilitated with automation, which enables this technique to be performed in a high throughput setting. The fragment patterns generated indicates both copy number and insertion site of IS6110 in the genome [8,9]. The patterns generated correlates directly with other independent markers and can be used for transmission investigation locally/internationally. Specific fragments are common in genetically related lineages and do not occur in other groups (e.g. spoligotype groups such as Beijing and the Euro-American lineage which contain the Latin American Mediterranean (LAM), Haarlem, S, T and X spoligotype groups). The patterns generated correlate directly with other independent markers and can be used for transmission investigation locally/internationally. Principal Genetic Groups (PGGs) can be assigned to Mtb strains based on the combination of polymorphism located at katG codon463 and gyrA codon95 in the respective genomes [10] or spoligotypes [11] or global phylogeny classification based on whole genome sequences [12].

Correspondence to: Catherine Arnold, Genomic Services and Development Unit, Public Health England, London, UK, Tel: 0208-327-6068; E-mail: Catherine.arnold@phe.gov.uk

Key words: tuberculosis, IS6110, FAFLP, rifampicin resistance, TB lineages

Received: June 01, 2016; Accepted: June 27, 2016; Published: June 30, 2016
Limited data are available on the characterisation of Mtb strains and genotypes circulating in Nepal. A key factor is the geographical location of Nepal, interlocked between China and India, two countries, which together account for approximately a third of annual global new cases (11% and 24%, respectively) [1]. A recent study of 261 Nepalese isolates found any drug resistance (any drug resistance has been defined as resistance to isoniazid, rifampicin, streptomycin, ethambutol, fluoroquinolones, and/or aminoglycosides) in 12.8% of Mtb strains that were new untreated cases, with the most frequent lineages reported as CAS/Delhi (40.6%), East Asian (including Beijing) (32.2%), Euro-American (15.7%) and Indo-oceanic (11.5%) [13]. To gain further insight into the characteristics and diversity of Mtb in Nepal, our study aimed to categorise isolates for the first time using IS6110 FAFLP PCR and to assign them to different genetic lineages. Second, the level of MDR would be characterized in the population.

Methods

Strains

Sputum samples from 176 consecutive new TB patients over one year were collected between 2007 and 2008 and cultured alongside routine diagnostics from two Nepalese tuberculosis reference centres located in the Kathmandu valley: the National Tuberculosis Centre (NTC) and the German Nepal Tuberculosis Project (GENETUP). The patient population represented local and referred cases from across Nepal. Bacterial genomic DNA from isolated strains was extracted by the Cetyltrimethylammonium Bromide (CTAB) method [18] at the Mycobacterial Research Laboratories (MRL) in Anandaban Hospital. Informed consent was not required at the time of this study, as samples were collected with routine clinical care and all patient identifiers were anonymized; however, all patients were provided an explanation and were only included upon provision of verbal informed consent. Study procedures were reviewed and approved by NTC and GENETUP. The results for the drug sensitivity tests were unavailable during the entire duration of this study.

IS6110 FAFLP PCR, Fragment Sizing and Analysis

Genomic DNA was digested with the restriction enzymes MseI and TaqI followed by ligation with double stranded TaqI restriction site specific adaptors. The adaptor ligated DNA was amplified following previously published PCR conditions using four fluorescently labelled adaptor specific Taq forward primers: 5'-CGATGAGTCCTGACCAGA*C/T/G each labelled with a single unique selective nucleotide at the 3' end and an IS6110 sequence specific reverse primer- 5'-CTGACATGACCCGACCATCCTTTT [9]. In a total volume of 20 µl, 1 µl of the adaptor ligated DNA was added to the reaction containing 1X reaction buffer, 1.5 mM MgCl₂, 0.2 mM dNTPs (Invitrogen, UK), 1 µM of labelled Taq I forward primer, 1 µM of IS6110 reverse primer and 1 U of recombinant Taq polymerase (Invitrogen, UK). The following PCR conditions were carried out in a Veriti thermocycler (Applied Biosystems, UK): 94°C for 15min followed by 35 cycles of 94°C for 20s, 66°C for 30s and 72°C for 2 min with the 66°C annealing temperature reducing by 1°C every cycle for nine cycles and the last 25 cycles at 56°C. Finally, an extension of 72°C for 60min was carried out before further manipulations. The fragments were separated on an ABI genetic analyser 3730XL (Applied Biosystems, UK), sized using PeakScanner v1.0 software (Applied Biosystems) and identified using their fluorescent tag (Figure 1). The four-dye FAFLP data collected from the different profiles were then recorded and compared with a reference collection of Mtb isolates [19] using BioNumerics software v6.1 (Applied Maths Inc., Belgium). Fragments common to different lineages (defined as being present in >50% of strains in a particular genetic lineage) were recorded for each Nepalese strain and compared with a fully characterised global collection as detailed by Thorne et al. [20]. These data were then used to build a dendrogram using the Dice coefficient of similarities to compare the similarity matrix and Unweighted Pair Group Method with Arithmetic Mean (UPGMA) derived cluster analysis with cophenetic correlation for the branch quality.

