A Phylogenomic- and Proteomic Investigation into the Evolution and Biological Characteristics of the Members of the Group 2 Latin-American Mediterranean (LAM) Genotype of *Mycobacterium tuberculosis*

By

Anzaan Dippenaar, M.Sc

Thesis presented for the degree Doctor of Philosophy (Molecular Biology)
in the Faculty of Medicine and Health Sciences, at Stellenbosch University

Supervisor: Prof N.C. Gey van Pittius
Co-supervisor: Prof R.M. Warren

April 2014

DECLARATION

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

Signature:

Date:

ABSTRACT

A PHYLOGENOMIC- AND PROTEOMIC INVESTIGATION INTO THE EVOLUTION AND BIOLOGICAL CHARACTERISTICS OF THE MEMBERS OF THE GROUP 2 LATIN-AMERICAN MEDITERRANEAN (LAM) GENOTYPE OF *MYCOBACTERIUM TUBERCULOSIS*

Mycobacterium tuberculosis (*M. tuberculosis*) is the causative agent of tuberculosis (TB), a disease that affects millions of people world-wide. The species *M. tuberculosis* consists of a large number of different strains that can be grouped into at least 40 different known strain families. Many of the strains present with different pathogenic characteristics and host adaptations. The F11 LAM strains and Beijing strains currently have a nearly equal representation in the population of Cape Town, making up a total of 45% of all strains in this setting. The Latin-American Mediterranean (LAM) family of *M. tuberculosis* is proved to be the cause of a large percentage of TB cases worldwide and it is the predominant strain in high-prevalence regions such as the Western Cape and KwaZulu-Natal regions of South Africa, Zambia, Zimbabwe, and South America.

This project aimed to investigate the evolution and biological characteristics of the members of the principle genetic group (PGG) 2 Latin-American Mediterranean (LAM) genotype of *M. tuberculosis* using a combination of whole genomic and proteomic analyses, coupled to mycobacterial molecular epidemiological techniques.

The evolution of *M. tuberculosis* strain families from the Western Cape Province of South Africa proved to be consistent with previous evolutionary scenarios for *M. tuberculosis* isolated from other parts of the world. This genome-wide SNP-based phylogeny for the evolution of *M. tuberculosis* offers novel insight into the unique global representation of the *M. tuberculosis* isolates from the Western Cape, South Africa. The evolutionary scenario presented confirms six LAM sub-lineages, namely IS*6110* RFLP families F9, F11, F13, F14, F15, and F26. A subset of sub-lineage defining SNPs was determined for each of the six LAM sub-lineages.

The genomic changes in the LAM genotype strains observed through the SNP analysis presented here mostly occur in the genes involved in the cell wall, cell processes, intermediary metabolism and respiration. The same phenomenon was observed when the non-redundant SNPs of the non-LAM isolates were functionally annotated. The functional classification of the regulated proteins in the representative of the LAM RD^{Rio} lineage of *M. tuberculosis* suggests that proteins involved in the lipid metabolism, intermediary metabolism and respiration may be the key to the pathogenic effectiveness of the RD^{Rio} LAM lineage. A combination of the LAM SNP analysis and the LAM RD^{Rio}/non-RD^{Rio} comparison showed that the overall genomic- and proteomic features involved in the cell wall and cell processes of the LAM genotype differ to a large extent from what is seen in the reference strain, *M. tuberculosis* H37Rv. This genome wide phylogenetic study is the first of its kind in a South African context, and not only presents a robust phylogeny of the *M. tuberculosis* strain families, and specifically the LAM lineage, but also gives the first ever insight into the protein differences which distinguishes RD^{Rio} and non-RD^{Rio} *M. tuberculosis* strains from each other.

OPSOMMING

‘N FILOGENOMIESE EN –PROTEOMIESE ONDERSOEK IN DIE EVOLUSIE EN BIOLOGIESE EIENSKAPPE VAN LEDE VAN DIE GROEP 2 LATYNS-AMERIKAANS MEDITERREENSE (LAM) GENOTIPE VAN *MYCOBACTERIUM TUBERCULOSIS*

Mycobacterium tuberculosis (*M. tuberculosis*) is die mikrobiese agent wat tuberkulose (TB), 'n siekte wat miljoene mense wêreldwyd affekteer, veroorsaak. Die spesie *M. tuberculosis* bestaan uit 'n groot aantal verskillende stamme wat in ten minste 40 verskillende bekende stam-families gegroepeer word. Baie van die stamme toon verskillende patogeniese eienskappe en gasheer aanpassings. Die F11 LAM stam en Beijing stam het tans 'n byna gelyke verteenwoordiging in die bevolking van Kaapstad, wat 'n totaal opmaak van 45% van stamme wat in hierdie gebied gevind word. Die Latyns-Amerikaanse Meditereense (LAM) familie van *M. tuberculosis* is bewys om die oorsaak van 'n groot persentasie van TB-gevalle wêreldwyd te wees, en dit is die oorheersende stam in hoë voorkoms streke soos die Wes-Kaap en KwaZulu-Natal streke van Suid-Afrika, Zambië, Zimbabwe en Suid-Amerika.

Hierdie projek het ten doel gehad om die evolusie en biologiese eienskappe van die lede van die basiese genetiese groep (BGG) 2 Latyns-Amerikaanse Meditereense (LAM) genotipe van *M. tuberculosis* te ondersoek deur gebruik te maak van 'n kombinasie van heel genoom en proteoom analise, gekoppel aan mikobakteriële molekulêre epidemiologiese tegnieke.

Die evolusie van *M. tuberculosis* stam families van die Wes-Kaap Provinsie van Suid-Afrika blyk om in ooreenstemming te wees met vorige evolusionêre scenario's vir *M. tuberculosis* wat in ander dele van die wêreld geïsoleer is. Die genoom-wye enkelnukleotied polimorfisme-gebaseerde filogenetiese hipotese vir die evolusie van *M. tuberculosis* bied nuwe insig in die unieke wêreldwye verteenwoordiging van die *M. tuberculosis* isolate van die Wes-Kaap, Suid-Afrika. Die evolusionêre scenario wat hier aangetoon word bevestig ses LAM sub-lyne, naamlik IS*6110* RFLP families F9, F11, F13, F14, F15, en F26. 'n Versameling sub-lyn definiërende enkelnukleotied polimorfismes was bepaal vir elk van die ses LAM sub-afstammelinge.

Die genomiese veranderinge wat waargeneem is in die LAM-genotipe isolate deur die enkelnukleotied polimorfisme analise wat hier aangebied word, is meestal in die gene wat betrokke is in die selwand, selprosesse, intermediêre metabolisme en respirasie. Dieselfde verskynsel is waargeneem wanneer die nie-oorbodige enkelnukleotied polimorfismes van die nie-LAM isolate funksioneel geannoteer is. Die funksionele klassifikasie van die gereguleerde proteïene in die verteenwoordiger van die LAM

RD^{Rio}-lyn van *M. tuberculosis* dui daarop dat die proteïene wat betrokke is in die lipiedmetabolisme, intermediêre metabolisme en respirasie die sleutel tot die patogene doeltreffendheid van die RD^{Rio}-LAM-lyn kan wees. 'n Kombinasie van die LAM enkelnukleotied polimorfisme analise en die LAM-RD^{Rio}/nie-RD^{Rio} vergelyking het getoon dat die totale genomiese- en proteomiese kenmerke wat verwant is aan selwand en selprosesse van die LAM genotipe tot ’n groot mate verskil van wat gesien word in die verwysing stam, *M. tuberculosis* H37Rv. Hierdie genoom-wye filogenetiese studie is die eerste van sy soort in 'n Suid-Afrikaanse konteks, en bied nie net ‘n robuuste filogenie van die *M. tuberculosis* stam families, en spesifiek die LAM genotipe van *M. tuberculosis* nie, maar gee ook die eerste keer ooit insig in die proteïen verskille wat RD^{Rio} en nie-RD^{Rio} *M. tuberculosis* stamme van mekaar onderskei.

ACKNOWLEDGEMENTS

This study was made possible by the input of numerous people. Without their contribution, effort and encouragement this work would not have been possible.

I would like to express my sincere gratitude to the following people:

Prof N.C. Gey van Pittius, Prof R.M. Warren and Prof P.D. van Helden, my project supervisors and research host, for their guidance, help, patience, trust and encouragement during the past few years.

All my colleagues and friends in the Mycobactomics research group for their friendship, support and help through the year.

Dr S. Smit, for her guidance and support in the proteomic analyses included in this thesis.

Everyone who were involved and contributed to the experience I have gained in being trained as a bioinformatician, Margaretha de Vos, Prof. Alan Christoffels, Mmakamohelo Direko, Anita Schürch, Keith Siame, Michelle Daya, Ruben van der Merwe, Dr. Taane Clark.

The National Research Foundation (NRF) and The Harry Crossley Foundation, for the financial support that made this study possible.

The financial assistance of the National Research Foundation (NRF) towards this research is hereby acknowledged.Opinions expressed and conclusions arrived at, are those of the author and are not necessarily to be attributed to the NRF.

My parents, Stones and Ansa Steenkamp, Isak and Carol Dippenaar, for their unconditional love and on-going support.

My husband, Riaan Dippenaar, for his love, support, understanding and encouragement every single day.

To the Lord, who has blessed me with the strength to endure through the tough times and the grace to rejoice in the good times, and whose grace and blessings carried me through this journey.

LIST OF ABBREVIATIONS

°C	Degrees Celcius
AA	Amino acid
A	Adenine or adenosine, one-letter code for alanine
ABC	Ammonium bicarbonate
ACN	Acetonitrile
AIDS	Acquired immune deficiency syndrome
bam	binary alignment map
BCG	Bacillus Calmette-Guérin
BFAST	Blat-like fast accurate Search tool
BLAST	Basic local alignment search tool
bp	Base pair
BSA	Bovine serum albumin
BSL3	Biosafety level 3
BWA	Burrows-Wheeler Aligner
C	Cytosine or cytidine, one-letter code for cysteine
CAS	Central Asian strains
CDS	Coding sequence
dATP	Deoxyadenosine triphosphate
DC	Dextrose catalse
ddH_2O	Double distilled water
DNA	Deoxyribo-nucleic acid
DNase	Deoxyribonuclease
DNS	Deoksieribo-nukleïensuur
dNTP	Deoxynucleoside triphosphate
DR	Drug resistant, direct repeat
DS	Drug sensitive
DTT	Dithiothreitol
DVR	Direct variable repeat
EAI	East African Indian
EDTA	Ethylenediaminetetraacetic acid
ESI	Electro spray ionisation

et al.	*et al*ii (and others)
F	Family (reffering to IS*6110* RFLP)
FA	Formic acid
g	Gram, gravity
G	Guanine or guanosine, one-letter code for glycine
GATK	Genome analysis toolkit
GTR	General time reversal
HGT	Horizontal gene transfer
HIV	Human immunodeficiency virus
i.e.	id est (that is)
IAA	Iodoacetamide
IL	Interleukin
In/del	Small insertions and deletions
IS*6110*	Insertion sequence *6110*
kb	Kilobase
KZN	KwaZulu Natal
LAM	Latin-American Mediterranean
LC	Liquid chromatography
LCC	Low copy clade
LJ	Löwenstein-Jensen
LSP	Large sequence polymorphism
M.	*Mycobacterium*
M	Molar / relative molecular weight
Mb	Megabase
MDR	Multi drug resistant
MGIT	Mycobacterial growth indicator tube
MIRU/VNTR	Mycobacterial interspersed repeat elements / variable number tandem repeats
ml	Milliliter
mm	Milimeter
mM	Millimolar
MMR	Miss match repair
MOPS	3-(N-morpholino)propanesulfonic acid (buffer)

mRNA	Messenger ribonucleic-acid
MS	Mass spectrometry
MTBC	*Mycobacterium tuberculosis* complex (excluding *M. canettii*)
MTC	*Mycobacterium tuberculosis* complex (including *M. canettii*)
MW	Molecular weight
NHLS	National Health Laboratory Service
NGS	Next generation sequencing
nt	Nucleotide
OD	Optical density
ORF	Open reading frame
PAGE	Poly acrylamide gel electrophoresis
PCR	Polymerase chain reaction
PGG	Principle genetic group
pH	Potential of hydrogen
QS	Quality score
QV	Quality value
rcf	Relative centrifugal force
RD	Region of difference
RFLP	Restriction fragment length polymorphism
RNA	Ribonucleic-acid
rpm	Revolutions per minute
sam	sequence alignment map
SAWC	South Africa Western Cape
SB	Sodium borate (buffer)
SDS	Sodium dodecyl sulphate
SNP	Single nucleotide polymorphism
Spoligotyping	Spacer oligonucleotide typing
STB	Smooth tubercle bacilli
T	Thymine or thymidine
TAE	Tris Acetate EDTA (buffer)
TB	Tuberculosis
TCA	Tri-carboxylic acid
TDR	Totally drug resistant

TE	Tris EDTA
T_m	Melting temperature
TNF	Tumour necrosis factor
us.	Upstream
vcf	Variant call format
V	Volt
WGS	Whole genome sequencing
WHO	World Health Organization
www	Wold wide web
XDR	Extremely drug resistant
ZN	Ziehl-Neelsen

LIST OF SYMBOLS

µ	Micro
η	Nano
%	Percent/ percentage
˚ C	Degrees Celsius
®	Registered trademark
™	Trademark
©	Copyright
Δ	Delta
~	Approximately

CONTENTS

CHAPTER 1

INTRODUCTION

1.1 BACKGROUND

Mycobacterium tuberculosis (*M. tuberculosis*) is the causative agent of tuberculosis (TB), a disease which is a major threat to the health of millions of people worldwide (WHO, 2011). The species *M. tuberculosis* is part of the *M. tuberculosis* complex (MTC), and consists of a large number of different strains falling into at least 40 different known strain families, with many presenting with different pathogenic characteristics and host adaptations (Brudey *et al*., 2006).

The different species and strains of the MTC is thought to have evolved from a common mycobacterial ancestor (Brosch *et al*., 2002). The evolution of the MTC has been hypothesized to have occurred in parallel with the evolution of anatomically modern humans and the expansion of the pathogen is ascribed to the increased human population density in the Neolithic period (Comas *et al*., 2013). The global spread of human-adapted *M. tuberculosis* strains coincides with modern human migration patterns out of Africa (Wirth *et al*., 2008). During the evolutionary process, members of the *M. tuberculosis* complex are thought to have lost the ability to undergo horizontal gene transfer, and as such, the only genetic changes that can still take place in these organisms are deletions, duplications, genetic rearrangements, transposon insertions and single nucleotide polymorphisms. Notwithstanding the high clonality of these MTC members, different lineages, sub-lineages and strains have evolved over time through the mechanisms mentioned above (Brosch *et al*., 2002).

The different strains of *M. tuberculosis* vary greatly in pathogenicity and fitness and are found historically associated with different human host populations from different regions of the world (Brown *et al*., 2010; Comas *et al*., 2013; Gagneux *et al*., 2006). Due to the fact that South Africa was located in a geographically central position in the historical trade route between East and West for hundreds of years, we observe a spectrum of strains from both European and Eastern origin as the dominant strain types in this country (Mokrousov *et al*., 2005). Over time and in association with different host populations, different strains of *M. tuberculosis* have acquired different phenotypic properties through evolution, resulting in strains which are, for example, hyper- or hypovirulent (de Souza *et al*., 2010). Furthermore, it was observed that specific strain families seem to outcompete others in a certain setting, suggesting host-pathogen compatibility (Gagneux *et al*., 2006). Co-evolutionary events between the host and pathogen are further supported by the association between human leukocyte antigen types and specific *M. tuberculosis* strain genotypes (Salie *et al*., 2013).

The Latin-American Mediterranean (LAM) lineage of *M. tuberculosis* is prevalent in Latin-America and the Mediterranean region. More recently, the LAM lineage has been shown to be the predominant cause of TB in Russia and in high-prevalence African regions such as the Western Cape and KwaZulu-Natal regions of South Africa, Zambia, and Zimbabwe (Chihota *et al*., 2012; Mulenga *et al*., 2010). The RD^{Rio} subset of LAM genotype strains shows effective transmissibility and are thought to be hyper-virulent, compared to non-RD^{Rio} strains (Gibson *et al*., 2008; Lazzarini *et al*., 2007, 2008). The LAM genotype is responsible for approximately 15% of TB cases globally (Brudey *et al*., 2006). The F11 LAM strains and Beijing strains currently have a nearly equal representation in the population of Cape Town, making up a total of 45% of all strains in this setting (Victor *et al*., 2004). We observe changing *M. tuberculosis* population patterns, largely due to the increase in the incidence of cases caused by specific strains of *M. tuberculosis* in a local South African setting. This highlights the ability of strains to acquire genotype-specific pathogenic characteristics (van der Spuy *et al*., 2009).

The fact that LAM strains are a common cause of TB worldwide highlights the importance of the LAM genotype and suggests that this strain lineage is a global threat that should be specifically targeted by public health resources.

1.2. PROBLEM IDENTIFICATION

In order to control the TB epidemic, it is essential to understand the dynamics of the epidemic. We need to be able to accurately identify the lineages, clusters and strains to which *M. tuberculosis* isolates belong in order to identify their genomic background, to understand the changing population patterns, and track outbreaks. Furthermore, we need to understand the biological grounds for these strain differences in order to efficiently design drugs and vaccines to combat the disease.

The Beijing and LAM genotypes of *M. tuberculosis* have been recognized as the major clades to cause TB globally (Brudey *et al*., 2006). The Beijing genotype has been the subject of various *in vivo* and *in vitro* studies over the last decade, due to its apparent hyper-virulence leading to increased transmissibility and ability to cause disease (Hanekom *et al*., 2011). Even though the LAM family of *M. tuberculosis* has been shown to be the cause of a large percentage of TB cases worldwide, it has not been as extensively studied and much remains to be uncovered about the mechanisms of pathogenicity of this strain family. We hypothesize that the epidemiological success of this strain family reflects physiological changes through evolution of the proteome as a result of altered gene expression (which may be derived from genomic mutations). As such, an understanding of the evolution and diversification of these strains, as well as the biological reasons underlying their fitness, are crucial to the combat of this disease.

1.3. AIMS

This project aims to investigate the evolution and biological characteristics of the members of the principle genetic group (PGG) 2 Latin-American Mediterranean (LAM) genotype of *M. tuberculosis* using a combination of whole genomic and proteomic analyses, coupled to mycobacterial molecular epidemiological techniques.

Whole genome sequencing and high throughput mass spectrometry will be employed in conjunction with a large set of molecular markers (which includes spoligotyping, IS*6110*-RFLP, deletions, and single nucleotide polymorphisms selected from previous studies in our department, from literature, and from the results of publicly-available whole genome sequencing) to screen all LAM strains available to the study. The results will be used to generate a robust phylogenetic reconstruction of the evolutionary history of the *M. tuberculosis* LAM genotype. A set of markers will be identified that could be used for the rapid screening, delineation and identification of lineages, sublineages and strains of LAM *M. tuberculosis*, and identify the biological basis for their variation in fitness and pathogenicity.

Overall aim of this study:
To use whole genome sequencing in association with highly sensitive mass spectrometry techniques to describe the genomes and proteomes of strains representative of the different sub-lineages of the *M. tuberculosis* LAM genotype.

1.4. OBJECTIVES

- To culture strains representative of the different strain families of the *M. tuberculosis* LAM genotype.
- To extract genomic DNA from strains representative of the different strain families of the *M. tuberculosis* LAM genotype for whole genome sequencing.
- To establish a pipeline for the bioinformatic analysis of next generation sequencing (NGS) data.
- To do bioinformatic analyses of the whole genome sequence data to identify changes that have occurred during the evolution of the LAM genotype.
- To generate a comprehensive phylogenetic tree of the LAM genotype.
- To extract whole cell lysate proteins from strains representative of different strain families of the *M. tuberculosis* LAM genotype for protein mass spectrometry analysis.