rpoB analysis

The 81 bp Rifampicin Resistant Determining Region (RRDR) of the rpoB gene of all strains were sequenced using published primers [21] and analysed in BIOEDIT software using ClustalW alignment parameters. The PCR was carried out in a total volume of 50 µl where 1 µl of the DNA was added to the reaction containing 1xPCR reaction buffer, 1.5 mM MgCl₂, 0.2 mM dNTPs (Invitrogen, UK), 20 µM each of both rpoB-RRDRforward (5'-CGATCACACCGCAGACGTTGA) and reverse primers (5'-GGCCAGCTCAGTGAGCAGGTC) and 5U recombinant Taq polymerase (Invitrogen, UK). The following PCR conditions were carried out using a Veriti thermocycler (Applied Biosystems, UK): 94°C for 2 min followed by 35 cycles of 94°C for 30 sec, 60°C for 30 sec and 72°C for 1 min. Finally, an extension of 72°C for 10 min was performed before cleaning the products using AmpureXP magnetic beads (Beckman Coulter, UK) according to the manufacturer’s protocol and sequencing using the forward primer, rpoB-RRDR forward.

![Figure 1](image-url)
Results

Analysis of Data using BioNumerics software v6.1

Of the 176 DNA extracts from isolates analysed, the majority of the samples (97 (55.4%) belonged to either the spoligotype-defined Central Asian Strain (CAS) lineage (64 i.e., 36.6%) or the Beijing lineage (33 i.e., 18.8%) grouping under PGG1 and the rest of the samples group under either PGG2 (1.7% S, 3.97% X, 7.95% Haarlem and 2.27% LAM, 2.27% T-Uganda) or PGG3 (2.27% of T) (Table 1). Forty three samples (24.4%) grouped under “unassigned” group. Common fragments seen were exactly the same as the earlier published report by Thorne et al., (2011) except for an additional fragment, 78.4 G, for the CAS lineage. A dendrogram was generated using only the IS6110 FAFLP data (Figure 2) confirming again the above mentioned lineages in relation to the PGGs.

rpoB Analysis

Of 176 DNA extracts analysed for rpoB mutations, seven samples (3.9%) had a single non-synonymous base change which would likely confer resistance to rifampicin (Table 2). Six of these seven samples showed a second base mutation in a codon triplet whereas sample N70 showed a first base mutation.

Discussion

It has been demonstrated previously that IS6110 FAFLP PCR can be used to delineate the phylogeny of MTBC as shared common fragments can determine the different lineages in a geographical location by comparison with a reference database collection [7]. As limited lineage information is available from strains in Nepal, we have applied the IS6110 method published recently on mapping the IS6110 sites in H37Rv [9] and also carried out rpoB sequencing to further characterize strains from this important region.

Fifty five percent of the 176 Nepalese strains analysed belong to the CAS (36.6%) and Beijing (18.8%) modern genetic spoligotypes (PGG1). The remaining 24.4% of the samples belong to the PGG2 and PGG3 groups (Haarlem, LAM, S, X, T-Uganda and T). However, a limitation of this technique is its difficulty to characterise the samples with less than 4-5 copies of IS6110 as seen in the unassigned group (24.4%) in figure 2, which can be overcome by the use of other typing techniques like Mycobacterial Interspersed Repetitive Units- Variable Number Tandem Repeats (MIRU-VNTR) [20]. The geographical position of Nepal is likely to have influenced this distribution, with a mixture of predominantly Beijing lineage from the North of the Himalayas and the CAS lineage from the south [22].

Table 1: Common fragments identified using IS6110 FAFLP PCR in TB genetic lineages between the 176 bacterial DNA isolates in Nepal.

<table>
<thead>
<tr>
<th>PGG/spoligotype/ sub-lineage</th>
<th>Common fragment sizes</th>
<th>No. Of Nepal strains</th>
</tr>
</thead>
<tbody>
<tr>
<td>PGG1-CAS</td>
<td>78.4G, 92.0B, 117.9R, 286.2G , 275.1R</td>
<td>64/176</td>
</tr>
<tr>
<td>PGG1-Beijing</td>
<td>101.7B, 102.5Y, 139.1R, 180.7Y, 254.8G, 332.4R, 353.5B</td>
<td>33/176</td>
</tr>
<tr>
<td>PGG2-Haarlem</td>
<td>87.0Y, 89.4G, 148.7B, 300.2R (H/X), 445.7Y</td>
<td>14/176</td>
</tr>
<tr>
<td>PGG2-LAM</td>
<td>71.5Y, 105.2R, 116.1R</td>
<td>4/176</td>
</tr>
<tr>
<td>PGG2-S</td>
<td>88.0G, 112.9R, 217.2R, 445.3G</td>
<td>3/176</td>
</tr>
<tr>
<td>PGG2-X</td>
<td>83.8Y, 300.2R (H/X)</td>
<td>7/176</td>
</tr>
<tr>
<td>PGG2-T Uganda</td>
<td>88.9Y, 119.5G, 227.7Y, 228.4Y, 266.8R</td>
<td>4/176</td>
</tr>
<tr>
<td>PGG3- T</td>
<td>81.3R, 192.4R, 360R</td>
<td>4/176</td>
</tr>
<tr>
<td>Ungrouped</td>
<td>43/176</td>
<td></td>
</tr>
</tbody>
</table>

Where B-Blue coloured fragment R-Red coloured fragment G-Green coloured fragment and Y-Black/ Yellow coloured fragment seen in the electropherogram. PGG represents Principal Genetic Groups according to Sreevatsan et al. [10], spoligotypes follow spolDB4 classification[4] and sub-lineages are grouped following Gagneux’s classification [8].

Table 2: List of mutations seen in rpoB Rifampicin Resistance-Determining Region (RRDR) of rifampicin resistant M. tuberculosis isolates from Nepal.

<table>
<thead>
<tr>
<th>Sample</th>
<th>Mutated locus</th>
<th>Nucleotide modification</th>
<th>Amino acid modification</th>
<th>FAFLP derived</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. N70</td>
<td>516</td>
<td>GAC &gt; TAC</td>
<td>(Asp&gt;Tyr)</td>
<td>PGG1-Beijing group</td>
</tr>
<tr>
<td>2. N10</td>
<td>522</td>
<td>TCG &gt; TTG</td>
<td>(Ser&gt;Leu)</td>
<td>PGG1-CAS group</td>
</tr>
<tr>
<td>3. N25</td>
<td>526</td>
<td>CAC &gt; CTC</td>
<td>(His&gt;Leu)</td>
<td>PGG3-like</td>
</tr>
<tr>
<td>4. N63</td>
<td>526</td>
<td>CAC &gt; CGC</td>
<td>(His&gt;Arg)</td>
<td>PGG3-like</td>
</tr>
<tr>
<td>5. N34</td>
<td>531</td>
<td>TCG &gt; TGG</td>
<td>(Ser&gt;Thr)</td>
<td>PGG3- T group (H37Rv-like)</td>
</tr>
<tr>
<td>6. N46</td>
<td>531</td>
<td>TCG &gt; TTG</td>
<td>(Ser&gt;Leu)</td>
<td>PGG3 (H37Rv-like)</td>
</tr>
<tr>
<td>7. N62</td>
<td>531</td>
<td>TCG &gt; TTG</td>
<td>(Ser&gt;Leu)</td>
<td>Ungrouped</td>
</tr>
</tbody>
</table>