- To identify proteins that have changed in abundance that may be involved in the development of the high virulence phenotypes of *M. tuberculosis*.
- To develop a ‘map’ of the changes in the genomes and proteomes to correlate mutation with protein abundance.

CHAPTER 2

LITERATURE REVIEW

M. TUBERCULOSIS EVOLUTION: THE GROUP II LATIN-AMERICAN MEDITERRANEAN (LAM) GENOTYPE

2.1. THE CURRENT SITUATION OF TUBERCULOSIS GLOBALLY

Mycobacterium tuberculosis (*M. tuberculosis*) is the causative agent of tuberculosis (TB) – an infectious disease that has pestered human-kind for centuries. Robert Koch was awarded the Nobel Prize for his discovery of *M. tuberculosis* at the turn of the 19th century. However, the pathogen is still responsible for a worldwide TB epidemic, which reached an all-time high in 2009 when the total number of tuberculosis cases worldwide was estimated at 9.4 million (WHO, 2011). Suggested factors that contribute to the extremely high global TB burden include HIV co-infection, the emergence of drug resistant (DR)-, multidrug resistant (MDR)-, extremely drug resistant (XDR)- and recently reported, totally drug resistant (TDR) TB (Chiang *et al*., 2010; Devaux *et al*., 2009; Migliori *et al*., 2009; Schaaf *et al*., 2009).

The TB burden in Africa, which also largely applies to South Africa, is extremely high compared to other parts of the world. This is striking when it is considered that the population size is smaller than e.g. South East Asia, yet the incidence (cases arising during a certain period), prevalence (new and existing cases at a given point in time) and mortality due to tubercle disease is second highest in the world in all three categories. The mortality rate due to TB in Africa (excluding HIV) was 50 per 100 000 during 2009. HIV contributes greatly to the increase in incidence of tuberculosis all over the world, especially in Africa. Among HIV-negative TB patients, 1.3 million deaths were reported worldwide in 2009 amongst HIV negative TB patients (WHO, 2012).

2.2. EARLY ORIGINS OF *M. TUBERCULOSIS*

The *Mycobacterium tuberculosis* complex (MTC) is comprised of a number of *Mycobacterium* species that cause tubercle disease in mammals. Members include *M. tuberculosis, Mycobacterium africanum (M. africanum), Mycobacterium canettii (M. canettii), Mycobacterium microti, Mycobacterium bovis (M. bovis)* and *Mycobacterium bovis* Bacillus Calmette-Guérin (BCG), *Mycobacterium pinnipedi,* *Mycobacterium caprae, Mycobacterium suricattae,* the dassie bacillus and the oryx bacillus, recently named *Mycobacterium orygis* (Brosch *et al*., 2001; Parsons *et al*., 2002, 2013; van Ingen *et al*., 2012; Walker *et al*., 2013). These species are all able to cause disease in mammals, including humans but tend to favour disease in a certain host (e.g. dassie bacillus is mostly found in Cape Hyrax and *M. pinnipedi* in seals) (Smith, 1898). Members of the MTC are highly clonal, are said to share 99.9% similarity in their nucleotide sequence. Little evidence of horizontal or lateral gene transfer has been observed (Baker *et al*., 2004). The 16s RNA sequences of the various members are highly conserved, except in the ancestral *M. canettii* (Dos Vultos *et al*., 2008).

Studies performed during the last decade prove that the ancestor of *M. tuberculosis* is *M. canettii*, and not *M. bovis* as previously assumed (Brosch *et al*., 2002; Sreevatsan *et al*., 1997). The different mycobacterial species comprising the MTC are thought to have evolved from this common ancestor. Some controversy exists as to the exact age of mycobacteria and when it evolved into the different mycobacterial species that are known to date (Wirth *et al*., 2008). A recent coalescent study estimated that the MTC emerged roughly 70 000 years ago and was spread by anatomically modern humans migrating out of Africa (Comas *et al*., 2013). Epidemiologic and phylogenetic studies have shown that tubercle bacilli emerged in Africa, and consequently underwent initial diversification (Comas *et al*., 2013; Walker *et al*., 2013). A successful pathogenic clone of the MTC, *M. tuberculosis,* expanded and spread to the rest of the world, most likely with the early human migration patterns driven by the historical trade route to and from Africa (Walker *et al*., 2013).

Comparative genomics have revealed that even the most distantly related *Mycobacterium species* (non-MTC members) share 60% DNA homology and that the loss of genetic material in particular contributes towards the continuous evolution of the slow-growing pathogenic species of mycobacteria (Brosch *et al*., 2001). It is commonly accepted that *M. tuberculosis* do not undergo horizontal gene transfer, although available data suggest that lateral genetic exchange has occurred in the precursor of the MTC, commonly referred to as *M. prototuberculosis* (Brisse *et al*., 2006; Supply *et al*., 2003; Walker *et al*., 2013). It is said that modern mycobacteria are continuously and rapidly evolving as apparently successful pathogens after the evolutionary ancestral bottleneck that gave rise to the different members of the MTC known to date (Namouchi *et al*., 2012).

Variations in *M. tuberculosis* were already described in 1982 (Collins *et al*., 1982) when five variants where identified on the basis of differences in oxygen requirement (aerobic or microaerophilic), nitrate reductase activity, susceptibility to pyrazinamide and susceptibility to thiophene-2-carboxylic acid hydrazide. These variants were then termed classical human, Asian human, bovine, African I and African II, and this terminology was used in early epidemiological studies of *M. tuberculosis* infection

(Collins *et al*., 1982). Current strain classification is based on various genetic markers (elaborated upon in Section 2.3).

Evolution may refer to change in the genetic properties and/or physical appearance of an organism over time. A clear distinction can be made between micro- and macro-evolution, where micro-evolution refers to small changes that occur in a shorter period of time and is undergone in one or more small events. Macro-evolution is achieved through rare, but large events and usually over a longer period of time (Behe, 1998). Genetic variation in a population makes it possible for that population to evolve when certain selective pressures are applied. The diversity among *M. tuberculosis* strains has largely been attributed to genetic drift, whilst the influence of natural selection on the development of diversity patterns is still poorly understood (Pepperell *et al*., 2013).

2.3. *M. TUBERCULOSIS* STRAIN DIVERSITY

Genetic divergence between the different *M. tuberculosis* strains underlies the variation in phenotype and disease presentation (Sarkar *et al*., 2012). *M. tuberculosis* lineages cause varying disease phenotypes that include duration and extent of disease, anatomic dispersal of tubercle lesions and response to treatment (Caws *et al*., 2008; Click *et al*., 2012; Hanekom *et al*., 2011; Hirsh *et al*., 2004). However, as a routine practice, the same drug regimen is administered to patients to treat tuberculosis regardless of the strain(s) of *M. tuberculosis* responsible for the infection, even though it has been shown that strains respond to treatment differently (Richardson *et al*., 2002; WHO, 2011). Even more alarming is the fact that tuberculosis treatment is often started even before any genetic characterisation (genotyping) of the disease-causing strain is performed, leading to a situation where patients with tuberculosis caused by DR-, MDR-, XDR strains together with drug sensitive (DS) strains are subjected to the same initial treatment. This may lead to the selection of DR-, MDR- or XDR strains in favour of the DS strain in the host. Alternatively, selective drug pressure may contribute to the development of a DR (or worse) strain if such a strain has a genetic predisposition to acquire resistance to a specific drug (Davies and Davies, 2010; Nachega and Chaisson, 2003). The on-going TB epidemic, the emergence of drug resistance and the consequent complications thereof have lead scientists to explore the genetic basis of *M. tuberculosis* variability in order to understand the underlying mechanisms of pathogenicity.

Strains of pathogens that cause severe disease in the host and cause rapid mortality are thought to undergo negative selection due to the reduction transmission time (de Souza *et al*., 2010; Mostowy and Behr, 2005). This common perception in the evolution of pathogens is challenged by the fact that natural selection may favour hyper-virulent strains (causing severe disease and rapid mortality) if these strains carry a remunerative advantage such as increased immediate rates of transmission, or high resistance to host defence-mechanisms (Lipsitch and Moxon, 1997). The perfect pathogen maintains a fine balance between virulence and a prolonged disease process in order to promote transmissibility through the host (Mostowy and Behr, 2005).

Several studies have shown that a small number of strains are frequently responsible for a large number of TB cases in a defined area (Chihota *et al*., 2007; van Soolingen *et al*., 1995). It can be speculated that inherent increased transmissibility of certain strains compared to others may favour the spread of that particular strain in an area. Inherent transmissibility may be due to smear positive disease caused by the strain in question, which is more prone to transmission, delayed onset of disease symptoms which cause patients to be infectious for a longer period of time, or increased virulence compared to other strains (Chihota *et al*., 2011).

It is evident that *M. tuberculosis* is more genetically diverse than previously assumed. In general, changes in the *M. tuberculosis* genome are caused by single nucleotide polymorphisms (SNPs), deletions, duplications, genetic rearrangements, and transposon insertions. The more recent changes may have occurred as a result of aberrant DNA repair, recombination and replication (3R); the *M. tuberculosis* genome contains no genes encoding components involved in the miss match DNA repair pathway (MMR) and as such, *M. tuberculosis* can be regarded as a natural mutator to some extent (Dos Vultos *et al*., 2008; Mestre *et al*., 2011).

2.4. GENETIC CHARACTERIZATION OF *M. TUBERCULOSIS*

Genetic markers are used to identify and distinguish different strains of *M. tuberculosis*, or to differentiate strains from the same lineage from one another to study the molecular epidemiology and the spread of disease, as well as to study the evolutionary history of *M. tuberculosis*. The strain-specific markers have different levels of discriminatory power and stability and various genetic typing methods are used to characterise clinical isolates of *M. tuberculosis* (Brosch *et al*., 2002; Brudey *et al*., 2006; Filliol *et al*., 2006; Frothingham and Meeker-O'Connell, 1998; Groenen *et al*., 1993; Gutacker *et al*., 2002; Moström *et al*., 2002; Sreevatsan *et al*., 1997). These typing methods rely on the characteristics of a small section of the genome of *M. tuberculosis* to make assumptions to identify clinical isolates in terms of lineage, family, or strain. The markers used in the different typing methods are diverse and vary in discriminatory power, and thus; the results obtained by different typing methods do not always agree (Filliol *et al*., 2006; Ford *et al*., 2012; Hanekom *et al*., 2008).

Various molecular epidemiologic genetic markers are used to characterise the MTC. These include IS*6110* – an insertion sequence unique to members of the MTC, guanine- and cytosine-rich elements present in the genome, spacer oligonucleotide typing (spoligotyping) of the direct repeat locus, mycobacterial interspersed repetitive units and variable-number tandem repeats (MIRU-VNTR) (Aranaz *et al*., 1996; Frothingham and Meeker-O'Connell, 1998; Hermans *et al*., 1991; Kamerbeek *et al*., 1997; Mazars *et al*., 2001; Ross *et al*., 1992; Van Embden *et al*., 1993).

The genetic typing method known as spoligotyping is based on determining the presence or absence of specific spacer sequences in the direct repeat (DR) locus of members of the MTC (Kamerbeek *et al*., 1997). The DR locus consists of directly repeated DNA sequences, 36 bases in length, interspersed by 34 - 40 bp non-repetitive spacer sequences (Kamerbeek *et al*., 1997).

It is hypothesised that the evolution of the DR locus of *M. tuberculosis* is unidirectional; spacers in this region are lost rather than gained (Van Embden *et al*., 1993). Spoligotype families are defined by certain spacer patterns which may evolve autonomous from each other in different strains causing homoplasy. This could be problematic as an epidemiologic tool, as it is possible that the spoligotype pattern that a strain presents with, may not represent the true identity of that strain (Warren *et al*., 2002). Spoligotyping is thus not a reliable typing method on its own, due to it being limited to a small portion of the genome. It is becoming clear that with the rise of convergent spoligotype patterns, one typing method is not sufficient to determine the true strain identity of *M. tuberculosis* isolates (Comas *et al*., 2009; Flores *et al*., 2007; Gibson *et al*., 2008; Reed *et al*., 2009; Schürch and van Soolingen, 2011).

2.5. LINEAGE-SPECIFIC GENETIC MARKERS OF *M. TUBERCULOSIS*

Single nucleotide polymorphisms (SNPs) are single nucleotide changes that occur in the genome and can be synonymous (polymorphisms that do not change the amino acid sequence) or non-synonymous (polymorphisms that alter amino acid sequence of the encoded protein). Some of the first classification systems of the MTC were based on SNPs. Members of the MTC are classified in three principle genetic groups based on different combinations of SNPs in the *katG* and *gyrA* genes at the following positions; *katG463* and *gyrA95*. Although *katG* and *gyrA* genes are both associated with drug-resistance, the abovementioned specific SNPs do not confer resistance (Sola *et al*., 2001; Sreevatsan *et al*., 1997).

Large sequence polymorphisms (LSPs), better known as regions of difference (RDs), are commonly detected in species of mycobacteria and are mostly genomic deletions, although insertions or genetic duplications also occur (Alland *et al*., 2007). TbD1 is an LSP in the genome of some *Mycobacterium* species that distinguishes ancient mycobacterial species and strains from their modern counterparts (Brosch *et al*., 2002). The TbD1 region is present in the genomes of all animal-adapted strains and some of the ancient human-adapted strains (*M. africanum* type 1 isolates, MANU and EAI) but deleted from most modern strains of *M. tuberculosis* (Baker *et al*., 2004). The TbD1 deletion of 2.1 Kb thus pre-dates the *katG* $_{G1388T}$ and *rpoB* $_{T3243C}$ mutations, and is estimated to have been deleted from the ancestral *M. tuberculosis* lineage before the 18th century (Baker *et al*., 2004; Brosch *et al*., 2002). Genes present in the TbD1 region (which include *mmpS6* and *mmpL6*, both encoding membrane proteins), are thus not present in modern *M. tuberculosis* strains (Cole, 2002).

Lineage-specific polymorphisms are unlikely to occur simultaneously, and can thus be used to infer phylogeny or an order of evolutionary informative events. SNP-based phylogenies of the MTC groups the members of the complex into multiple lineages that separate the animal- and human-adapted strains from each other and further subdivide the *M. tuberculosis* strains (Baker *et al*., 2004; Gutacker *et al*., 2002). Genome wide polymorphisms in genes including *oxyR*, *katG* and *rpoB*, have been used to subdivide *M. tuberculosis* strains in four lineages or phylogenetic groups, in a study by Baker *et al.* (2004) - *M. bovis* strains accounted for a fifth lineage (Baker *et al*., 2004). These phylogenetic groups, based on polymorphisms at multiple loci, gave similar groupings to the principle genetic groups distinguished by the *katG463* and *gyrA95* polymorphisms, IS*6110* RFLP patterns and spoligotype data (Bhanu *et al*., 2002; Filliol *et al*., 2003; van Soolingen *et al*., 1995).

M. tuberculosis strains causing disease in South Eastern Asia are mostly restricted to lineage I, whilst TB in Europe is mainly caused by members of lineage II, whereas lineage III is strongly associated with the Indian subcontinent. However, lineage IV is globally distributed with the exception of Europe. African-born TB patients however, do not show an association with a single lineage of *M. tuberculosis*, but a wide array of *M. tuberculosis* strains cause disease in this high incidence setting (Baker *et al*., 2004). The mixed bag of *M. tuberculosis* strains is partly explained by the African continent's central position in the historical trade routes between East and West. Although the most ancient members of the MTC traces back to Africa, this continent and its people were also exposed to various strains of *M. tuberculosis* from across the world (Baker *et al*., 2004).

LSPs may be located in regions of the *M. tuberculosis* genome that are likely to undergo insertion or deletion events (so-called hotspot regions) and are often flanked by the IS*6110* insertion element (Alland *et al*., 2007; Tsolaki *et al*., 2004). These LSPs have the potential to be somewhat inexpedient for the pathogen, but in some cases they are speculated to be advantageous. Possible beneficial effects of large deletions include the ability of the pathogen to escape the host immune system, reducing the load of microbial mobile genetic elements, or even to promote drug resistance. Deletions can also render a state of latency in the microbe to promote long term illness and thus increased opportunity for transmission (Tsolaki *et al*., 2004). LSPs have been shown to provide considerable diversity within mycobacteria and are considered to be more stable genetic markers than point mutations as the latter are subject to reversion (Alland *et al*., 2007; Brosch *et al*., 2002). These LSPs are often used to infer the phylogeny of *M. tuberculosis* strains and can thus differentiate between certain groups of strains if the LSP is a result of a unique event polymorphism (UEP) (Alland *et al*., 2007; Gagneux *et al*., 2006; Hirsh *et al*., 2004; Huard *et al*., 2003). UEPs are polymorphisms that are irreversible and do not display homoplasy (convergent evolution) (Hirsh *et al*., 2004). LSPs are very useful for defining robust phylogenetic groups, and are particularly informative and discriminatory for ancient *M. tuberculosis* strains and W-Beijing strains that present with a spoligotype profile where nearly all spacers are absent (Flores *et al*., 2007).

A significant association was shown in several studies between the patients' origin (country of birth) and the lineage of the disease-causing strain of *M. tuberculosis* (Brown *et al*., 2010; Gagneux *et al*., 2006). Host-pathogen compatibility was inferred from the fact that different lineages of *M. tuberculosis* have been associated with the continent of birth of the patients in which TB is caused (Gagneux *et al*., 2006). Regardless of the differences in naming convention and strain distribution of *M. tuberculosis* lineages described in different studies, certain strains show predominance in specific geographical areas of the world. SNPs were used to infer four lineages of *M. tuberculosis* (see Table 2.5.1) (Baker *et al*., 2004). The various lineages of *M. tuberculosis* have consequently been named in conjunction with the geographical location in which a specific strain was first isolated and where it is thought to have originated. It has been shown that certain lineages prevail as the dominant strain responsible for causing disease in a specific geographical setting and the associated host population (Gagneux *et al*., 2006).

A subsequent study grouped the MTC into six phylogenetic lineages (one to six) and one lineage consisting of animal-adapted strains of mycobacteria on the basis of a phylogenetic analysis of over 300 SNPs (Hershberg *et al*., 2008). The study performed by Hershberg *et al.* (2008) referred to the *M. tuberculosis* complex as MTBC as they did not include *M. canetti* in their analysis since it is more distantly related to any species or strain within the complex than any two members of the MTC from each other (Hershberg *et al*., 2008). This phylogenetic grouping was more comprehensive than the initial SNP-based lineages and was also concordant with that obtained with LSPs, including the presence and absence of the TbD1 deletion, and spoligotype family groupings (Baker *et al*., 2004; Gagneux *et al*., 2006). Lineage one includes the East-African-Indian (EAI) or the Indo-Oceanic strains. The well-described Beijing genotype of *M. tuberculosis* correlates with lineage two, Central-Asian (CAS) and Delhi strains belong to lineage three, Haarlem-, Latin-American Mediterranean (LAM)-, X- and U families segregated to form several sublineages of lineage four, better known as the Euro-American lineage. The widely used reference strain; *M. tuberculosis* H37Rv is a member of lineage four, and is thus a member of PGG3. *Mycobacterium africanum* AFR2 correspond to lineage five and

Mycobacterium africanum AFR1, to lineage six (Chuang *et al*., 2010; Comas *et al*., 2009). A summary of the SNP- and LSP-based lineages of *M. tuberculosis* are shown in Table 2.5.1. Data mining approaches have identified seven basal/primary strain groups of *M. tuberculosis*, which include Beijing, LAM, EAI, Haarlem, CAS, X, and T, which can all be assigned to one of the major lineages as described above (Brudey *et al*., 2006; Filliol *et al*., 2002; Sebban *et al*., 2002; Sola *et al*., 2001). Strong evidence also suggests that geographic structuring in *M. tuberculosis* populations could play a role in the evident global variation in success of the *M. bovis* BCG vaccine (Baker *et al*., 2004; Brown *et al*., 2010; Gagneux *et al*., 2006).