* PGG represents Principal Genetic Group.
percentage of mainly European lineages (Haarlem, LAM and T, 12.5%) indicates that there has also been mixing of the different lineages over an extended time and that European travellers/migrants to South East Asia and Nepal may have transmitted European strains to the local population. According to Malla et al. [13], fifty strains had any drug resistance and sixteen (6.1%) out of 261 isolates were MDR. Among the fifty any drug resistant strains, 29 cases were previously treated and twenty-one were new untreated cases (8.0% of 261 total strains and 12.8% of 164 new untreated cases). In this study, MDR-TB was tested by using rifampicin as the resistance marker and 7 isolates out of 176 were found (3.9%) from new untreated cases possessing drug resistance genotypes. However, their reported MDR percentage was based on the total number of isolates, of which 37.2% of isolates were from previously treated cases, which may enhance a probability of drug resistance development compared to untreated cases. Our results indicate that the prevalence of Rifampicin (RIF) resistant TB (surrogate marker for MDR) was higher than the nationally reported 2.2% MDR in new untreated cases. Further, our results are concordant with a recently conducted study by Creswell et al. [23], where they have shown that the genotypic rifampicin resistance in newly diagnosed TB patients to be 3.3% in Nepal [23].

Conclusions

The IS6110 FAPLFP data from our study reiterates the fact that the geographic location of Nepal is the key for the circulation of PGG1 TB lineages, CAS and Beijing, which were predominant in India and China respectively. Further the RRDR study correlates with the recent work by Creswell et al. showing that prevalence of MDR-TB may be marginally higher than the national average in new untreated TB cases. As the monitoring of TB is important in Nepal, this simple and informative PCR-based molecular epidemiological technique would prove useful for the study of outbreaks of the disease and also to detect cross-contamination between different strains or isolates in resource poor settings. The most common mutation site in the RRDR is at codon 531 and parallels the findings of earlier studies [21,22].

Authorship

All authors mentioned above gave substantial intellectual contribution to this manuscript.

Acknowledgments

We thank gratefully to the staff of the NTC and GENETUP from the Kathmandu valley in Nepal. We would like to thank all staffs from MRL, Ananadaban Hospital in Lalitpur, Nepal for their technical help. We are also grateful to Dr. Ali Al-Shahib from Bioinformatics Department (PHE, UK) and Mr. Jon White from Medical Illustration Department (PHE, UK) for their technical expertise.

Funding information

The authors received no specific funding for this work.

Competing interest

The authors declare that they have no competing interests.

References

8. Borrell S (2009) Comparison of four-colour IS6110-FAPLFP with the classic IS6110-RFLP on the ability to detect recent transmission in the city of Barcelona, Spain. *Tuberculosis* (Edinburgh, Scotland) 89: 233-237. [Crossref]
Mycobacterium tuberculosis from isolates or directly from sputum. *Clin Microbiol Infect* 11: 122-130. [Crossref]


**Copyright:** ©2016 Moganeradj K. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.
Sequence analysis of the rifampicin resistance determining region (RRDR) of rpoB gene in multidrug resistance confirmed and newly diagnosed tuberculosis patients of Punjab, Pakistan

Salma Hameed1,2, Kartyk Moganeradj3*, Nasir Mahmood4, Timothy D. McHugh2, Muhammad Nawaz Chaudhry1, Catherine Arnold3

1 College of Earth and Environmental Sciences, University of the Punjab, Lahore, Pakistan, 2 Centre for Clinical Microbiology, Division of Infection and Immunity, Royal Free Campus, University College London, London, United Kingdom, 3 Genomic Services and Development Unit, Public Health England, Colindale, United Kingdom, 4 Department of Biochemistry; Department of Human Genetics and Molecular Biology, University of Health Sciences, Lahore, Pakistan

* Kartyk.moganeradj@phe.gov.uk

Abstract

Molecular screening of new patients suspected for TB could help in the effective control of TB in Pakistan as it is a high TB burden country. It will be informative to understand the prevalence of multi drug resistance for a better drug regimen management in this geographical area. The Rifampicin resistance determining region (RRDR) sequencing was used to identify mutations associated with drug resistance in DNA extracts from 130 known multidrug resistant (MDR) cultured strains and compared with mutations observed in DNA extracts directly from 86 sputum samples from consecutive newly diagnosed cases in Lahore, Pakistan. These newly diagnosed samples were positive for smear microscopy, chest X-ray and presumed sensitive to first line drugs. In the known MDR group the most frequent mutations conferring resistance were found in rpoB531 (n = 51, 39.2%). In the newly diagnosed tuberculosis group with no history of MDR, mutations in rpoB531 were seen in 10 of the samples (11.6%). Collectively, all mutations in the RRDR region studied were observed in 80 (61.5%) of known MDR cases and in 14 (16.3%) of the newly diagnosed cases. Using the RRDR as a surrogate marker for MDR, sequences for the newly diagnosed (presumed sensitive) group indicate much higher levels of MDR than the 3.9% WHO 2015 global estimate and suggests that molecular screening directly from sputum is urgently required to effectively address the detection and treatment gaps to combat MDR in this high burden country.

Introduction

Pakistan is among the top 20 countries with a high TB and MDR-TB burden [1]. Chest X-ray, acid fast bacilli (AFB) smear microscopy and culture on Lowenstein-Jensen (LJ) media are the
conventional methods of investigation for tuberculosis [2] but require additional analysis to define the species of Mycobacteria as well as the mechanism of drug resistance. *M. tuberculosis* drug resistance detection using conventional methods is by culture of bacilli on a medium containing antibiotic and can require several weeks. However, with the development of rapid molecular methods it is possible to detect mutations in genes associated with resistance in a much shorter time [3]. Resistance to two first line predominant anti-TB drugs i.e., isoniazid (INH) and rifampicin (RIF), is termed as ‘multidrug resistance tuberculosis’ [4]. Molecular methods for this are diverse and each method has its benefits and drawbacks; for example PCR-RFLP [5] and allele-specific PCR [6]. Several molecular techniques have been evolved to detect the gene mutation related to resistance. These include hybridization methods; single strand polymorphism, DNA sequencing and other PCR based methods [5, 7, 8]. Multiplex Allele Specific (MAS) PCR, a rapid and cost-effective method simultaneously detects INH, RIF and Ethambutol (EMB) resistance associated genetic mutations [9]. PCR technology can provide many advantages over traditional techniques. Many PCR tests can be rapidly performed and interpreted on the same day of submission of samples. A major advantage of PCR over traditional techniques includes the ability to rapidly identify organisms that are difficult to culture and the DNA of interest can be amplified with the DNA from just one cell. The sensitivity of PCR is also its major disadvantage since very small amounts of contaminating DNA (from a different sample) can also be amplified. One major limitation of PCR is that prior information about the target sequence is necessary in order to generate the primers that will allow its selective amplification [10]. Whole Genome Sequencing (WGS) sequences the whole genome rather than specific genes. So, drug resistance prediction from the whole genome sequence is possible using publically available software which rapidly analyses all known gene targets and identifies mutations associated with resistance thus enabling targeted treatment [11] but requires culture and is currently prohibitively expensive for high burden countries. Identification of MDR-TB is a crucial step as treatment of multi drug resistant tuberculosis (MDR-TB) is a considerable challenge. Globally in 2015, an estimated 3.9% (95% confidence interval [CI]: 2.7–5.1%) of new cases and 21% (95% CI: 15–28%) of previously treated cases had MDR/RR-TB [1]. Resistance to rifampicin is the result of mutations in the rifampicin resistance determining region (RRDR) of *rpoB*, particularly mutations at codons 516, 526 and 531. MDR-TB is defined as resistance to rifampicin and isoniazid, the two most effective anti-TB drugs. In December 2010, WHO recommended the use of the GeneXpert MTB/RIF to detect and infer resistance to rifampicin directly from sputum [12]. In May 2016, WHO issued guidance that “people with TB resistant to rifampicin, with or without resistance to other drugs, should be treated with an MDR-TB treatment regimen.” Together with MDR-TB, these are referred to as MDR/RR-TB.

This work was carried out in Pakistan to characterize mutations associated with rifampicin resistance directly from sputum samples from newly diagnosed (ND) patients with no history of drug resistance and identify their key risk factors in this setting. Identifying resistance in presumed resistant samples and inferring resistance profiles directly from sputum may enable a better tailored drug regimen where possible. It will also inform patient management more rapidly and consequently reduce the rate of onward transmission of MDR tuberculosis in high burden countries such as Pakistan.