SNP-typing of mycobacteria has been extensively used in phylogenetic studies to determine the evolutionary structure of the MTC and, more specifically, *M. tuberculosis* and its large array of strains (Baker *et al*., 2004; Filliol *et al*., 2006; Gutacker *et al*., 2006, 2002). Hence, various SNP clusters and SNP based phylogenies have been published to date, the groupings mostly coincide with each other but due to the non-uniform naming scheme of these clusters or phylogenetic groupings, some confusion may exist (Baker *et al*., 2004; Chuang *et al*., 2010; Filliol *et al*., 2006; Gutacker *et al*., 2006, 2002). Table 2.5.1 shows a summary of the lineages of *M. tuberculosis* as defined by SNPs, LSPs, and spoligotyping.

Table 2.5.1. A summary of the lineages of *Mycobacterium tuberculosis* and *Mycobacterium africanum* as determined by various genetic markers in different studies

Markers used (Study)	Lineage 1	Lineage 2	Lineage 3	Lineage 4	Lineage 5	Lineage 6
SNPs (Sreevatsan *et al*., 1997)	PGG1	PGG1	PGG1	PGG2, 3	PGG1	PGG1
SNPs (Baker *et al*., 2004)	Lineage IV	Lineage I	Lineage III	Lineage II	Not done	Not done
LSPs (Gagneux *et al*., 2006)	Indo-Oceanic lineage	East Asian lineage	East African-Indian lineage	Euro-American lineage	West African lineage I	West African lineage II
SNPs (Gutacker *et al*., 2006)	Cluster I	Cluster II	Cluster II.A	Clusters III-VII	Not done	Not done
SNPs (Filliol *et al*., 2006)	Cluster group 1	Cluster group 2	Cluster group 3a	Cluster groups 3b-6b	Not done	Not done
Spoligotyping (Brudey *et al*., 2006)	EAI	Beijing	CAS	Haarlem, LAM, T, X, U	AFRI2	AFRI1
LSP marker	RD239	RD105	RD750	Pks15/1 - 7bp	RD711	RD702
SNP marker	*OxyR* C37T	*Rv3815c* G81A	*RpoB* T2646G	*KatG* T1388G, *RpoB* C3243T	Not known	Not known
Geographical association	East Africa, South-East Asia, South India	East Asia, Russia, South Africa	East Africa, North India, Pakistan	Americas, Europe, North Africa, middle east	Ghana, Benin, Nigeria, Cameroon	Senegal, Guinea-Bissau, The Gambia
Other information	Ancestral strains (TBD1 +)	-	-	-	*M. africanum*	*M. africanum*

PGG: principle genetic group, LAM: Latin-American Mediterranean, LCC: low copy clade, CAS: Central Asian strains, EAI: East African Indian, AFRI1: *M. africanum* type 1, AFRI2: *M. africanum* type 2. Table adapted from Gagneux and Small 2007 (Gagneux and Small, 2007).

2.6. WHOLE GENOME SEQUENCING OF *M. TUBERCULOSIS*

With the rapid advances in next generation sequencing techniques and the decrease in its cost, several of the MTC members, various strains of *M. tuberculosis* and other *Mycobacterium* species have been subjected to whole genome sequencing (WGS), which provides genetic information with a much greater resolution than previous methods (Bentley *et al*., 2012; Chan *et al*., 2012; Cole *et al*., 1998; Fleischmann *et al*., 2002; Gardy *et al*., 2011; Pan *et al*., 2011).

When the first whole genome sequence of *M. tuberculosis* (of the laboratory strain H37Rv) was first published in 1998, it was envisaged that phenotypic variation would be readily explained by the underlying genetic features. However, the genome sequence of *M. tuberculosis* in combination with various genetic typing techniques has not provided an unambiguous understanding of the underlying mechanisms of pathogenicity. Recent advances in the field of TB research together with the rapid decline in the cost of WGS technologies made genetic characterisation of different strains much more accessible, plausible and feasible in a research and diagnostic environment (Ford *et al*., 2012). These advances have also brought about the characterisation of various strains of *M. tuberculosis* and the differences on multiple biological levels between these strains are currently being extensively explored. The biological grounds of pathogenicity and disease presentation due to *M. tuberculosis* infection are largely unknown, therefore research on both bacterial and host level continues.

The tools that are currently employed to monitor and control TB outbreaks do not provide sufficient evidence about the directionality and the sequence of transmission events for cases arising over a short period of time and fail to capture the true dynamics of an outbreak (Roetzer *et al*., 2013). The unprecedented resolution that WGS offers, enables the study of molecular epidemiology over short periods of time as well as the micro-evolutionary events observable during transmission (Walker *et al*., 2013). WGS permits the assessment of the aptitude of *M. tuberculosis* to mutate over the course of an infection in its natural host context with minimum bias and optimum sensitivity (Bryant *et al*., 2013; Ford *et al*., 2012). Whole genome sequencing of MTC members is quickly becoming the gold standard for various research applications for the simple reason that it provides an unparalleled means to detect genetic diversity between isolates.

To date, whole genome sequencing has been applied to evolutionary studies, transmission investigations and preliminary studies to observe the dynamics of host-pathogen co-evolution. The potential of whole genome sequencing as a molecular epidemiology tool has been recognised due to its sensitivity to detect intermittent genetic events, and the comprehensive ability to detect several forms of genetic variation. It has been demonstrated that whole genome sequencing has the capability to replace current genotyping methods used in molecular epidemiology. Future technologies aim to improve the chemistry in order to produce longer read lengths with the aim to resolve these repeat elements and decrease the cost involved.

2.7. THE LATIN AMERICAN-MEDITERRANEAN GENOTYPE

2.7.1. INTRODUCTION

Epidemiological studies have revealed two major families of *M. tuberculosis* to be the cause of TB globally, the Beijing- and the LAM family. The Beijing family of *M. tuberculosis* has been extensively reviewed and is the topic of many TB-related studies – present and past. The LAM-genotype is clearly of great importance due to its high prevalence, not only in South-Africa but also in other African countries such as Zambia and Zimbabwe as well as Latin-America and the Mediterranean regions

(Chihota *et al*., 2007; Lazzarini *et al*., 2007), and the need for an in-depth investigation of this genotype is apparent.

The LAM family of strains has been the subject of many studies during the last decade due to its world-wide prevalence. Its position in the evolution of *M. tuberculosis* is shown in Figure 2.6.1, which represents a summary of the general evolution of *M. tuberculosis* as previously reported (Baker *et al*., 2004; Gutacker *et al*., 2002; Hershberg *et al*., 2008). A wide range of phenotypic and genetic characteristics are particularly associated with the LAM lineage of *M. tuberculosis*.

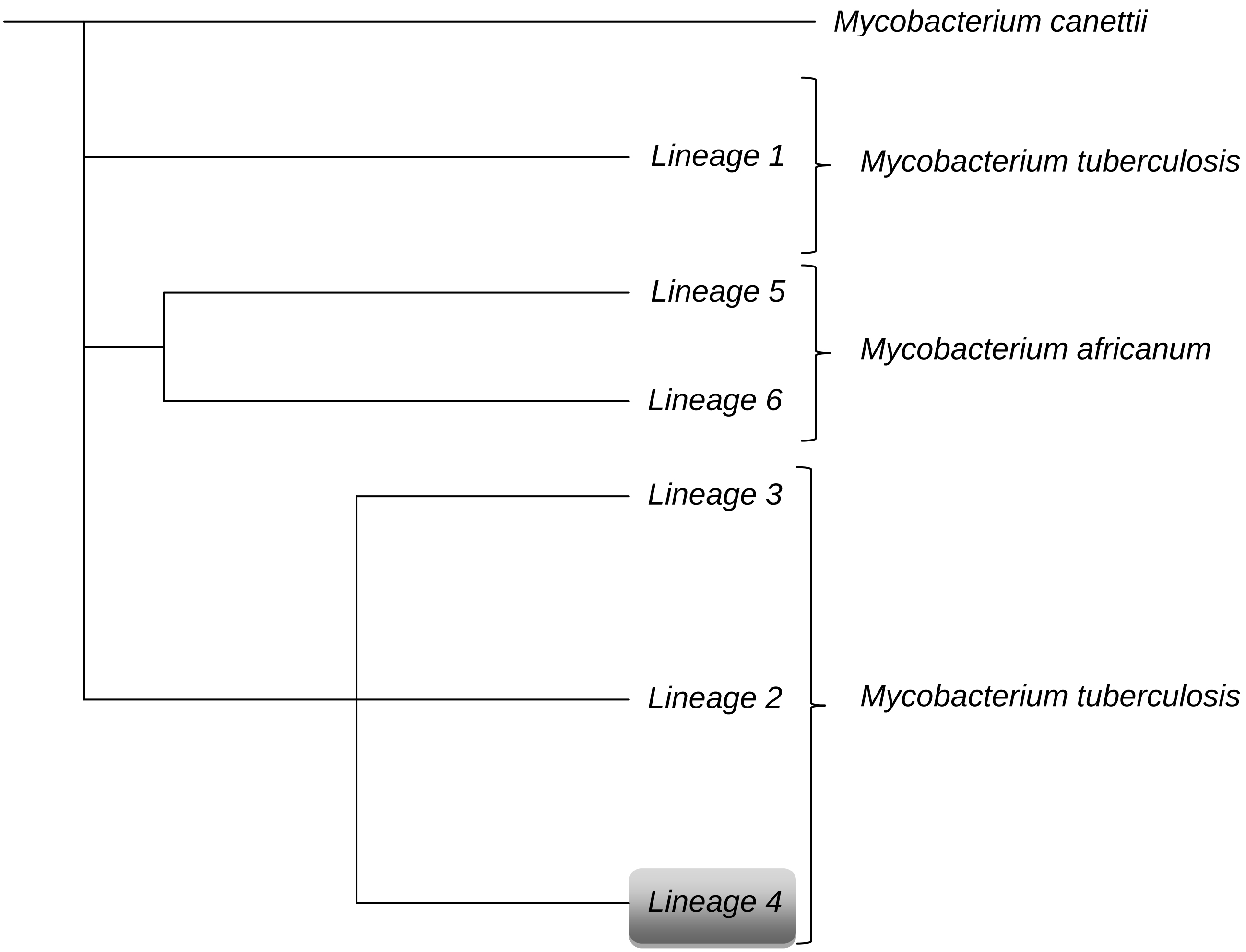


Figure 2.6.1. A summary of the evolution of *M. tuberculosis*, showing the evolutionary position of lineage 4 (to which the LAM genotype belongs)

2.7.2 EPIDEMIOLOGY

The LAM lineage is endemic to Latin-America and the Mediterranean region, which includes coastal regions of three continents; Europe, Asia and Africa, and was later introduced into North America

(Fanning, 1994; Lazzarini *et al*., 2007). Worldwide, the group 2 LAM genotype strains of *M. tuberculosis* are the cause of approximately 15% of all TB cases (Brudey *et al*., 2006). Tuberculosis caused by a subset of the LAM genotype of *M. tuberculosis*, termed RD^{Rio} *M. tuberculosis* was found in four continents out of five examined and is thus not confined to a specific area (Gibson *et al*., 2008). LAM and Beijing genotypes of *M. tuberculosis* commonly cause tuberculosis in Russian prisons (Drobniewski *et al.*, 2002; Shemyakin *et al.*, 2004; Ignatova *et al.*, 2006). The majority of LAM strains have been shown to contain an IS*6110* in a unique position in the *plcA* gene. This specific IS*6110* insertion site serves as a genetic marker for the LAM-RUS family (Dubiley *et al*., 2007).

It has recently been shown that the Euro-American lineage, which includes the LAM lineage, is largely accountable for MDR-TB outbreaks in Latin-America and Spain (Ritacco *et al*., 2011). The LAM genotype is a prevalent cause of tuberculosis in South Africa, especially in the Western Cape and KwaZulu-Natal provinces of the country (Streicher, 2007; Victor *et al*., 2004). In Cape Town, Western Cape, South Africa, the strain responsible for causing the highest number of TB cases is LAM strain family F11. A 2004 study reported that 21.4% of TB cases in this area are attributed to this strain (Victor *et al*., 2004). F11 was also shown to cause TB in several other countries around the world, on various continents (Victor *et al*., 2004). The LAM3 lineage was only first observed to cause tuberculosis in Asia in 2008, therefore it was accepted that LAM strains were not a very common cause of TB in Asia (Chuang *et al*., 2008).

2.7.3. GENETIC CHARACTERISTICS

A 7-bp region of difference in the *pks15/1* (polyketide synthase) gene was shown to differentiate the major Euro-American lineage of *M. tuberculosis* (Gagneux *et al*., 2006; Gibson *et al*., 2008; Marmiesse *et al*., 2004). Interestingly, the Euro-American lineage consists of both members of PGG2 and PGG3 of *M. tuberculosis*, but not ancestral strains (i.e. members of PGG1) (Huard *et al*., 2006; Marmiesse *et al*., 2004). Furthermore, it was shown that the 7-bp deletion in the *pks15/1* gene is associated with the *kat*G463 SNP, CGG (Gibson *et al*., 2008).

Genome-wide SNP analysis has been used to infer phylogeny of the strains of *M. tuberculosis* (Baker *et al*., 2004; Filliol *et al*., 2006; Gutacker *et al*., 2006). Nine phylogenetic lineages were distinguished by Gutacker *et al.* (2006), SNP cluster VI consists of ~ 80% LAM strains and a subset of T strains (Gutacker *et al*., 2006). Large sequence polymorphisms (LSPs) were used by Gagneux *et al.* (2006) to create a global phylogenetic tree, which gave similar results. A SNP has been identified that is situated in the *mgtC* gene that distinguishes Haarlem strains from non-Haarlem strains, which are closely related to the LAM genotype of *M. tuberculosis* (Alix *et al*., 2006; Gagneux *et al*., 2006). A sub-group of the LAM lineage, designated the LAM-RUS family, are characterised by a unique IS*6110* insertion in the *plcA* gene, that encodes a phospholipase C protein (Dubiley *et al*., 2007). The LAM-RUS family has been associated with multidrug-resistance and is a prevalent cause of disease in prisons in Tula and Moscow in Russia (Ignatova *et al*., 2006; Shemyakin *et al*., 2004). A LAM lineage specific SNP was identified in the gene encoding *Ag85C* (*Rv0129c*) at position 103 (Musser *et al.*, 2000). This SNP was later shown to be present in most LAM strains tested, as well as in a small number of non-LAM strains, the specificity and sensitivity of the Ag85C^{103} polymorphism in a global collection of strains are relatively high (Gibson *et al*., 2008). A study investigating SNP markers for antibiotic resistance has shown that a SNP in the *thyA* gene (Thr202Ala) does not confer resistance to *para*-aminosalicylic acid as hypothesized, but instead serves as a marker for the LAM lineage of *M. tuberculosis* (Feuerriegel *et al*., 2010). A synonymous LAM-specific SNP in *ligB1212* was reported in a lineage assignation study employing high-throughput SNP typing of the 3R genes (Abadia *et al*., 2010). Recently cell wall biosynthesis-associated genes were screened for SNPs to further uncover the phylogeny of *M. tuberculosis* strains. A SNP in *fbpC* codon 103 (synonymous SNP, GAG to GAA) was identified in all LAM isolates and correlated with the findings of Gibson *et al.* (2008) (Chuang *et al*., 2010; Gibson *et al*., 2008). A SNP in codon 107 (synonymous SNP, GGC to GGT) of the *pimB* gene was identified as a robust marker for lineage IV, which includes LAM, Haarlem and T strains of *M. tuberculosis* (Chuang *et al*., 2010).

In 2007, a sublineage of LAM strains was identified that is the major cause of TB in Rio de Janeiro, Brazil. These strains are distinguished by a 26-kb deletion, the largest genomic deletion described in *M. tuberculosis* to date (Lazzarini *et al*., 2007). Ten genes are affected by the RD^{Rio} deletion, including PPE55 and PPE56. Recently RD^{Rio} LAM genotype strains of *M. tuberculosis* have been associated with MDR-TB in Ibero-America as well as in Portugal (Ritacco *et al*., 2011). RD^{Rio} is associated with the West African sub-lineage of the major Euro-American lineage of *M. tuberculosis* (Gibson *et al*., 2008).

RD174 is associated with the RD^{Rio} genetic marker found in a subset of members of the LAM family of *M. tuberculosis*, and is not present in any other (non-RD^{Rio}) LAM strains (Gagneux *et al*., 2006; Gibson *et al*., 2008). The 3 650-bp region that is deleted in RD174 includes genes involved in the hypoxia-induced regulon (Tsolaki *et al*., 2004). RD174 was identified prior to the massive RD^{Rio} deletion and was shown to be associated with RD149 in a study characterising a collection of strains from San Francisco. RD149 includes the open reading frames (ORFs) *Rv1572c – Rv1586c* (Kato-Maeda *et al*., 2001; Musser *et al*., 2000; Tsolaki *et al*., 2004). The RD^{Rio} subset of LAM genotype strains is commonly associated with LAM9 and LAM1 (Lazzarini *et al*., 2007). Furthermore, LAM1 strains exclusively possess the RD^{Rio} deletion (Gibson *et al*., 2008). These were concordant with the South

African F9 and F13 families (respectively LAM11 and LAM1) also commonly associated with the RD^{Rio} genetic marker (Gibson *et al*., 2008).

The transmission and evolution of an XDR (previously MDR) F15/LAM4/KZN strain of *M. tuberculosis* in a South African setting were described in a study by Pillay and Sturm (2007). The rapid evolution from MDR to XDR and persistence of the strain was attributed to its effective transmission ability and the selective pressure of the standard TB treatment received by patients presenting with pulmonary TB symptoms (Pillay and Sturm, 2007).

A group of strains, named the South Africa 1 (SAF1) family, was identified, and proved to be the predominant cause of TB in Zimbabwe (it caused 47.2% of TB cases included in the study) and Zambia (65%) (Chihota *et al*., 2007). These strains belong to the South African F9 family of LAM strains and are thus part of the RD^{Rio} sublineage of the LAM lineage of *M. tuberculosis* (Chihota *et al*., 2007; Gibson *et al*., 2008). Gagneux *et al.* (2006) proposed that certain strains of *M. tuberculosis* are adapted to cause disease in a specific human population which may in part explain the selection and dissemination of SAF1 in Zambia and Zimbabwe (Chihota *et al*., 2007; Gagneux *et al*., 2006). This again suggests that there is some level of host-pathogen compatibility contributing to the predominance of a specific strain of *M. tuberculosis* largely causing disease in a geographically related community (Gagneux *et al*., 2006).

In 2010, a large duplication of 350-kb was reported in two strains of the Beijing family of *M. tuberculosis.* This duplication event spans *Rv3128c* to *Rv3427c* and includes the *dosR* regulon (Domenech *et al*., 2010). A subsequent study found that this duplication is also present in lineage two and four of *M. tuberculosis*, and is thus not limited to the Beijing family. The large duplication has different boundaries in different strains, which indicates that it originated from distinct duplication events (Weiner *et al*., 2012).

In summary, the LAM lineage of *M. tuberculosis* possess the following genetic characteristics: i) LAM strains are members of PGG2 as defined by the *katG* and *gyrA* polymorphisms, and phylogenetic cluster VI as defined by Gutacker *et al.* (2002), ii) contain a 7-bp deletion in the *pks15/1* gene, iii) the Ag85C103 SNP, iv) the site-specific IS*6110* insertion element, and v) a typical spoligotype pattern where spacers 21 to 24 and -33 to 36 are deleted (Brudey *et al*., 2006; Gibson *et al*., 2008; Gutacker *et al*., 2006, 2002; Sreevatsan *et al*., 1997). However, in a lineage assignation study utilising a high-throughput 3R SNP typing method and spoligotypes for lineage calling ability, it was shown that spacers 21 and 22 are not as definitively informative for LAM strains as the deleted spacer 23 (Abadia *et al*., 2010).

The genetic markers (SNPs, RDs and IS*6110* elements) associated with the group 2 LAM genotype of *M. tuberculosis* are summarised Table 2.7.1.

Table 2.7.1. Summary of the genetic characteristics of the group 2 LAM genotype.

Characteristic	Details	Position	Start	Stop	Size	Gene/genes involved	Reference
RD115	*M. tuberculosis* KZN 4207, 1435,605, V2475, R506; 98-R604 (LAM5); F11 (LAM3)	n.a	453364	455971	2 607 bp	*Rv0376c - Rv0378*	(Kato-Maeda *et al.*, 2001; Tsolaki *et al.*, 2004)
RD761	Unique to a South African strain from the Euro-American lineage	n.a	1502787	1503881	1094 bp	*Rv1334-Rv1336*	(Gagneux *et al.*, 2006)
RD149	RD174 is associated with the presence of RD 149, subgroup of LAM strains	n.a	1779264	1788512	9248 bp	*Rv1572c-Rv1578c*	(Kato-Maeda *et al.*, 2001; Tsolaki *et al.*, 2004)
RD174	Associated with the RD^{Rio} deletion	n.a	2237049	2240699	3650 bp	*Rv1992c-Rv1997*	(Gibson *et al.*, 2008)
Deletion in *pks15/1*	Micro deletion specific Euro-American lineage	n.a	-	-	7 bp	*Rv2946c - Rv2947c*	(Marmiesse *et al.*, 2004)
RDRio	Subgroup of LAM strains	n.a	-	-	26 314 bp	*Rv3345c/PE_PGRS50, Rv3346c, Rv3347c/PPE55, Rv3348, Rv3349c, Rv3350c/PPe56, Rv3351c, Rv3352c, Rv3353c, Rv3355c*	(Lazzarini *et al.*, 2007)
Deletion in *gidB*	KZN MDR and XDR				130 bp	*gidB*	(Ioerger *et al.*, 2009)
IS*6110* in *plcA*	IS*6110* in *plcA* gene, specific to LAM-RUS family	n.a	-	n.a	n.a	*plcA*	(Dubiley *et al.*, 2007)
SNP in *katG*	PGG2 defining SNP	codon 95	n.a	n.a	n.a		
SNP in *mgtC*	Distinguishes Haarlem strains from non-Haarlem strains (GCG to CAC)	Codon 182	n.a	n.a	n.a	*MgtC*	(Alix *et al.*, 2006)
SNP in *fbpC*	LAM lineage specific, GAG (Glu) to GAA (Glu) SNP in fbpC	Codon 103	n.a	n.a	n.a	*FbpC*	(Chuang *et al.*, 2010; Gibson *et al.*, 2008; Musser *et al.*, 2000)

Table 2.7.1. Summary of the genetic characteristics of the group 2 LAM genotype. (Continues from previous page)

Characteristic	Details	Position	Start	Stop	Size	Gene/genes involved	Reference
SNP in *rrs*	C-to-T transition, 491 of the rrs gene Specific to F11	491 (nt	n.a	n.a	n.a	*rrs*	(Victor *et al.*, 2001)
SNP in *ligB*	TCC (ser) to TCG (ser)	1212 (nt)	n.a	n.a	n.a	*LigB*	(Abadia *et al.*, 2010)
SNP in *Rv0129c*	LAM lineage specific SNP	codon 103	n.a	n.a	n.a	*Ag85C* (*Rv0129c*)	(Gibson *et al.*, 2008; Musser *et al.*, 2000)
SNP in *pimB*	Robust marker for Lineage IV	codon 107				*pimB*	(Chuang *et al.*, 2010)
SNP in *thyA*	Thr202Ala Phylogenetic marker for LAM genotype	codon 20			n.a	*thyA*	(Feuerriegel *et al.*, 2010)
Duplication of dosR region	Duplication in various *M. tuberculosis* lineages	-	-	-	350 Kb	*Rv3128c – Rv3427c*	(Weiner *et al.*, 2012)

2.7.4. PHENOTYPIC INFORMATION

It has been shown that the genetic diversity observed in *M. tuberculosis* strains has important clinical relevance, in that different strains may present with differing disease phenotypes (Thwaites *et al*., 2008). In addition to causing a severe form of pulmonary disease, the LAM family of strains exhibits a fitness advantage over non-LAM strains when growth rates were compared *in vitro*, emphasising the success of this group of strains (Von Groll *et al*., 2010).

It was shown that different lineages of *M. tuberculosis* have specific patterns of cytokine induction and growth. Lineage four strains induce a more severe immune response in the host shown by high levels of interleukin (IL) 12p40 and tumour necrosis factor (TNF), in comparison with lineage 2 and 3. This suggests that strains have differing *in vivo* capabilities to optimally utilise the host milieus in different human populations (Sarkar *et al*., 2012).

An association was shown between *M. tuberculosis* RD^{Rio} infection and pulmonary cavitation, indicating that the severity of disease caused by the RD^{Rio} LAM genotype of *M. tuberculosis* was increased in a certain setting. The phenomenon was attributed to a strategy of the strain to increase transmission (Lazzarini *et al*., 2007; Weisenberg *et al*., 2011). In an HIV co-infected population it was concluded that TB caused by SAF1/RD^{Rio} *M. tuberculosis* strains is associated with smoking, but strangely not with cavitary lung disease (Chihota *et al*., 2011).

A description of the transmission and evolution of an XDR (previously MDR) F15/LAM4/KZN strain of *M. tuberculosis* in a South African setting showed rapid evolution from MDR to XDR. Persistence of the strain was attributed to its effective transmission ability and the selective pressure of the standard TB treatment received by patients presenting with pulmonary TB symptoms (Pillay and Sturm, 2007).

LAM10 has shown a strong association with isoniazid resistance (Brown *et al*., 2010). The LAM F11 strain has been identified as the most successful TB-causing strain in communities of the Western Cape Province in South Africa (Warren *et al*., 1999, 2002). LAM F11 drug-resistant isolates were described in a study setting in the Western Cape of South Africa (Victor *et al*., 2004).

2.8. CONCLUSION

Decades of research on tuberculosis and the etiologic agent of the disease (*M. tuberculosis*), its mechanisms of pathogenesis and persistence, have shown it to be an extremely successful pathogen of both humans and animals. In order to combat a disease of any sort, the biological means in which the agent exhibits pathogenicity has to be understood. Research on the evolution and the biological grounds that account for strain differences and varying disease phenotypes of different strains of *M. tuberculosis* is thus potentially of great public health benefit.

The LAM lineage of *M. tuberculosis* does not only pose a world-wide threat, but is also extremely prevalent in local settings, including central southern Africa (Zambia and Zimbabwe) and the Western Cape and KwaZulu-Natal provinces of South-Africa (Chihota *et al*., 2007; Mulenga *et al*., 2010; Viegas *et al*., 2010). A significant knowledge gap remains in the understanding of the evolution and diversification of the LAM lineage of *M. tuberculosis* strains, and the genetic and proteomic traits underlying its fitness advantage. Further delineating the strains of *M. tuberculosis* will aid in the comprehension of host-pathogen compatibility, population patterns, and the evolution of drug resistance and could ultimately advance drug- and vaccine development to combat the disease.

CHAPTER 3

MATERIALS & METHODS

3.1. INTRODUCTION

All procedures involving the handling of live *M. tuberculosis* cultures were done in the Biosafety level 3 (BSL3) laboratory in the Division of Molecular Biology and Human Genetics, 4th floor, FISAN building, Faculty of Medicine and Health Sciences, Stellenbosch University, Tygerberg, South Africa. Strict BSL3 safety procedures were followed as described in the in-house departmental BSL3 safety manual. Care was taken to prevent any contamination of the bacterial cultures or human exposure to pathogenic mycobacterial cultures.

The work described in the present study forms part of a large on-going project which received ethical approval from the Stellenbosch University Health Research Ethics Committee under the title: "An investigation into the evolutionary history and biological characteristics of the members of genus *Mycobacterium*, with specific focus on the different strains of *M. tuberculosis*, other members of the *M. tuberculosis* complex and non-tuberculosis Mycobacteria (NTM)", ethics reference number: N10/04/126.

Methods for preparing buffers and reagents used in the experiments described in this chapter are included in Appendix A. All oligonucleotide sequences used to verify nucleotide variants identified by whole genome analysis are listed in Appendix B.

The whole genome sequencing data analysis pipeline described in this dissertation was developed by the candidate to analyse the data generated for this project, as well as other ongoing research projects. All bioinformatics was done by the candidate unless stated otherwise (e.g. scripts written by colleagues).

Sputum specimens from all new and retreatment tuberculosis cases were collected from patients who attended primary healthcare clinics and who resided in an epidemiological field site in Cape Town, Western Cape, South Africa, from January 1993 to December 2004. These samples were subjected to sputum smear microscopy (fluorescent staining) and culturing in a BACTEC 460, MGIT 960 and Löwenstein-Jensen (LJ) medium by the National Health Laboratory Service (NHLS) or Stellenbosch

University laboratories to confirm the presence of *M. tuberculosis*. Confirmed *M. tuberculosis* isolates were subject to standard genetic characterisation by IS*6110* RFLP analysis and spoligotyping (Kamerbeek *et al*., 1997; Van Embden *et al*., 1993). Well characterized drug susceptible *M. tuberculosis* isolates (n=50) were selected from this established sample bank for the present study. This isolate panel represents a mixture of different genotypes found in the Western Cape setting.

3.2. SAMPLE SELECTION

M. tuberculosis strains representative of the three principle genetic groups (PGG) were selected from the longitudinal sample bank. These strains were cultured for DNA extraction and the DNA was subjected to whole genome sequencing. The genetic population structure of this sample set is summarised in Table 3.2.1.

Table 3.2.1 Summary of the *M. tuberculosis* strains subjected to whole genome sequencing.

Local *Mycobacterium tuberculosis* strains analysed				
Principle genetic group	**IS*6110* family**	**SAWC number**	**Strain ID**	**Spoligotype pattern**
PGG 1	34	995	CAS1/Delhi	■■■□□□□■■■■■■■■■■■■■■■□□□□□□□□□□□□■■□□■■■■■
	29	1116	Typical Beijing, sublineage 4	□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□■■■■■■■■■
	29	1125	Typical Beijing, sublineage 6	□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□■■■■■■■■■
	27	1453	Atypical Beijing, sublineage 3	□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□■■■■■■■■■
	unidentified	1659	EAI	■■■■□■■■■■■■■■■■■■■■■■■■■■■□□□□■□■■■■■■■■■
	27	2701	Typical Beijing, sublineage 2	□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□■■■■■■■■■
	25	3385*	Cas1/Delhi/Kili	■■■□□□□■□■■■■■■■■■■■■■□□□□□□□□□□□□■■■■■■■■■
	33	3740	Cas/Kili	■■■□□□□■■□■■■■■■■■■□□□□□□□□□□□□□□□□■■■■■■■■
	20	4370	Cas1/Delhi	■■■□□□□■■■■■■■■■■■■■■■□□□□□□□□□□□□□□□■■■■■■
	29	4437	Typical Beijing, sublineage 7	□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□■■■■■■■■■
	29	4570	Typical Beijing, sublineage 5	□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□■■■■■■■■■
	31	6519	Atypical Beijing, sublineage 1	□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□■■■■■■■■■
PGG 2	32	1123*	-	■■■■■■■■■■■■■■■■■■■■■■□□□□□□□□□■□□□□■■■■■■■
	28	1595	Quebec	■■■■■■■■□□■■■■■■■■■■■■■■■■■■■■■■□□□□■■■■■■■
	9	1955	LAM 11/ZWE	■■■■■■■■■■■■■■■■□□■■□□□□■■□□□□■■□□□□■■■□■■■
	30	2005*	-	■■■■■■■□□□■■■■■■■■□□□□□□□■■■■■■□□□□□□■■■■■■
	24	2026*	Pre-Haarlem/X1	■■■■■■■■■■■■■■■■■□■■■■■■■■■■■■■■□□□□■■■■■■■
	26	2262	U	■■■■■■■■■■■■■■■■■■■■□□□□■■■■■■□□□□□□■■■■■■■■
	2	2282	Haarlem 3	■■■■■■■■■■■■■■■■■■■■■■■□□■■■■■■□■□□□□■■■■■■■
	140	2336*	LCC, 4 bander	■■■■■■■■■■■■■■■■■□■■■■■■■■■■■□□□□□□□■■■■■■■

	13	2511*	LAM 1	■■□■■■■■■■■■■■■■■■■■□□□□■■■■■■■■□□□□■■■■■■■
	15	2576*	LAM 4, KZN	■■■■■■■■■■■■■■■■■■■■□□□□■■■■■■■■□□□□■■■□■■■
	13	2904	LAM 1	■■□■■■■■■■■■■■■■■■■■□□□□■■■■■■■■□□□□■■■■■■■
	14	3100*	T1, T1-Tusc, T5/Rus	■■■■■■■■■■■■■■□□□□□□□■□□■■■■■■□□□□□□■■■■■■■
	130	3200*	LCC, 3 bander	■■■■■■■■■■■■■■■■■□■■■■■■■■■■■■■■□□□□■■■□□■■
	14	3388	T1, T1-Tusc, T5/Rus	■■■■■□□□□■■■■■□□□□□□□■□□■■■■■■□□□□□□■■■■■■■
	1	3448*	Haarlem 1	■■■■■■■■■■■■□■■■■■■■■■■■■□□□□□□■□□□□■■■■■■■
	9	3517	LAM 11/ZWE	■■■■■■■■■■■■■■■■■■■■□□□□■■□□□□■■□□□□■■■■■■■
	14	3651	T1, T1-Tusc, T5/Rus	■■■■■■■■■■■■■■□□□□□□□□□■■■■■■■■■□□□□■■■■■■■
	26	3656*	U	■■■■■■■■■■■■■■■■■■■■□□□□■■■■■■□□□□□□■■■■■■■
	110	3933*	LCC, 1 bander, T1	■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□□□■■■■■■■
	6	4046	Haarlem-like T4/Ceu 1	■■■■■■■■■■■■■■■■■■□■■■□□■■■■■■■■□□□□■■□■■■■
	160	4336*	LCC, 6 bander	■■■□□□□□□□□□□□■■■□■■■■■■■■■■■■■■□□□□■■■■■■■
	11	4498	LAM 3	■■■■■■■□□□□■■■■■■■■□□□□□■■□■■■■■□□□□■■■■■■■
	150	4972*	LCC, 5 bander	■■■■■■■■■■■■■■■■■□■■■■■■■■■■■■■■□□□□■■■■■■■
	7	4978*	Haarlem-like T1	■■■■■■■■■■■□□□□□■■■□□□□□■□■■■■■■□□□□■□□■■■■
	11	5165	LAM 3	■■■■■■■■□□□■■■■■■■■■□□□□■■■■■■■■□□□□■■■■■■■
	26	5260*	U	■■■■■■■■■■■■■■■■■■■■□□□□■■■■■■□□□□□□■■■■■■■
	13	5276*	LAM 1	■■□■■■■■■■■■■■■■■■□■□□□□■■■■■■■■□□□□■■■□■■■
	4	6680	Haarlem 3	■■■■■■■■■■■■□■■■■■■■■■■■■■■■■■■■□□□□■■■■■■■
	10	7367	Haarlem 1	■■■■■■■■■■■■■■■■■■■■■■■■■□□□□□□■□□□□■■■■■■■
	22	337*	T1	■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□□□■■■■■■■
PGG 3	5	478*	T2	■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□□□■■■□■■■
	17	1397*	T1	■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□□□■■■■■■■
	16	1543*	-	■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□□□■■■■■■■
	18	3839*	T1	■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□□□■■■■■■■
	8	4302	U	■■■■■■□□□□□□□□□□□□■■■■■■□□□□□□□□□□□□□□□□□□□
	23	4472*	T	■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□□□■■■■■■■
	21	5330*	T5	■■■■■■■■■■■■■■■■■■■■■■□■■■■■■■■■□□□□■■■■■■■
	12	5440*	-	■■■■■■■■■■■■■■■■■■■■■■□■■■■■■□□□□□□□□□□□□□■■

An asterisk (*) next to the SAWC number of an isolate indicates that the spoligotype signature was not derived from traditional spoligotyping, but was determined using the WGS data for that isolate using SpolPred (Coll *et al*., 2012).

3.3. MYCOBACTERIAL CULTURING FOR DNA EXTRACTION

Clinical isolates of *M. tuberculosis* that were genetically characterised by IS*6110* DNA fingerprinting and spoligotyping using the internationally standardised methods, were used. The selected isolates were initially cultured in BACTEC 460 tubes (12B media) or mycobacterial growth indicator tubes

(MGITs) at 37°C from frozen stock cultures until positive growth was detected by means of the BACTEC 460, 960 instruments (BD Biosciences, USA), respectively.

Upon detection of growth the isolates were allowed to incubate until visible growth was observed. The BACTEC 460 or MGIT tubes were briefly centrifuged to collect bacterial growth. Two LJ slants were prepared for each isolate, by adding 100 µl of the bacterial growth collected from the bottom of the tube to each of two clearly marked LJ slants. The LJ slants were incubated at 37˚C with regular aeration for approximately three weeks or until sufficient growth was observed.

3.4. EXTRACTION OF DNA

DNA was extracted as previously described (Warren *et al*., 2006). In summary, after sufficient growth was observed, the outer surface of the LJ slants were decontaminated with Incidin Plus (Ecolab, Minnesota, USA) and placed in a biosafety autoclave bag and transferred to a pre-warmed fan oven at 80°C for 1 hour to ensure heat killing. Colonies were gently scraped with a disposable 10 µl loop from the surface of the media and transferred to a 50 ml tube containing approximately 20 glass beads (4mm in diameter) and 6 millilitres of extraction buffer (5% sodium glutamate, 50 mM Tris-HCl (pH 7.4) and 25mM EDTA)). The 50 ml tubes containing the extraction buffer cell suspension were vortexed to disrupt the colonies. Lysozyme (25 mg, Roche, Germany) and RNAse A (25 µg, Roche, Germany) were then added and incubated after gentle mixing for 2 hours at 37°C to degrade the cell wall and to digest RNA. Thereafter, 600 µl of 10x proteinase K buffer (5% sodium dodecyl sulphate, 100nM Tris-HCl (pH 7.8), 50mM EDTA) and 1.5 mg proteinase K (150 µl of a 10 mg/ml stock solution) were added and the suspension was incubated for a further 16 hours at 45°C to digest all bacterial proteins. An equal volume of phenol/chloroform/isoamyl alcohol (24:23:1) was added and mixed intermittently over a period of 2 hours at room temperature. After centrifugation at 3000 x g for 20 minutes, the aqueous phase was aspirated and an equal volume of chloroform/isoamyl alcohol (24:1) was added. After centrifugation at 3000 g for 20 minutes, the extracted DNA was precipitated with the addition of 600 µl 3 M sodium acetate (pH 5.2) and an equal volume of isopropanol, and the precipitate immediately collected on a glass rod. The DNA was washed with 70% ethanoland was left to dry on the glass rod. The air-dried DNA was re-dissolved in 300 µl TE (10 mM Tris-HCl (pH 8.0), 1mM EDTA) and stored at -20°C. The concentration of the extracted DNA was measured using a NanoDrop spectrophotometer (Thermo Fisher Scientific, Waltham, Massachusetts, USA).