**Materials and methods**

**Sampling**

Tuberculosis patients attending the Ghulab Devi Chest Hospital, Lahore, Pakistan in collaboration with University of Health Sciences, Lahore were enrolled in this study over 18 months.
between 2013 and 2015, based on the following inclusion criteria. For the MDR group, patients diagnosed previously with TB, and with a history of resistance to first line anti-tuberculosis drugs, were included. Sputum samples were taken from this group and culture and Drug Susceptibility Testing (DST) were performed. The second group of patients included freshly diagnosed consecutive cases, presumed drug susceptible with clinical symptoms of TB, positive in sputum smear microscopy AFB, chest X-ray positive and no history of resistance to any first line tuberculosis drugs. Sputum samples were taken from this group and culture was not performed. Not all individuals were included; those with clinical complications in addition to tuberculosis were not included in this study.

Patient history
The patient’s history was collected using a proforma and included age, gender, area, economic status (earning less than 300 US dollars per month), information of previous anti-tuberculosis therapy, chest x-ray, AFB test and family history of TB. The environmental parameters studied were animal contact, source of drinking water, un-boiled milk use and smoking or drug use.

GeneXpert testing
GeneXpert testing was carried out only for six samples of the MDR group only due to lack of global funding, according to the manufacturer’s instructions. Newly diagnosed presumed susceptible samples were not tested due to the reason that these samples have no history of drug resistance. This test was performed in order to confirm the samples of MDR group before carrying out DST.

Initial sputum culturing on drug free LJ medium
For the MDR group, sputum suspension of each patient was made by mixing 0.5ml sputum in equal volume of autoclaved deionized water under aseptic conditions. 0.1ml of sputum sample was spread on LJ medium for Mycobacterium tuberculosis culture under strict aseptic conditions. The colonies appeared on the LJ medium after 4–6 weeks of culturing at 37˚C.

Drug susceptibility testing (DST)
Sub culturing of Mycobacterium tuberculosis colonies from the MDR group was carried out on LJ medium containing the different first line anti-tuberculosis drugs. The concentrations of drugs added were: rifampicin (40 μg/ml); isoniazid (0.2 μg/ml); ethambutol (2 μg/ml); pyrazinamide (50 μg/ml) and streptomycin (4 μg/ml). Following a sterility check by incubating the culture bottles at 37˚C for one week, the bottles were inoculated with an Mycobacterium tuberculosis suspension of the previous culture. After 4–6 weeks incubation at 37˚C in incubator, growth on a drug-free control medium was compared with growth on culture media containing each concentration of anti-tuberculosis drug. Any growth of Mycobacterium tuberculosis colonies on drug containing LJ media were designated drug resistant while samples where no growth was observed on drug containing LJ media, were declared to be drug susceptible Mycobacterium tuberculosis.

DNA isolation and quantification
Sputum samples from the presumed susceptible group of patients were collected and then DNA extraction was carried out. Both cultures (described above) and sputum samples were extracted using the column based TIANamp genomic DNA isolation kit (TIANGEN Biotech Beijing, China) method. Quantity and quality of the isolated genomic DNA was determined by NanoDrop (Thermo Scientific, USA) using 1μL sample of purified DNA.
**rpoB analysis**

The fragment containing 81bp Rifampicin Resistant Determining Region (RRDR) of the rpoB gene of all strains were sequenced using published primers [13] and analysed in BIOEDIT software using ClustalW alignment parameters (BioEdit version 7.2.5). The PCR was carried out in a total volume of 50μl where 1μl of the DNA was added to the reaction containing 1xPCR reaction buffer, 1.5mM MgCl₂, 0.2mM dNTPs (Invitrogen, UK), 20μM each of both rpoB-RRDR forward (5’ - CGATCACACCGCAGACGTTGA) and reverse primers (5’ - GCCAC GTCACGTCGACGACC) and 5U recombinant Taq polymerase (Invitrogen, UK). The following PCR conditions were carried out using a Veriti thermocycler (Applied Biosystems, UK): 94˚C for 2 min followed by 35 cycles of 94˚C for 30 sec, 60˚C for 30 sec and 72˚C for 1 min. Finally, an extension of 72˚C for 10 min was performed before cleaning the products using AmpureXP magnetic beads (Beckman Coulter, UK) according to the manufacturer’s protocol and forward and reverse sequencing performed.