3.5. NEXT GENERATION SEQUENCING (NGS) OF ISOLATES3.5.1. NGS INTRODUCTION

Sequencing and library preparation was done in collaboration with Dr Arnab Pain and Dr Abdallah M. Abdallah from the King Abdullah University of Technology (KAUST), Saudi Arabia, and Dr Ruth McNerney and Dr Taane Clark from the London School of Hygiene and Tropical Medicine (LSHTM), UK. The DNA isolated from the *M. tuberculosis* isolates selected for the present study were subjected to whole genome sequencing on the Illumina HiSeq2000 (Illumina, California, USA) platform using a paired-end approach, with 500 base fragment sizes which resulted in insert sizes between 350 and 550 bases. The depth of coverage for all the isolates was estimated to be at least above 100 times. This, together with the excellent quality of the sequencing data, ensured a high level of confidence for identifying variation in the genomes of interest.

3.5.2. DATA SOURCES

The fastq whole genome sequence data files of the sequenced *M. tuberculosis* isolates were provided by KAUST and were shared via a secure file transfer protocol (FTP). The reference genomes (*M. tuberculosis* H37Rv and *M. tuberculosis* F11) in fasta format, genome summary information and summary of gene sequences were downloaded from the TB database (TBDB) at http://genome.tbdb.org/annotation/genome/tbdb/MultiDownloads.html. Other previously published *Mycobacterium sp.* genome sequences that were included in the analyses were downloaded from publically available databases and are summarised in Table 3.5.1

Table 3.5.1 Published *Mycobacterium* sp. genomes included in the present study

Strain name	Reference	Source
***M. tuberculosis* KZN 4207 (DS)**	"*M. tuberculosis* Comparative Sequencing Project, Broad Institute of Harvard and MIT (http://www.broadinstitute.org/)"	ftp://ftp.broad.mit.edu/pub/annotation/mtuberculosis/kzn
***M. tuberculosis* KZN 1435 (MDR)**	"*M. tuberculosis* Comparative Sequencing Project, Broad Institute of Harvard and MIT (http://www.broadinstitute.org/)"	ftp://ftp.broad.mit.edu/pub/annotation/mtuberculosis/kzn
***M. tuberculosis* KZN 605 (XDR)**	"*M. tuberculosis* Comparative Sequencing Project, Broad Institute of Harvard and MIT (http://www.broadinstitute.org/)"	ftp://ftp.broad.mit.edu/pub/annotation/mtuberculosis/kzn
***Mycobacterium africanum* lineage 5 (5444)**	(Comas *et al*., 2010)	ftp://ftp.broad.mit.edu/pub/annotation/mtuberculosis/diversity
M. tuberculosis	(Comas *et al*., 2010)	ftp://ftp.broad.mit.edu/pub/annotation/mtu

(LAM) 1503		berculosis/diversity
***Mycobacterium africanum* lineage 5 (11821_03)**	(Comas *et al*., 2010)	ftp://ftp.broad.mit.edu/pub/annotation/mtuberculosis/diversity
***Mycobacterium africanum* lineage 6 (0981)**	(Comas *et al*., 2010)	ftp://ftp.broad.mit.edu/pub/annotation/mtuberculosis/diversity
***Mycobacterium canettii* (K116)**	(Comas *et al*., 2010)	ftp://ftp.broad.mit.edu/pub/annotation/mtuberculosis/diversity
***M. tuberculosis* Erdmann**	(Miyoshi-Akiyama *et al*., 2012)	ftp://ftp.ddbj.nig.ac.jp/ddbj_database/dra/fastq/SRA008/SRA008875/SRX005707/
***M. tuberculosis* lineage 7 (Mt256)**	(Blouin *et al*., 2012)	http://www.ebi.ac.uk/ena/data/view/ERP001885

3.5.3. BIOINFORMATICS ANALYSIS OF WGS DATA

Computational analysis was done in collaboration with Professor Alan Christoffels and Ms Mmakamohelo Direko (South African National Bioinformatics Institute (SANBI), University of the Western Cape, South Africa), (initial Linux training and advice on software packages to use for NGS data analysis), Dr Margaretha de Vos, Dr Ruben van der Merwe, and Mrs Michelle Daya (Department of Molecular Biology and Human Genetics, Stellenbosch University), (writing of scripts for data processing and annotation). Due to the large amounts of data included in the present study and the lack of feasibility to verify the magnitude of variants present in the array of isolates analysed with PCR-Sanger sequencing, a multi-software approach was used to minimise the identification of false positive variants. Only variants (polymorphisms) called by all software packages employed were regarded as high-confidence polymorphisms.

The workflow for the analysis of the whole genome sequences produced on the Illumina platform is summarised in Figure 3.5.1.

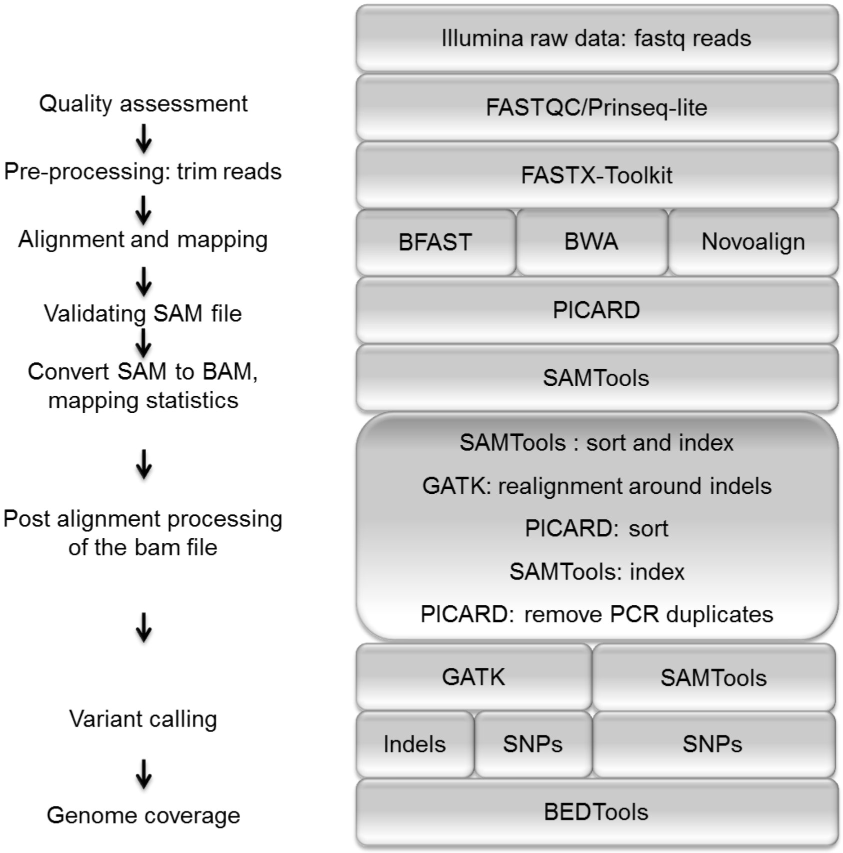
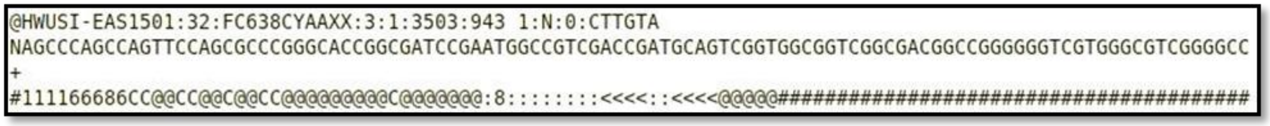

Illumina raw data: fastq reads
Quality assessment
FASTQC/Prinseq-lite
Pre-processing: trim reads
FASTX-Toolkit
Alignment and mapping
BFAST
BWA
Novoalign
Validating SAM file
PICARD
Convert SAM to BAM, mapping statistics
SAMTools
Post alignment processing of the bam file
SAMTools : sort and index
GATK: realignment around indels
PICARD: sort
SAMTools: index
PICARD: remove PCR duplicates
Variant calling
GATK
SAMTools
Indels
SNPs
SNPs
Genome coverage
BEDTools

Figure 3.5.1. The summarised workflow of the computational analysis of whole genome sequencing data.

3.5.4. FASTQ FILE FORMAT

The fastq format (Figure 3.5.2) is a text-based format for the storing of biological sequences and the corresponding quality information as it is recorded from the sequencer (Cock *et al*., 2010).



```
@HWUSI-EAS1501:32:FC638CYAAXX:3:1:3503:943 1:N:0:CTTGTA
NAGCCCAGCCAGTTCCAGCGCCCGGGCACCGGCGATCCGAATGGCCGTCGACCGATGCAGTCGGTGGCGGTCGGCGACGGCCGGGGGTCGTGGGCGTCGGGGCC
+
#111166686CC@@CC@@C@@CC@@@@@@@@@C@@@@@@@:8::::::::<<<<::<<<<@@@@@#######################################
```

Figure 3.5.2. A representative of a typical read from a fastq file produced by the Illumina sequencing platform version 1.5.

The data contained in these files begin with an '@' character, which is followed by the sequence identifier. On the second line, the sequence is listed in raw sequence letters. The end of the sequence is indicated by a new line and the third line starts with a '+' character and is optionally followed by the same sequence identifier. The quality values of the sequence follow in the fourth line.

3.5.5. PHRED-SCALED QUALITY VALUES

Phred quality scores are assigned to each base called during an automated sequencing run, such as those produced by NGS sequencing technologies (Cock *et al*., 2010). A phred score is a log value of the probability that a base is called incorrectly by the sequencer, such that:

$$Q = -10 \log_{10} P$$

Where:

Q = phred score

P= error probability

Illumina 1.5 and Sanger phred quality scores are depicted as ASCII (American standard code for informational change) characters with an offset of +33. The system correlates a character with a number. For example, the character "=" represents a phred quality score of 28, that correlates to an error probability of 0.00158. Table 3.5.2 contains some of the quality scores and their corresponding ASCII characters found in the datasets analysed in the present study.

Table 3.5.2. Examples of phred quality scores and the corresponding ASCII characters

Phred quality score (Q)	Error probability (P)	Corresponding ASCII character
02	0.63096	#
10	0.10000	+
15	0.03162	0
20	0.01000	5
25	0.00316	:
28	0.00158	=
30	0.00100	>

3.5.6. NGS DATA ANALYSIS PIPELINE

An extensive selection of specialised software is freely available to analyse high through-put next generation sequencing (NGS) data. However, the software available is often generic and publicly available pipelines do not take organism-specific genomic traits into account. These traits include genome size, GC-content, and characteristic repeat regions. A combination of software packages was used with optimised parameters, together with in-house developed scripts to compile a customised pipeline for the analysis of mycobacterial genomes.

3.5.6.1. QUALITY CONTROL

The raw sequencing data for each isolate analysed was subject to quality control. The quality of the Illumina reads was assessed with FastQC and Prinseq-lite to validate the quality of the reads and to determine whether the data included any factors that would influence the subsequent analysis. FastQC is a java-based program that takes fastq files as input and the results obtained are produced in HTML format. FastQC analyses the raw data through a 7 step module that includes the following:

1. The general statistics, including a description of the platform used, input file type, the amount of reads processed, read length and percentage GC content.
2. Calculating the per read quality score using the mean scores of all the reads.
3. Calculating the per base sequence content to determine the distribution of the four nucleotides throughout the reads.
4. Calculating the GC content throughout the reads and comparing it to a theoretical normal distribution and mean value.
5. Calculating the possibility of contamination in the reads.
6. Calculating the number of uncalled bases throughout the reads (number of “N”s).
7. Calculating the amount of duplicate sequences.

PRINSEQ is a rapid quality control and data pre-processing tool that also produces summarised statistics of fastq files in tabular and graphical format (Schmieder and Edwards, 2011). A locally installed version of PRINSEQ, PRINSEQ-lite, was used to generate graphs presented in Chapter 4 (Section 4.1.). The data generated by PRINSEQ include information on read length, GC content, read quality, sequence complexity, duplicated sequences, the presence of Ns (uncalled bases) and poly-A tails.

3.5.6.2. TRIMMING OF THE READS:

Appropriate trimming of the 3’ ends of the reads was carried out with a simple clipper function in FASTX-Toolkit. While it is important to maintain good quality (phred score > 28) bases in the reads, read length must remain sufficient. A balance should thus be maintained between good quality bases in reads (phred score > 28) and read length. These trimmed, high quality fastq files were used for subsequent analyses. Table 3.5.3 shows the read lengths before and after trimming.

Table 3.5.3. Read length before and after trimming

Read	Read length before trimming (bases)	Number of bases trimmed from 3' end	Resulting length of read (bases)
Forward (R1)	105	35	70
Reverse (R2)	105	35	70

3.5.6.3. ALIGNMENT AND MAPPING

Three independent mapping software packages were used that utilise different algorithms for mapping short sequencing reads to a reference genome, in this case *M.tuberculosis* H37Rv or *M. tuberculosis* F11. The mapping software packages that were used included BWA, which uses a Burrows-Wheeler transform algorithm (Li and Durbin, 2009), Novoalign release 2. 07.18 (Needleman-Wunsch algorithm) (Novocraft Technologies http://www.novocraft.com), and Blat-like Fast Accurate Search Tool (BFAST) which employs a hash table- based algorithm (Homer *et al*., 2009). Various free software packages were used to post-process the alignment files and perform quality control.

Using three different aligners minimised the identification of false positive variants since the aligners use different algorithms. All the aligners employed produced an output in the Sequence Alignment/Map (SAM) format. This format is compact in size and allows for most operations on the alignment to work on a stream without loading the entire alignment into the memory. This format can also be indexed, allowing the efficient and fast retrieval of all reads at a specific chromosomal locus.

BURROWS-WHEELER ALIGNER (BWA)

BWA aligns relatively short reads to a reference genome by implementing two algorithms: BWA-short (for reads shorter than 200bp) and BWA-SW (for longer reads around 100kbp). BWA-short was used for the analyses in the present study since it does gapped global alignment, supports paired-end reads and produces results with a low error rate (< 3%). The program requires an indexed reference genome for the alignment step. This was done by using the "faidx" command from SAMTools (Li *et al*., 2009) and the "index" command in BWA. BWA takes fastq reads as input and uses the "aln" command to find suffix array (SA) coordinates of good hits for every read. The BWA "aln" command was run separately for the forward and reverse reads. Subsequently the "sampe" command was used to convert the SA coordinates into chromosomal coordinates, combining the forward and reverse read alignments, and thereby producing a SAM file. Default command line parameters were used for this alignment procedure.

NOVOALIGN

Novoalign aims to produce fast and accurate alignments of short reads in fastq format to a reference genome in fasta format. This software also requires the reference genome to be indexed. This was done by the "novoindex" command, using a k-mer indexing size of 13 and an indexing step size of 1. Novoalign takes fastq reads as input and uses the Needleman-Wunsch algorithm to compute the alignment. The software does global gapped alignment and for this analysis the default value of 6 was used, thus allowing six mismatches per alignment, producing a SAM file as output.

BFAST

BFAST is optimised to rapidly align DNA sequencing reads with a length of between 25 and 100 bases, to a reference genome. This algorithm creates flexible and efficient whole genome indexes to map reads to candidate alignment locations. Firstly, the reference genome in fasta format was made compatible with BFAST by running the "bfast fasta2brg" command. Ten arbitrary multiple independent indexes were created for each reference genome used (*M. tuberculosis* F11 and *M. tuberculosis* H37Rv), which allows robust alignment to guard against erroneous mapping of low quality reads and sequence variants. The pair of fastq files for each isolate was merged with the use of a python script that interleaves two fastq files by writing a read and its corresponding mate one after the other together with the corresponding quality information in a new fastq file ([CSL STYLE ERROR: reference with no printed form.]). The indexes are searched by using the "bfast match" command. The final local alignment uses a Smith-Waterman method that allows for gapped alignment, and is employed by using the command; "bfast localalign". Finally, the alignment is converted to SAM format with the "bfast postprocess" command. The alignment produced by the series of commands in BFAST does not contain read groups that are necessary for downstream analysis. Therefore the Picard tool "AddReadGroups" (http://picard.sourceforge.net/) were used to add read groups to the SAM file.

3.5.6.4. SAM FILE VALIDATION

The "ValidateSamFile" command from Picard (http://picard.sourceforge.net) was used to report on the validity of the SAM file. This tool verifies the presence of read groups within the SAM file and provides a verification of the primary step in the genomic analysis pipeline.

3.5.6.5. CONVERTING THE SAM FILE TO A BAM FILE

SAMTools (http://samtools.sourceforge.net) provides a set of various utilities for manipulating and post-processing of SAM files (Li *et al*., 2009). The software package includes tools to sort, merge, and index SAM files. The SAMTools commands "view" and "sort" was used to convert the SAM alignment files into a sorted binary alignment (BAM) format.

3.5.6.6. ALIGNMENT STATISTICS

SAMTOOLS

SAMTools was used to compute statistics about the alignment process. The “flagstat” command calculates the total number of reads used as input in the mapping software. Additionally it also calculates: the number of duplicate reads; the number of reads that mapped to the reference genome and the number of reads that properly paired with its mate in the opposite direction when aligned to the reference genome.

QUALIMAP

Qualimap was used to generate more comprehensive alignment statistics for the isolates analysed in Chapter 4, Section 4.1. The program inspects the sequence alignment in an input SAM or BAM file and provides a holistic report of the data with regards, to depth of coverage of the reference genome, nucleotide distributions and mean and median values of the insert size (García-Alcalde *et al*., 2012).

3.5.6.7. POST ALIGNMENT PROCESSING OF BAM FILES

BAM files were subsequently processed to correct for errors incorporated during the alignment step.

COORDINATE SORTING AND INDEXING OF BAM FILES

The SAMTools functions “sort” and “index” were used to convert the BAM file in a format that is easy to manage and read. The BAM file is sorted by coordinate to avoid loading extra alignments into the computational memory.

LOCAL REALIGNMENT AROUND IN/DELS (INSERTIONS AND DELETIONS)

Insertions and deletions might influence the alignment of reads to the reference genome, which may result in many bases mismatching the reference near the misalignment (which could be misinterpreted as single nucleotide polymorphisms (SNPs). To minimize the number of mismatching bases across all reads, the Genome Analysis Toolkit (GATK) was used to locally realign misaligned sequencing reads (DePristo *et al*., 2011; McKenna *et al*., 2010). The first step in this process includes the identification of small intervals which were misaligned, by using the “RealignerTargetCreator” tool. Subsequently, the “IndelRealigner” tool was used to realign the suspicious intervals to the reference genome, and thereby correcting the misaligned reads.

CO-ORDINATE SORTING AND INDEXING OF REALIGNED BAM FILES

The realigned BAM files were sorted with Picard’s “sortsam” function and indexed with the “index” function of the SAMTools software package.

REMOVAL OF PCR DUPLICATES

Duplicate reads may be produced by PCR amplification during library construction. To lessen the bias introduced by PCR amplification, the Picard command “MarkDuplicates” was used to locate duplicate reads in the BAM file, which was then flagged in the output BAM file.

3.5.6.8. VARIANT CALLING

Two independent variant callers were used to identify SNPs and short insertions and deletions (in/dels) with regards to the reference genome used for the alignment. The three different mapping files obtained from the previous step were analysed with two SNP callers – SAMTools and GATK (Li *et al*, 2009; McKenna *et al*, 2010; DePristo *et al*, 2011). The resulting variants are contained in six variant call format (vcf) files. The GATK was also used to identify small insertions and deletions from each of the mappings, resulting in three vcf files containing possible in/dels for each isolate analysed. The use of three mapping algorithms with two variant callers minimised the identification of false positive variants.

GATK

The “UnifiedGenotyper” tool from the GATK was used for SNP and indel calling and produced an output in the variant call format (vcf) format. The Stand_call_conf value was set to 50. This allows a minimum phred-scaled confidence threshold of 50, variants with a confidence value greater than or equal to 50 are reported as polymorphic sites. The Stand_emit_conf value was set to 10. Variants with a phred-scale confidence value greater than or equal to 10 but less than the calling threshold of 50 are reported but marked as filtered. The output vcf file contains information about the position, the alternative sequence and the phred scaled probability that the polymorphism exists at this position. The phred scale (Q) is defined as a property which is logarithmically related to the probability that there might be an error in the actual base-calling. $Q = -10\log_{10}$, a value of 10 indicates that there is a 1 in 10 chance of error. The vcf file also contains information specific to the alternative base, including the number of reads bridging that position, as well as the number of reads that contained the reference and alternative base at that position.

SAMTOOLS

The SAMTools function; “mpileup” was used to create a pileup of the reads relative to the reference genome and subsequently identifies SNPs relative to the reference sequence. Default parameters were employed. In/dels called by SAMTools were excluded from the subsequent analysis. The vcf file produced by SAMTools contains information about the position, the alternative sequence and quality score in phred scale of each variant. The file also contains specific information about the variant, including the number of reads aligning to that position.

3.5.6.9. EXTRACTION OF OVERLAPPING SNPS GENERATED BY THE VARIOUS PIPELINES

The SNPs called by the combination of all the mappers and variant callers that overlap in both position and base identity were written to a high-confidence vcf file with the use of an in-house script, for each reference strain used. The in-house script is called *run_snp_overlap.sh* and was written by Mrs. Michelle Daya. This script makes use of the R statistical package to compare the SNPs identified by each pipeline (R Development Core Team, 2005). These high confidence SNPs were then used in subsequent analyses.

As discussed previously, three different aligners (BWA, Novoalign and BFAST) and two different variant callers (GATK and SAMTools) were used to minimize the identification of false positive variants (see Figure 3.1). Therefore, six strategies (pipelines) where used to identify SNPs, these were BWA-GATK, BWA-SAMTools, Novoalign-GATK, Novoalign-SAMTools, BFAST-GATK and BFAST-SAMTools. For indel calling, only three pipelines were used: BWA-GATK, Novoalign-GATK and BFAST-GATK. Thus variants were only considered in further analysis if identified with all pipelines used. In-house script *run_snp_overlap.sh* was used to extract overlapping SNPs generated by all six pipelines. For this purpose, SNPs present in both vcf files created by BWA-GATK and BWA-SAMTools were first extracted and written to a new vcf file containing all SNPs identified in the BWA-mapping with high confidence. Likewise, SNPs present in both of the vcf files created by Novoalign-GATK and Novoalign-SAMTools were also extracted, and the same procedure was followed for vcf files created by BFAST-GATK and BFAST-SAMTools. SNPs present in all three of these newly created vcf files were extracted to create a final vcf file that contains high confidence SNPs identified by all six pipelines. A similar approach was followed to extract in/dels present in three vcf files created by BWA-GATK, Novoalign-GATK and BFAST-GATK. The statistical package R (http://www.r-project.org) was used to create Venn diagrams to illustrate the distribution of variants created with the respective pipelines.

3.5.6.10. ANNOTATION AND FUNCTIONAL CLASSIFICATION OF SNPS/IN/DELS

Perl scripts written by Mrs Margaretha de Vos were used to: 1) annotate the identified high confidence SNPs, 2) calculate the consequent amino acid change caused by SNPs located within genes (if any), 3) annotate the identified in/dels 4) calculate the effect of in/dels on the reading frame of the corresponding genes. The genes in which the variants occur were classified according to its cellular function as described on the TubercuList knowledge base (Cole *et al*., 1998; Lew *et al*., 2011). The scripts are included in Appendix E.

3.5.6.11. CONSTRUCTION OF A PHYLOGENETIC TREE

A python script (see Appendix E) written by Mrs. Michelle Daya was used to create a concatenated sequence of all high confidence SNPs identified for each isolate. The principle of this method is illustrated by the following example:

Reference strain partial genome sequence:	ATGCAGTTGCGCACAGCTGCGGAT
Strain A partial genome sequence:	ATCCAGTACCGCACCGCTGCGGAT
Strain B partial genome sequence:	ACGCAGTTCCGCACAGGTGCGCTT

Concatenated SNP strings:

Reference:	TGTGACGAT
Strain A:	TCACCCGAT
Strain B:	CGTCAGCTT

The concatenated sequences containing informative sites were saved in multi-fasta format and aligned using ClustalW (Larkin *et al*., 2007). The alignment was saved in the Phylip format (.phy) and was used as input in Modelgenerator to determine the optimal substitution model that fits the data structure (Felsenstein, 1989; Keane *et al*., 2006). The general time reversal (GTR) model scored the lowest in the hierarchical likelihood ration test; Bayesian information criterion (BIC), and thus described the substitution pattern occurring in the dataset most accurately. The GTR model of substitution was subsequently applied to construct a maximum likelihood phylogeny of the isolates included in this analysis with MEGA 5.2, RaxML and PhyML with 1000 bootstrap replicates (Guindon and Gascuel, 2003; Stamatakis, 2006). The phylogenetic trees produced by the different algorithms were compared for similarity of the nodes.

3.5.6.12. IDENTIFICATION OF LARGE DELETIONS

Genome-wide depth of coverage was determined with a specialised function in the BEDTools software package, for each BAM file produced by BWA, Novoalign and BFAST, respectively (Quinlan and Hall, 2010). The depth of coverage files were searched for regions that have zero depth of coverage, i.e. no reads mapped to these regions on the reference genome used during mapping. This step outputs a file with a list of positions that indicate possible deletions. However, visual inspection of every region was required in order to determine if the region indicated is a true deletion, since false deletions are commonly picked up with this approach. Sequencing in highly repetitive regions or regions with a high GC content may present some difficulty and are often erroneous. Mapping stringency may contribute to the fact that these regions are often flagged as deletions in this approach, which is why visual inspection is of the essence. Again, large deletions present in at least two of the three mapping files, were considered as high confidence deletions and were subjected to PCR verification and sequencing to determine the exact loci of the deletion boundaries.

3.5.6.13. VALIDATION OF HIGH CONFIDENCE VARIANTS

High confidence variants (large deletions) identified in the analysis of sequenced genomes were first validated by visualisation using Artemis (Carver *et al*., 2011; Rutherford *et al*., 2000). Subsequently primers were designed to amplify and/or sequence the surrounding regions of the variants as described in Section 3.7.

3.6. POLYMERASE CHAIN REACTION (PCR)

PCR reactions were performed using the HotStar-Taq system (Qiagen, Venlo, Limburg, Netherlands) in 25 µl reaction volumes. The PCR master-mix consisted of 1x Q-solution, 1x reaction buffer, 2mM $MgCl_2$, 200µM of each deoxyribonucleotide triphospate (dNTP), 50 µM forward primer, 50µM internal reverse primer, 50µM reverse primer, 1.25 U Hotstar Taq polymerase, 1 µl of the DNA template (50 ηg/µl), and H_2O to make up the reaction to a final volume of 25 µl. PCR reactions took place in a thermal cycler (GeneAmp PCR System 2400, Applied Biosystems, Foster City, CA, USA) under the following thermo-cycling conditions: an initial denaturing step at 95ºC for 15 minutes, followed by 35 cycles of: (a) a denaturing step at 94ºC for 1 minute, (b) an annealing step at the T_m of the primers included in the specific reaction for 1 minute, (c) an extension step at 72ºC for 1 minute (depending on the expected size of the PCR products, allocating 1 minute per 1000 bp), and a final extension step at 72ºC for 15 minutes. All PCR experiments included a negative control containing no DNA template and a positive control (*M. tuberculosis* H37Rv genomic DNA) which does not contain the deletion being tested for.

3.7. GEL ELECTROPHORESIS

A 1.5% agarose solution was prepared by dissolving 1.5 g of agarose in 100 ml 1x sodium borate (SB) buffer (pH 8.3) and heating the solution in a microwave until fully disolved. Once the mixture has cooled, 5 µl ethidiumbromide (10 mg/ml) was added and the agarose was cast into a gel tray and allowed to cool to room temperature. The PCR product (25 µl) was mixed with 5 µl of blue loading dye (0.25% Xylene Cyanol, 30% glycerol) and a 5 µl aliquot wasloaded onto the gel. A 100 bp Plus DNA ladder (GeneRuler, Thermo Scientific, USA) was loaded into a well adjacent to the samples to allow for the determination of the size of the amplified DNA samples. The gel was run at 120 V for approximately 3-4 hours in 1x SB buffer and visualized under ultra violet light using the Kodak Digital Science Electrophoresis Documentation and Analysis System 120 (VilberLourmat, France).

3.8. PREPARATION OF *MTB.* CULTURES FOR PROTEIN EXTRACTION

Clinical isolates of *M. tuberculosis* that were genetically characterised by IS*6110* DNA fingerprinting and spoligotyping using the internationally standardised methods, were used. The selected isolates were initially cultured in mycobacterial growth indicator tubes (MGITs) at 37°C from frozen stock cultures until positive growth was detected by means of the BACTEC 960 instrument (BD Biosciences, USA).

MGIT tubes were briefly centrifuged to collect bacterial growth at the bottom of the MGIT tube. A dilution series of the bacterial suspension was prepared in Middlebrook 7H9 (Becton, Dickinson and Company, Sparks, USA) liquid medium, supplemented with 0.2% glycerol, 10% albumin, dextrose, catalase (ADC) (Merck Laboratories, Saarchem, Gauteng, SA) and 0.1% Tween 80 (Becton, Dickinson and Company, Sparks, USA). For a ten times dilution, 100 µl of the bacterial suspension was added to 900 µl of supplemented 7H9 liquid medium. From this dilution, 100 µl was added to 900 µl supplemented 7H9 liquid medium in order to obtain a 100 times dilution. For each isolate included in this study, 100 µl of concentrated bacterial suspension from the MGIT culturing tube was spread out on a Middlebrook 7H10 (Becton, Dickinson and Company, Sparks, USA) agar plate, supplemented with 0,5% Glycerol and 10% ADC aseptically. Also, 100 µl of the ten- and 100 times dilution were spread out on supplemented Middlebrook 7H10 plates, respectively, for each isolate. The culturing plates were incubated at 37°C for approximately seven days. The duration of the incubation period varied slightly between strains/isolates, as fitness vary amongst the members of different strains included for investigation in this study. The period of incubation was extended until single colonies were easily visible on the surface of the media, if necessary. Five single colonies were picked for each isolate under investigation. Single colonies were carefully removed from the culturing media with a disposable pipette tip and inoculated in 5 ml of supplemented 7H9 liquid media, in 25 cm^2 culturing flasks (Greiner Bio-one, Maybachstreet, Germany). The culturing flasks were incubated at 37˚C without shaking. Cultures were incubated until an optical density (OD) reading of between 0.6 and 1.0, measured as arbitrary units, was reached at the wavelength of 600 ηm. Multiple glycerol freezer stocks were made by adding 500 µl of the bacterial culture to 500 µl 50% glycerol. These glycerol stock cultures were stored at -80°C until needed for inoculation of cultures for protein extraction.

3.9. ZIEHL-NEELSEN STAINING

In order to ensure that the rest of the experiments proceed successfully, Ziehl-Neelsen (ZN) staining was performed on each isolate, every isolate was also plated on a blood agar plate to confirm the absence of contamination. Blood agar plates with mycobacterial culture were incubated at 37°C for two to three days. If no growth was present on the blood agar plates, and the Ziehl-Neelsen procedure confirmed the absence of non-acid fast bacteria, cultures were considered to be pure and free of contamination. For ZN staining, a small amount of liquid culture from each isolate was fixed individually on a microscope slide. The microscope slides with bacterial culture were heat fixed at 85°C for two hours after which the slides were flooded with carbolfuchsin. Flooded slides were heated with a flame until the liquid on the slides was steaming. Slides were flooded with carbolfuchsin again if necessary and left for 5 minutes. The slides were subsequently rinsed with water and drained of excess liquid. Slides were flooded with 1% acid alcohol and left to destain for approximately 2 minutes. The acid alcohol was removed by rinsing the slides with running water and draining the slides of excess liquid. Slides were flooded with methylene blue and left to stand for roughly 2 minutes, after which the slides were rinsed with running water and drained of excess water. Slides were left to dry completely before analysing the slides with a light microscope under an oil immersion objective lens. The absence of non-acid fast bacteria was confirmed.

3.10. MYCOBACTERIAL CULTURING FOR PROTEIN EXTRACTION

M. tuberculosis cultures selected for proteomic analysis were cultured from the glycerol stocks prepared as described above in Section 3.9. The analysis for each strain selected was performed in duplicate on biological level. For this purpose, two of the glycerol stock cultures (prepared from one single colony inoculated in liquid culture) were used to obtain two separate starter cultures, from each of which a final volume of 50 ml was inoculated for whole cell lysate protein extraction. Each 1 ml glycerol stock culture was inoculated into 10 ml 7H9 Middlebrook growth medium supplemented with 0.2% glycerol, 0.1% Tween 80 and 10% dextrose, catalase and incubated at 37°C until it reached an A_{600} (absorbance or optical density reading at 600 ηm) of between 0.6 and 0.8. (Note that no albumin was included in the growth media used to culture isolates subjected to proteomic investigation as the added protein may have an influence on the protein fingerprint obtained for each isolate via mass spectrometry.) One millilitre of the starter culture was inoculated into 50 ml albumin-free supplemented 7H9 Middlebrook medium and subsequently incubated at 37°C until the culture reached an OD of between 0.6 and 0.7 at 600 ηm. A growth curve experiment was performed for all strains cultured for whole cell lysate protein extraction by measuring the OD at regular time intervals for a period of 20 days. It was determined that the *M. tuberculosis* cultures used in the present study reached a mid-log (middle logarithmic) growth phase at an OD (A_{600}) of between 0.6 and 0.7. Care was taken to ensure that all cultured isolates used for whole cell lysate protein extractions and subsequent protein analysis were at a similar OD at the time of harvesting to ensure comparability between different strains evaluated by mass spectrometry.

3.11. PREPARATION OF CRUDE MYCOBACTERIAL EXTRACTS

The mycobacterial cells were cultured as described in Section 3.3 and harvested by means of centrifugation for 10 minutes, at 2500 g and 4°C. The supernatant was decanted into sterile 50 ml tubes and the cell pellet was resuspended in 1 ml of cold lysis buffer containing 10 mM Tris-HCl , pH 7.4 (Merck Laboratories, New Jersey, USA), 0.1% Tween 80 (Sigma-Aldrich, Missouri, USA), 200 µl of protease inhibitor cocktail (Roche Applied Bioscience, Penzburg, Germany). Resuspended cells were transferred into 2 ml cryogenic tubes with O-rings. The resuspension was centrifuged at 13 000 g for 1 minute, followed by one minute on ice, and the process was repeated. The supernatant was removed by pipetting and discarded. An equal volume of 0.1 mm glass beads (Biospec Products Inc, Bartlesville, OK) was added to the pellet of cells together with 300 µl of cold lysis buffer and 10 µl of RNase-free DNase I (Thermo Scientific, Massachusetts, USA). The cells were subsequently lysed mechanically by ribolysing 6 times for 20 seconds at a speed of 4.0, after each ribolysing step the cells were cooled on ice for 1 minute. The crude lysate was centrifuged at 13 000 g for one minute, kept on ice for one minute, and centrifugation was repeated. The supernatant was retained and filter-sterilised twice through a 0.22 micron Acrodisc 25mm PF syringe filter (Millipore, Massachusetts, USA), aliquoted in 100 µl aliquots and stored at -80˚C until further analysis. The culture media was filter sterilised in the same manner and stored at -80˚C for future analysis.

3.12. PROTEIN FRACTIONATION AND IN-GEL TRYPSIN DIGEST

The protein concentration was determined by using the commercial RC-DC Protein assay (Bio-Rad Laboratories, California, USA), following the manufacturer's instructions. A total of 60 µg of whole cell lysate protein extract made up to a total volume of 20 µl was mixed with 6 × Laemmli buffer and heated for 5 minutes at 95 ˚C. The total volume of protein extract and sample buffer was loaded on a 1.0 mm NuPAGE gel (Invitrogen, Carlsbad, CA, USA) with a gradient of 4 – 12%. For each biological replicate, the proteins were fractionated in duplicate by running the SDS-PAGE at 150 V for approximately 1 hour, or until the dye has reached the bottom of the gel. After electrophoresis, the gel was stained for one hour with Imperial™ Protein Stain (Thermo Scientific, Massachusetts, USA), whilst shaking. The gel was destained in double distilled water overnight to remove excess stain and the gel was inspected to make sure that the lanes were uniform and that proteins were not degraded in the previous steps.

Each lane of the gel was divided into 10 approximately equal fractions and each fraction was cut into $1mm^2$ sized cubes and added to a sterile 1.5 ml centrifuge tube. Fractions were washed with 300 µl HPLC grade water for 10 minutes at room temperature, after which the water was removed by pipetting. A volume of 300 µl 50% acetonitrile (ACN) (Sigma-Aldrich, Misourri, USA) was added to the fractions for 10 minutes at room temperature, after which the ACN was removed by pipetting. This was followed by 300 µl of 50 mM ammonium bicarbonate (ABC) (Sigma-Aldrich, Misourri, USA) and a further incubation step at room temperature for 10 minutes, after which the ABC was also removed. The wash cycle was repeated 3 times to remove excess dye from the gel pieces. Three hundred µl of 100% ACN was added to each fraction and was removed after 10 minutes. The fractions were dried in a centrifugal evaporator for approximately one hour. After the fractions were dried, 120 µl of 10 mM dithiothreitol (DTT) (Sigma-Aldrich, Misourri, USA) was added to each fraction to reduce the disulphide bonds in the proteins. The fractions were incubated at 56 ˚C and the DTT was removed after 1 hour, after which 120 µl of 55 mM iodacetamide (IAA) (Sigma-Aldrich, Misourri, USA) was added to each fraction and incubated in the dark at room temperature for 1 hour. IAA alkylates catalytic cysteine residues. The IAA was removed when the incubation time was over and the fractions were washed by adding 300 µl of 55 mM ABC and incubating the fractions at room temperature for 10 minutes. The ABC was removed and 300 µl of 50% ACN was added to each fraction. After 10 minutes of incubation at room temperature, the ACN was removed and the fractions were dried in a centrifugal evaporator for approximately 1 hour to remove the ACN. The reduced and alkylated proteins were digested with sequencing grade modified trypsin (Promega, Fitchburg, Wisconsin) by adding 80 µl of 10 ng/µl trypsin solution to each fraction, and incubating the mixture at 37˚C for 17 hours. To quench the reaction, 80 µl 70% ACN, 0.1% formic acid (FA) (Sigma-Aldrich, Misourri, USA) was added to each fraction. After 30 minutes, the supernatant of each fraction was removed and added to a sterile 1.5 ml centrifuge tube. The peptide mixtures were desalted using STAGE-tips packed with C18 resin and dried in a centrifugal evaporator for one hour.