**Ethics and consent**

The present research work was approved by the ethical committee of University of the Punjab, Lahore, Pakistan in accordance with the ethical standards of the responsible committee on human experimentation and with the latest (2008) version of Helsinki Declaration of 1975 [14]. The purpose of the study was explained and written consents from the patients or guardians were taken from all patients or from next of their kin, caretakers, or guardians/parents on behalf of all child participants.

**Results**

**GeneXpert testing**

GeneXpert testing was positive for six of the samples of MDR group and further verified by DST.

**Drug susceptibility testing (DST)**

Of the 130 MDR cultures tested for resistance to isoniazid (I), rifampicin (R), ethambutol (E), pyrazinamide (P) and streptomycin (S), 96 were resistant to IREPS, 26 were resistant to IREP, three were resistant to IR, two were resistant to IRES, one was resistant to IRP and two resistant to IRPS. Streptomycin was added so that to kill any other bacterial contamination in the culture in addition to drug sensitive mycobacteria.

**rpoB analysis**

An overview of the rpoB mutations seen in both groups is shown in Table 1. The isolated DNA from all samples quantification was carried out by nanodrop and DNA quantity was found to

### Table 1. Mutations seen in rifampicin resistance determining region (RRDR).

<table>
<thead>
<tr>
<th>SNP (AA change)</th>
<th>516 (WT = GAC- Asp)</th>
<th>526 (WT = CAC- His)</th>
<th>531 (WT = TCG- Ser)</th>
</tr>
</thead>
<tbody>
<tr>
<td>516 (WT = GAC- Asp)</td>
<td>TAC (Tyr)</td>
<td>GTC (Val)</td>
<td>GGC (Gly)</td>
</tr>
<tr>
<td>526 (WT = CAC- His)</td>
<td>TAC (Tyr)</td>
<td>AAC (Asn)</td>
<td>CCC (Pro)</td>
</tr>
<tr>
<td>531 (WT = TCG- Ser)</td>
<td>TGG (Leu)</td>
<td>TGG (Trp)</td>
<td>TGC (Cys)</td>
</tr>
</tbody>
</table>

| MDR group (n = 130) | 8 | 7 | 2 | 4 | 2 | 1 | 49 | 1 | 1 |
| ND group (n = 86) | 2 | 2 | 0 | 0 | 0 | 0 | 8 | 2 | 0 |

*Note: Multidrug resistant group (MDR), Newly diagnosed group (ND).*

https://doi.org/10.1371/journal.pone.0183363.t001
be in range of 70–90 ng/μl. Of the 130 MDR strains, 80 had mutations in the RRDR region of rpoB (61.3%) (S1 Fig). In order of mutation frequency, 49 strains carried a single mutation at position 531, TCG>TTG/Ser>Leu (65.3%), 8 strains carried a single mutation at position 516, GAC>TAC/Asp>Tyr (10%), 7 strains carried a single mutation at position 516 GAC>GTC/Asp>Val (8.8%), 4 strains carried a single mutation at position 526, CAC>TAC/His>Tyr (5%), 2 strains carried a single mutation at position 516, GAC>GCC/Asp>Gly (2.5%), a further 2 strains carried a single mutation at position 526, CAC>AAC/His>Gly (2.5%) and single strains carried mutations at positions 526, CAC>CCC/His>Pro; 531, TCG>TGG/Ser>Trp; TCG>TGC/Ser>Cys (all at 1.3%). The remaining mutations were as follows: one strain with a deletion of positions 516 and 517; one strain with two mutations, the first at position 516, GAC>GGC/Asp>Ala and the second at position 531, TCG>GCG/Ser>Ala; one strain with two mutations, the first at position 526, CAC>CAG/His>Gln and the second at position 533, CTG>CCG/Leu>Pro; one strain with a two amino acid deletion of 516/7 and two strains with mutations just upstream of the RRDR.