3.13. MASS SPECTROMETRY

The dried peptides obtained from the previous step were submitted to the Stellenbosch University Central Analytical Facility (CAF) for analysis by means of Mass Spectrometry by Dr. S. Smit. Dried peptides were dissolved in 5% acetonitrile in 0.1% formic acid and 10 µl injections were made for nano-LC chromatography. All experiments were done on a Thermo Scientific EASY-nLC II connected to a LTQ Orbitrap Velos mass spectrometer (Thermo Scientific, Germany) equipped with a nanoelectropsray source. For liquid chromatography, separation was performed on an EASY-Column (2 cm, ID 100µm, 5 µm, C18) pre-column followed by XBridge BEH130 NanoEase column (15 cm, ID 75 µm, 3.5 µm, C18) with a flow rate of 300 ηl/min. The gradient used was from 5-17% B in 5 min, 17-25% B in 90 min, 25-60% B in 10 min, and 60-80% B in 5 min and kept at 80% B for 10 min. Solvent A was 100% water in 0.1% formic acid, and solvent B was 100% acetonitrile in 0.1% formic acid. The mass spectrometer was operated in data-dependent mode to automatically switch between Orbitrap-MS and LTQ-MS/MS acquisition. Data were acquired using the Xcaliber software package (Thermo Scientific, Germany). The precursor ion scan MS spectra (*m/z* 400 – 2000) were acquired in the Orbitrap with resolution R = 60000 with the number of accumulated ions being 1 x 10^6. The 20 most intense ions were isolated and fragmented in linear ion trap (number of accumulated ions 1.5 x 10^4) using collision induced dissociation. The lock mass option (polydimethylcyclosiloxane; *m/z* 445.120025) enabled accurate mass measurement in both the MS and MS/MS modes. In data-dependent LC-MS/MS experiments, dynamic exclusion was used with 60 s exclusion duration. Mass spectrometry conditions were 1.8 kV, capillary temperature of 250°C, with no sheath and auxiliary gas flow. The ion selection threshold was 500 counts for MS/MS and an activation Q-value of 0.25 and activation time of 10 ms were also applied for MS/MS.

3.14. PROTEIN IDENTIFICATION

All acquired data was processed and analysed using MaxQuant 1.2.2.5 (Cox and Mann, 2008) as part of the service provided by CAF. MaxQuant is a united suite of algorithms that were specifically developed for the analysis of high-resolution, quantitative mass spectrometry data. The algorithms employ correlation analysis and graph theory to detect peaks, isotope clusters and labelled peptide pairs as three-dimensional objects in mass to charge ratio (m/z). It uses elution time and signal intensity space (area) to determine peptide sequences. The statistical plugin, Perseus, is used to do all downstream statistical analysis on the MaxQuant data files and peak intensity is used to infer relative abundance. The elution profiles of the peaks detected in the first mass spectrometry scan (MS1) is used to determine the area under the curve as a measure of abundance in order to compare peaks in samples that are being compared (Cox and Mann, 2008).

A total of 120 RAW files which accounted for a total of 240 hours of acquisition time were submitted for protein identification using the TbDB H37Rv (GB:AL123456) and custom database as previously reported (de Souza *et al*., 2011, 2010). Carbamidomethyl cysteine was set as fixed modification, and oxidized methionine, N-terminal acetylation (protein), deamidation (NQ), pyro-Glu(Gln) and pyro-Glu(Glu) as variable modifications. The precursor mass tolerance was set to 20 ppm, and fragment mass tolerance set to 0.8 Da. Two missed tryptic cleavages were allowed. The following criteria were applied: peptide and protein false discovery rate of 1% (0.01), minimal peptide length of six, and a protein posterior error probability (PEP) of 0.01 and peptide PEP of 0.15. These extremely strict parameters guarantee that proteins are identified with high confidence. Each protein identified was assigned to a functional category of proteins as defined by Cole *et al*. (1998) and summarised on the TubercuList knowledge base (Lew *et al*., 2011).

CHAPTER 4

RESULTS & DISCUSSION

4.1. QUALITY CONTROL

4.1.1. INTRODUCTION

Whole genome sequencing of clinical isolates from the Western Cape region in South Africa was done in collaboration with the King Abdullah University of Science and Technology (KAUST), and the sequencing process and data analysis pipeline are explained in detail in Chapter 3 (Section 3.8). Several measures were taken to ensure high quality sequencing and mapping of data in order to call variants with high confidence. In order to control for the quality of the sequencing technology and alignment strategies employed to analyse the large amounts of data generated during the present study, two clinical isolates from the local setting (SAWC 4498 and SAWC 5165) were used as controls to validate the NGS data analysis pipeline. These isolates were identified as fully drug susceptible *M. tuberculosis* F11 LAM strains by standard genotyping methods, and were selected based on the fact that they are members of the same *M. tuberculosis* strain family as the fully sequenced *M. tuberculosis* F11 strain. These isolates were subjected to whole genome sequencing on the Illumina HiSeq 2000 platform and the customised pipeline was used to analyse the WGS data. The reads obtained for these strains were aligned to the completed sequence of *M. tuberculosis* F11 (Section 4.1.1), sequenced by the Broad Institute (*M. tuberculosis* Diversity Sequencing Project, Broad Institute of Harvard and MIT (http://www.broadinstitute.org/)), as well as the completed genome sequence of the *M. tuberculosis* H37Rv laboratory strain (Section 4.1.2). The *M. tuberculosis* F11 strain was the first drug resistant strain to be sequenced and the second clinical isolate of *M. tuberculosis* of which the completed sequence was produced (Victor, 2005). The F11 strain has been isolated from TB patients around the world and is particularly prevalent in the Western Cape Province, South Africa. The genome sequence of the F11 strain revealed that it had multi-drug resistance conferring mutations in *katG*315 (AGC to ACC) and *rpoB*531 (TCG to TTG).

The *M. tuberculosis* F11 strain family lack spoligotype spacers 9 to 11, 21 to 24, and 33 to 36 (Warren *et al*., 1999), which is consistent with the spoligotype patterns of SAWC 4498 and SAWC 5165 (Table 3.1). F11 isolates are members of PGG2 based on the SNPs in the *katG*463 (CGG) and *gyrA*95 (ACC) genes. The evolution of the PGG-defining SNPs in *katG* and *gyrA* is shown in Figure 4.1.1.

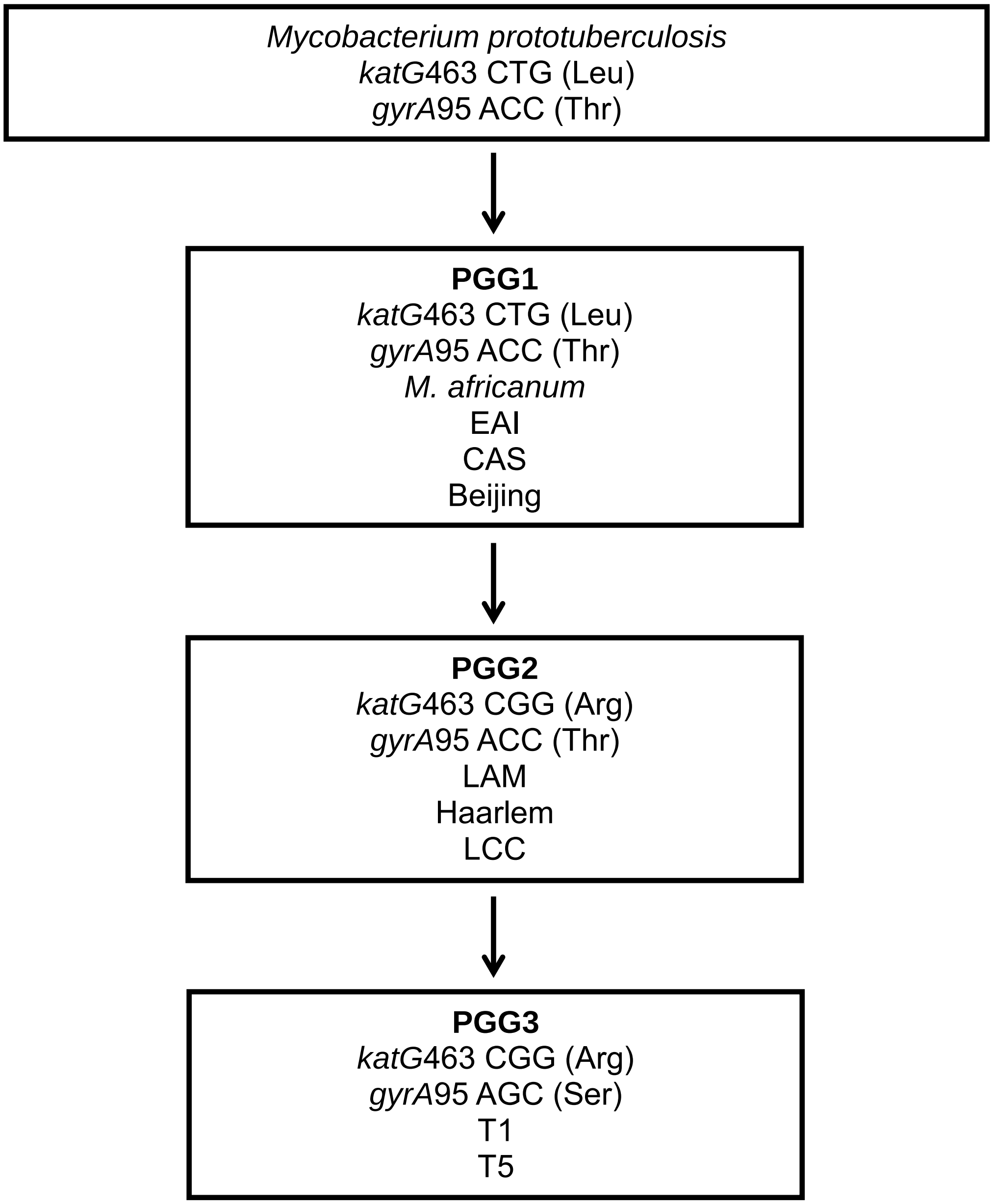


Figure 4.1.1. PGG-defining SNPs in *gyrA* and *katG*.

All members of F11 have a C – T polymorphism in *rrs*491 (Victor *et al*., 2001). From the published sequence of the *M. tuberculosis* F11 sequence, a 1094 bp deletion with respect to *M. tuberculosis* H37Rv was revealed. This region contains the genes *Rv1334* – *Rv1336*, and is known as RD761 (Gagneux *et al*., 2006; Victor *et al*., 2004). Other LAM specific, and F11 genetic markers with respect to the reference genome, *M. tuberculosis* H37Rv, are listed in Table 2.1.4, Chapter 2. All of these genetic markers characteristic to *M. tuberculosis* F11 were confirmed in the two *M. tuberculosis* F11 isolates (SAWC 4498 and SAWC 5165) analysed in this chapter of the present study.

4.1.2 READ ASSESSMENT AND TRIMMING

The quality of the raw sequencing reads in the forward and the reverse orientation were assessed for both of the *M. tuberculosis* F11 LAM strains using open-access quality control tools. FastQC and PRINSEQ-lite, were used to assess the quality of the reads and the consequent mapping steps ("Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data," n.d.; Schmieder and Edwards, 2011). The quality of the raw sequencing data was considered and trimming of the reads were done accordingly. Figure 4.1.2.A and 4.1.2.B shows quality scores for the raw reads obtained for SAWC 4498 in the forward and the reverse orientation, respectively. The reads were 105 bases long and the quality scores in phred scale deteriorated towards the ends of the reads, consistent with previous observations (Claesson *et al*., 2010). The average and mean values of the quality scores at the base position in the read indicated that 35 bases should be trimmed from the 5`-ends of the sequencing reads. Trimming resulted in reads that were 70 bases long with average phred scaled quality values above 28, which translates to a sequencing error probability of 0.00158. The quality of the trimmed reads in the forward and reverse orientation is depicted in Figure 4.1.2.C and 4.1.2.D, respectively.

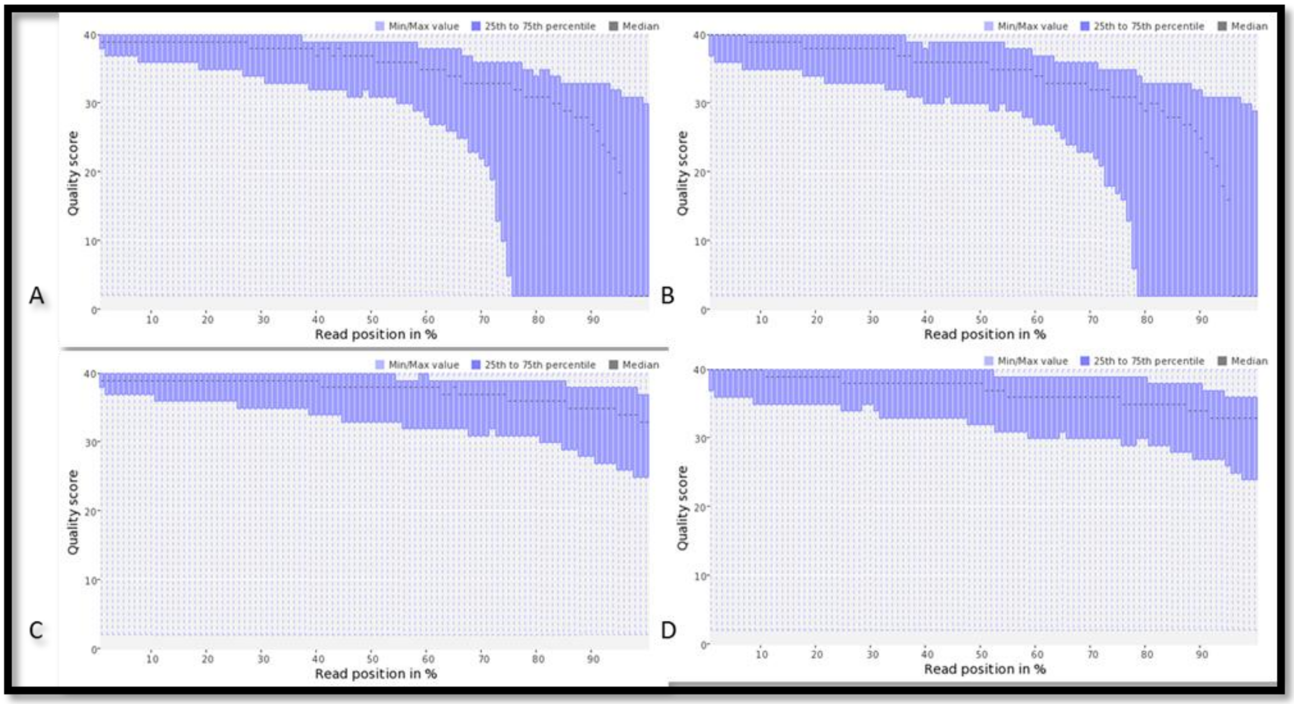


Figure 4.1.2. The per base quality of the sequencing reads for SAWC 4498.
A. The per base quality graph for the forward reads (R1) of SAWC 4498 before trimming.
B. The per base quality graph for the reverse reads (R2) of SAWC 4498 before trimming.
C. The per base quality graph for the forward reads (R1) of SAWC 4498 after trimming.
D. The per base quality graph for the reverse reads (R2) of SAWC 4498 after trimming.

Similarly, Figure 4.1.3.A and 4.1.3.B illustrates the quality of the raw sequencing reads respectively in the forward and reverse direction obtained for SAWC 5165. The average and mean values of the quality scores at the base position in the read were used to indicate that 35 bases should be trimmed from the 5`-ends of the sequencing reads. The trimming procedure resulted in reads that were 70 bases long with average phred scaled quality values above 28. The quality of the trimmed reads in the forward and reverse orientation is shown in Figure 4.1.2.C and 4.1.2.D respectively.

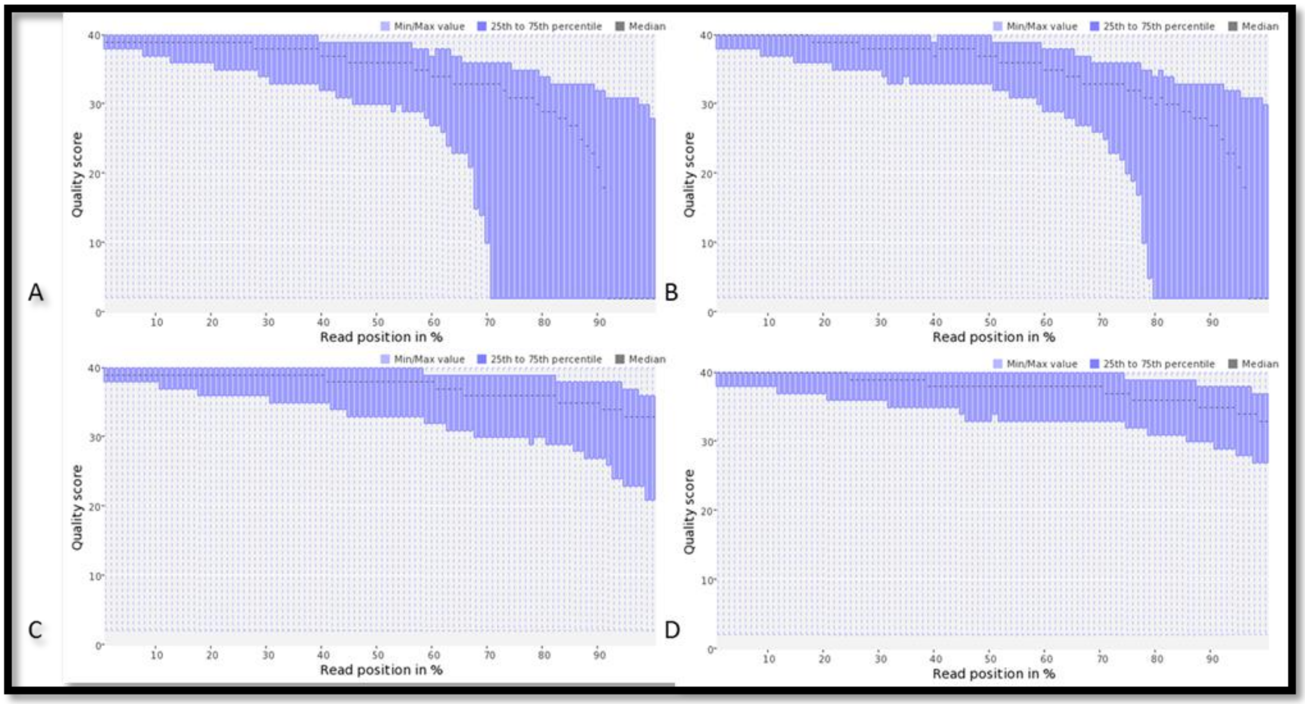


Figure 4.1.3. The per base quality of the sequencing reads for SAWC 5165.
A. The per base quality graph for the forward reads (R1) of SAWC 5165 before trimming.
B. The per base quality graph for the reverse reads (R2) of SAWC 5165 before trimming.
C. The per base quality graph for the forward reads (R1) of SAWC 5165 after trimming.
D. The per base quality graph for the reverse reads (R2) of SAWC 5165 after trimming.

The pre-processing and quality control of raw sequencing data is the first crucial *in silico* step to ensure that high quality data is used as input in the next generation sequencing (NGS) analysis pipeline. Table 4.1.1 summarises the number of sequencing reads obtained for SAWC 4498 and SAWC 5165, the read lengths before and after trimming, as well as the GC content of the reads. The number of reads can provide an approximate indication of the coverage of the genome of interest by multiplying the total number of reads (in the forward and reverse orientation) with the original read length to obtain the total number of bases generated by the sequencing run for a specific isolate. The expected size of the genome of interest divided by the total number of bases generated gives an estimated indication of the average coverage of the entire genome. This estimate can be used as a preliminary screen to evaluate whether enough data has been generated; however more comprehensive and reliable coverage statistics can be obtained after the pre-processed reads have been mapped to a reference genome. When members of a species with a high GC content is sequenced GC content of the reads should reflect the expected genomic GC content, as PCR- and sequencing errors can occur and may lead to sequencing bias where GC-rich areas of the genome of interest are not covered sufficiently.