For the ND group, 14/86 of the extracted DNAs (16.3%) carried mutations associated with MDR status; the remainder showed wild type RRDR sequence. Of the 14, 8 DNAs carried a single mutation at position 531, TCG>TTG/Ser>Leu (57.1%); 2 DNAs carried a single mutation at position 531, TCG>TGG/Ser>Trp (14.3%); 2 DNAs carried a single mutation at position 516, GAC>GTC/Asp>Val (14.3%) and a further 2 DNAs carried a single mutation at position 516, GAC>TAC/Asp>Tyr (14.3%).

Discussion
Molecular screening of M. tuberculosis–containing sputum samples for drug resistance, although recommended by the WHO, is expensive and inaccessible to many high incidence areas such as Pakistan. To gain further information about the prevalence of MDR in newly diagnosed patients in this area of Lahore, Pakistan, the RRDR region of the rpoB gene from two groups was sequenced; the first group comprised extracted DNA from 130 MDR strains from patients diagnosed previously with TB, and with a history of resistance to first line anti-tuberculosis drugs; the second group comprised 86, DNA extracts directly from sputum samples from consecutive newly diagnosed patients, with clinical symptoms of TB, positive in sputum smear microscopy AFB, chest X-ray positive and presumed drug susceptible with no history of resistance to any first line tuberculosis drugs. The most common mutation found in both groups was in line with other studies, i.e. at position 531, TCG>TTG/Ser>Leu, and at 61.3% and 57.1% for MDR and ND groups respectively, at a similar prevalence.

Studies indicate that 96.1% of the rifampicin resistant strains worldwide will have rpoB mutations (so a surrogate marker for MDR) studies [15, 16]. Comparison of the DST results with the RRDR data from the MDR group in this study suggests that only 61.5% of strains carry mutations in this region of the rpoB gene so sequencing the RRDR does not correlate with rifampicin resistance as successfully. The reasons for this may be that resistance to rifampicin is conferred by mutations in other parts of the gene or genome or that the DST testing was sub optimal in some way and indicated resistance when none was present, although all patients from this group had a history of resistance to first line anti-tuberculosis drugs. Antimicrobial resistance testing (AST) or Drug resistance testing was established in the 1960s [17] and there is no consensus reference method for MIC determination against which the different methods can be compared to determine common breakpoints.

The main finding of this study however is the high incidence of rifampicin resistance associated mutations, which is often used as a surrogate marker for MDR. At 16.3%, it is considerably higher than the WHO estimate of 3.9% of new cases of multi drug resistant tuberculosis.
[1]. This single piece of study needs to be justified with the help of future studies to support a high percentage of rif resistant strains. The cost of rpoB sequencing in this study was approximately $10, the cost of a GeneXpert test in this region. The time taken to carry out rpoB sequencing is slightly longer than GeneXpert testing but requires more complex testing and analysis. GeneXpert testing or rpoB sequencing for detection of MDR TB in sputum samples is much faster than waiting for results of culture and DST. The diagnosis and effective treatment for individuals with MDR needs to be tailored and administered quickly by rapid molecular tests and, as a consequence of that, the control of transmission of MDR tuberculosis will be tightened. Only then will MDR tuberculosis infection and transmission be effectively controlled in high incidence areas such as Pakistan, where it is most needed.

Supporting information

S1 Fig. Alignment of RRDR sequences from this research work.

(DOC)

Author Contributions

Conceptualization: Salma Hameed, Kartyk Moganeradj, Nasir Mahmood.
Data curation: Kartyk Moganeradj, Catherine Arnold.
Formal analysis: Salma Hameed, Kartyk Moganeradj, Catherine Arnold.
Funding acquisition: Nasir Mahmood.
Investigation: Salma Hameed, Kartyk Moganeradj.
Methodology: Salma Hameed, Kartyk Moganeradj, Catherine Arnold.
Resources: Kartyk Moganeradj, Nasir Mahmood.
Supervision: Kartyk Moganeradj, Nasir Mahmood, Timothy D. McHugh, Muhammad Nawaz Chaudhry, Catherine Arnold.
Writing – original draft: Salma Hameed, Kartyk Moganeradj.
Writing – review & editing: Salma Hameed, Kartyk Moganeradj, Nasir Mahmood, Timothy D. McHugh, Muhammad Nawaz Chaudhry, Catherine Arnold.

References