Table 4.1.1. Summary of sequencing reads for SAWC 4498 and SAWC 5165

		SAWC 4498		**SAWC 5165**	
	Read orientation	Forward	Reverse	Forward	Reverse
Before trimming	Number of reads	2 976 575	2 976 575	4 000 000	4 000 000
	Read length	105	105	105	105
	GC content	64%	65%	65%	65%
After trimming	Number of reads	2 976 575	2 976 575	4 000 000	4 000 000
	Read length	70	70	70	70
	GC content	64%	64%	64%	64%

4.1.3. READ ALIGNMENT AND MAPPING STATISTICS

Three autonomous mapping software packages were used that employ different algorithms for mapping short sequencing reads to the reference genomes used for the analysis in this chapter, i.e. *M. tuberculosis* F11 and *M. tuberculosis* H37Rv, respectively (http://genome.tbdb.org/annotation/genome/tbdb/MultiDownloads.html). The mapping software packages that were used included BWA that uses a Burrows-Wheeler transform algorithm (Li and Durbin, 2009), Novoalign release 2.07.18 (Needleman-Wunsch algorithm) (Novocraft Technologies http://www.novocraft.com), and Blat-like Fast Accurate Search Tool (BFAST) which employs a hash table-based algorithm (Homer *et al*., 2009), as described in Chapter 3, Section 3.8.6.

All the aligners used for mapping sequencing reads to a reference genome produced an output in the Sequence Alignment/Map (SAM) format. Qualimap, a Java application that considers sequence features and their genomic properties to provide graphical and statistical evaluation of the data in SAM or BAM format, was used to obtain mapping statistics for the *M. tuberculosis* F11 strains pertaining to the present analysis (García-Alcalde *et al*., 2012). Selected mapping statistics obtained with Qualimap for the alignments of SAWC 4498 and SAWC 5165, to the reference genomes *M. tuberculosis* F11 and *M. tuberculosis* H37Rv, are summarised in Table 4.1.2. The range of values obtained for each of the parameters listed in Table 4.1.2 reflects the divergent algorithms employed by the mapping software packages used. The repetitive nature of sections of the *M. tuberculosis* genome provides challenges for whole genome sequencing and the bioinformatics analysis of the data generated. Aligning algorithms may map a read to the wrong location in the reference genome, especially around repetitive areas of other low-complexity regions. Using multiple aligners thus minimises the identification of false positive variants called for each isolate, since different algorithms were used to map sequencing reads to the reference genome.

Table 4.1.2. Summary of alignment statistics for SAWC 4498 and SAWC 5165

		SAWC 4498			**SAWC 5165**		
		BFAST	**BWA**	**Novoalign**	**BFAST**	**BWA**	**Novoalign**
***M. tuberculosis* F11**	Percentage mapped reads	96.05%	95.85%	97.72%	96.51%	96.18%	98.22%
	Mean coverage	89.01 (+/- 23.87)	89.26 (+/- 21.07)	79.3 (+/- 21.98)	124.55 (+/- 34.19)	124.77 (+/- 30.44)	110.51 (+/- 31.67)
	Mean mapping quality	47.52	56.74	69.61	48.17	57.16	69.68
	Mean insert size	-	252.07	252.08	-	220.35	220.41
	Number of mapped bases	393817300	394935034	350853342	551051003	552049399	488927395
		BFAST	**BWA**	**Novoalign**	**BFAST**	**BWA**	**Novoalign**
***M. tuberculosis* H37Rv**	Percentage mapped reads	95.79%	95.37%	97.23%	96.25%	95.72%	97.77%
	Mean coverage	88.97 (+/- 24.2)	89.09 (+/- 21.74)	79.09 (+/- 22.46)	124.48 (+/- 34.59)	124.47 (+/- 31.29)	110.25 (+/- 32.25)
	Mean mapping quality	47.47	56.72	69.59	48.12	57.14	69.66
	Mean insert size	-	252.01	252.03	-	220.32	220.37
	Number of mapped bases	392515657	392754834	348621137	549169463	549108114	486392010

4.1.4. HIGH CONFIDENCE VARIANTS

4.1.4.1. *M. TUBERCULOSIS* F11 AS REFERENCE GENOME

Three different aligners (BWA, Novoalign and BFAST) and two independent variant callers (GATK and SAMTools) were used to minimize the identification of false positive SNPs. Therefore, six pipelines where used to identify SNPs, these were BWA-GATK, BWA-SAMTools, Novoalign-GATK, Novoalign-SAMTools, BFAST-GATK and BFAST-SAMTools. For in/del calling, only three pipelines were used: BWA-GATK, Novoalign-GATK and BFAST-GATK. SNPs present in both vcf files created by BWA-GATK and BWA-SAMTools were first extracted and written to a new vcf file containing all SNPs identified in the BWA-mapping with high confidence. Likewise, SNPs present in both of the vcf files created by Novoalign-GATK and Novoalign-SAMTools were also extracted, and the same procedure was followed for vcf files created by BFAST-GATK and BFAST-SAMTools. SNPs present in all three of these newly created vcf files were extracted to create a final vcf file that contains high confidence SNPs identified by all six pipelines. A similar approach was followed to extract in/dels present in three vcf files created by BWA-GATK, Novoalign-GATK and BFAST-GATK. The statistical package R (http://www.r-project.org) was used to create Venn diagrams to illustrate the distribution of variants identified with the respective pipelines. The Venn diagrams shown in Figure 4.1.4.A and -B depicts the number of SNPs that are identical in position and base identity and which overlap between the pipelines for SAWC 4498 and SAWC 5165 respectively. A total number of 79 high confidence SNPs were identified in SAWC 4498 compared to the *M. tuberculosis* F11 reference genome (Figure 4.1.4.A). The SNPs identified by only one or two of the three pipelines were not regarded as true variants. Upon further inspection it was shown that mapping quality at the positions of these SNPs only identified by one or two of the pipelines was low. The *M. tuberculosis* SAWC 5165 isolate differed from the *M. tuberculosis* F11 reference genome by 65 high confidence SNPs (Figure 4.1.4.B).

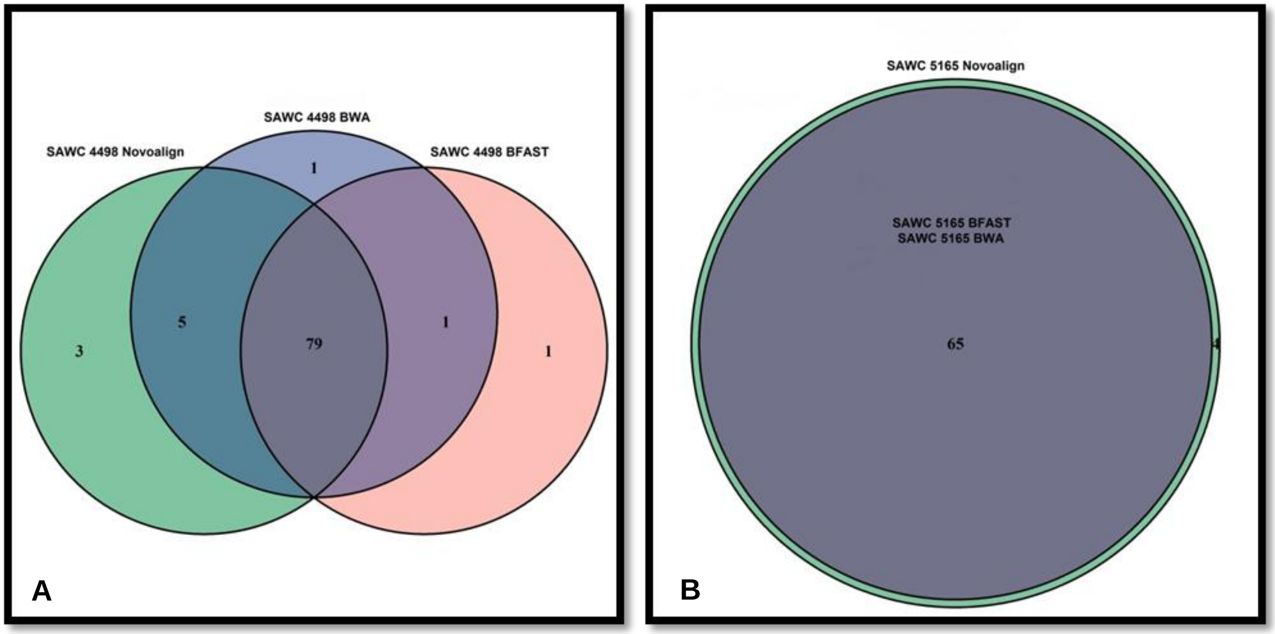


Figure 4.1.4. SNPs in SAWC 4498 and SAWC 5165. A. SNPs called with SAMTools and GATK using the alignments obtained from different mapping algorithms for SAWC 4498, employing *M. tuberculosis* F11 as a reference genome. B. SNPs called with SAMTools and GATK using the alignments obtained from different mapping algorithms for SAWC 5165, employing *M. tuberculosis* F11 as a reference genome.

The Venn diagrams shown in Figure 4.1.5.A and -B depicts the number of high confidence in/dels that overlap, and are unique to SAWC 4498 and SAWC 5165, respectively. A total number of 14 high confidence in/dels were identified in SAWC 4498 compared to the *M. tuberculosis* F11 reference genome (Figure 4.1.5.A). The in/dels identified by only one or two of the three pipelines were not regarded as true variants, upon further inspection it was shown that the alignment algorithms often did not agree on the start and end positions of the in/dels, despite realignment around in/dels being done.

Mapping quality around the positions of these in/dels only identified by one or two of the pipelines was low. The *M. tuberculosis* SAWC 5165 isolate differed from the *M. tuberculosis* F11 reference genome by 10 high confidence in/dels (Figure 4.1.5.B).

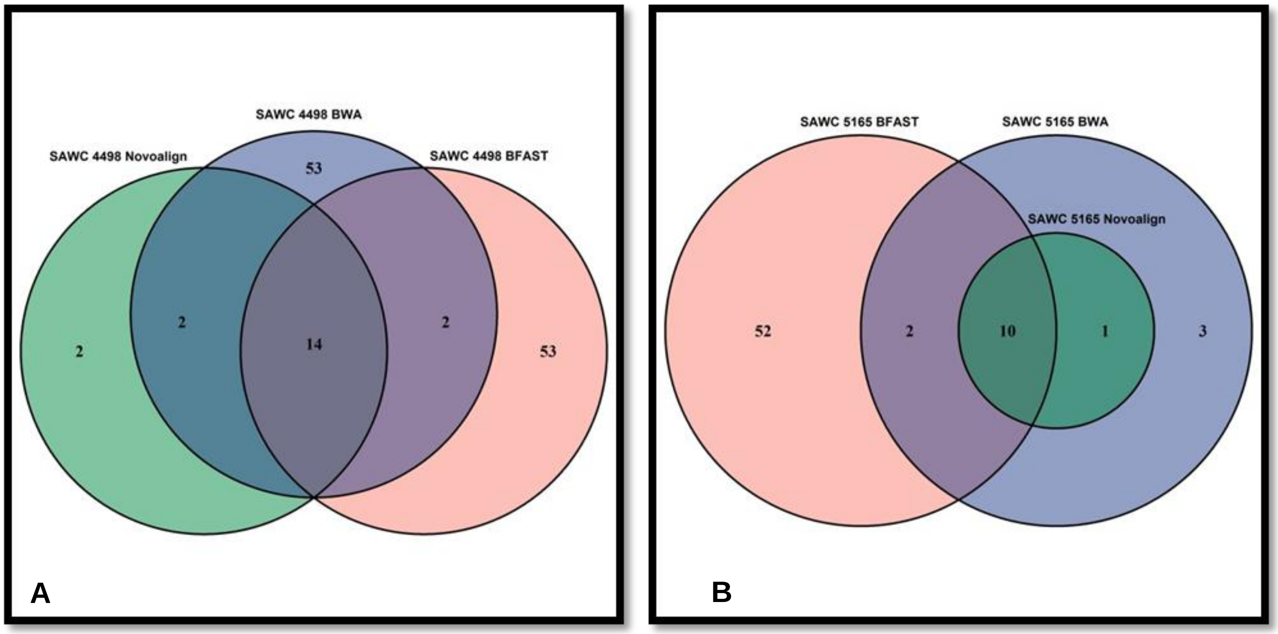


Figure 4.1.5. In/dels in SAWC 4498 and SAWC 5165. A. In/dels called with GATK using the alignments obtained from different mapping algorithms for SAWC 4498, employing *M. tuberculosis* F11 as a reference genome. B. In/dels called with GATK using the alignments obtained from different mapping algorithms for SAWC 5165, employing *M. tuberculosis* F11 as a reference genome.

The *M. tuberculosis* clinical isolates, SAWC 4498 and SAWC 5165, which were analysed for the purpose of this chapter were both confirmed to be members of the LAM family of strains, more specifically the F11 subfamily, as determined by their spoligotype patterns and characteristic *IS6110* RFLP fingerprints, and thus belong to the same family as the reference strain, *M. tuberculosis* F11. Even though these isolates are members of the same family as the reference strain and seemingly closely related to one another, a relatively large number of variants were identified between the isolates and the reference strain used in the mapping steps of the bioinformatics analysis. The diversity of the strains highlights the diversity of the pool of strains present in the high TB incidence setting from which these strains were isolated. Also, the extent to which whole genome sequencing provides an unparalleled means to detect genetic diversity between isolates is emphasised. Figure 4.1.6.A shows a Venn diagram of the SNPs that overlap and are unique to SAWC 4498 and SAWC 5165 with regards to the *M. tuberculosis* F11 reference strain. SAWC 4498 and SAWC 5165 share 33 common SNPs and have 46 and 32 SNPs respectively that are unique to the two strains (in position and base identity). SAWC 4498 and SAWC 5165 share six in/dels with respect to *M. tuberculosis* F11 and have eight and four unique in/dels, respectively (Figure 4.1.6.B).

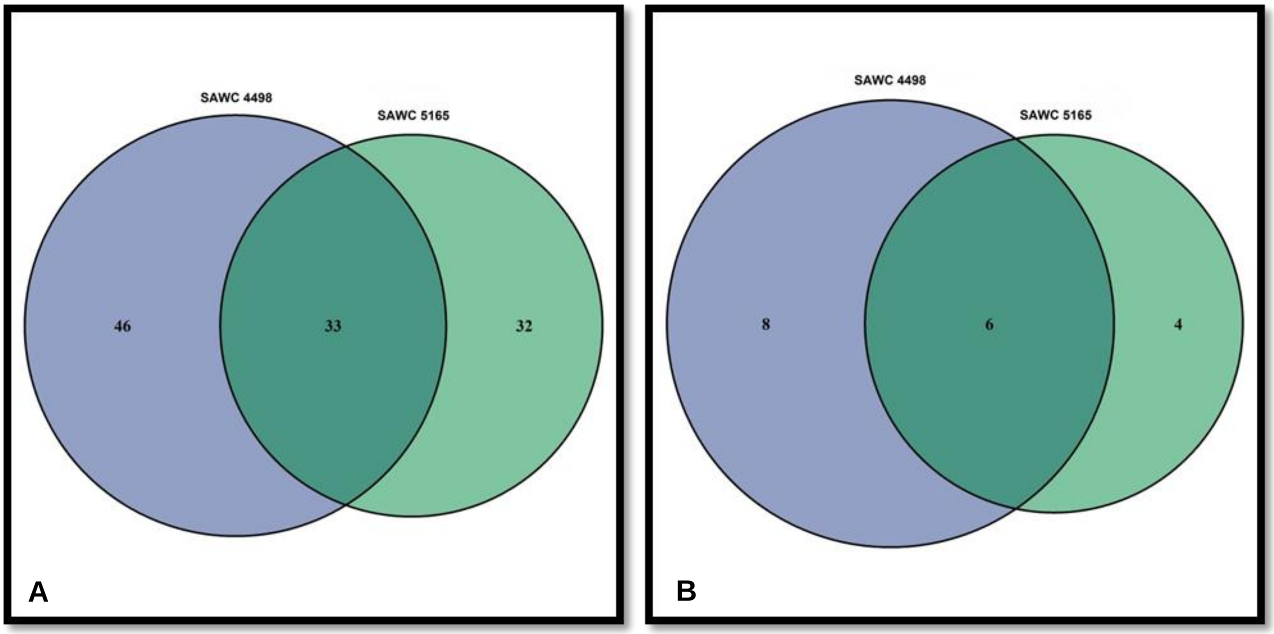


Figure 4.1.6. Overlapping variants between SAWC 4498 and SAWC 5165. A. SNPs that overlap in position and base identity between, or unique to, SAWC 4498 and SAWC 5165, employing *M. tuberculosis* F11 as a reference genome. B. In/dels that overlap in length, start- and end position and base identity between, or unique to, SAWC 4498 and SAWC 5165, employing *M. tuberculosis* F11 as a reference genome.

SNPs unique to SAWC 4498 and SAWC 5165 respectively, with regards to the *M. tuberculosis* F11 reference strain, and overlapping between these two isolates, were annotated. Annotation includes determining whether the SNP falls in between two coding regions (i.e. is situated in an intergenic region) or if it falls within a gene (i.e. it is intragenic). SNPs that occur intragenically were further assessed for being either non-synonymous (resulting in an amino acid change in the translated protein product) or synonymous (not resulting in an amino acid change in the translated protein product). Furthermore, the genes in which the SNPs occur were assigned to functional categories according to the TubercuList knowledge base as first described by Cole *et al*., 1998 (Lew *et al*., 2011). Table 4.1.3, 4.1.4 and 4.1.5 summarises the annotated SNPs unique to, and shared between SAWC 4498 and SAWC 5165. The in/dels unique to, and shared between SAWC 4498 and SAWC 5165 are summarised in Table 1, Appendix C.

Of the 46 high confidence unique SNPs identified in SAWC 4498 (not identified in SAWC 5165), five were intergenic. Of the remaining 41 SNPs, 27 led to non-synonymous changes whilst 14 were synonymous. A large proportion (roughly one third) of the genes of *M. tuberculosis* has an unknown function, as can be seen by the number of SNPs found in genes that are annotated as “conserved hypothetical” (12) or “hypothetical” (4), see Table 4.1.3.

Table 4.1.3 Analysis of SNPs unique to SAWC 4498

	All SNPs		**Non-synonymous SNPs**		**Synonymous SNPs**	
	Functional category of genes in which SNPs occur	**Number of SNPs**	**Functional category of genes in which SNPs occur**	**Number of SNPs**	**Functional category of genes in which SNPs occur**	**Number of SNPs**
SAWC 4498 only	Information pathways	4	Information pathways	2	Information pathways	2
	Conserved hypothetical proteins	12	Conserved hypothetical proteins	8	Conserved hypothetical proteins	4
	Cell wall and cell processes	4	Cell wall and cell processes	3	Cell wall and cell processes	1
	Intermediary metabolism and respiration	5	Intermediary metabolism and respiration	3	Intermediary metabolism and respiration	2
	Regulatory proteins	0	Regulatory proteins	0	Regulatory proteins	0
	Virulence, detoxification and adaptation	3	Virulence, detoxification and adaptation	1	Virulence, detoxification and adaptation	2
	Lipid metabolism	5	Lipid metabolism	4	Lipid metabolism	1
	PE/PPE protein families	1	PE/PPE protein families	0	PE/PPE protein families	1
	Insertion sequences and phages	3	Insertion sequences and phages	3	Insertion sequences and phages	0
	Hypothetical	4	Hypothetical	3	Hypothetical	1
	Intergenic	5	-		-	
	Total	46	Total	27	Total	14

Table 4.1.4 shows that one SNP of the total number of 32 high confidence SNPs that were identified uniquely for SAWC 5165 (not found in SAWC 4498) with respect to the *M. tuberculosis* F11 reference strain was intergenic. Of the remaining 31 SNPs, 20 led to non-synonymous changes and 11 were synonymous. A total number of 11 and 5 SNPs were annotated as “conserved hypothetical” and “hypothetical”, respectively.

Table 4.1.4 Analysis of SNPs unique to SAWC 5165

	All SNPs		Non-synonymous SNPs		Synonymous SNPs	
SAWC 5165 only	**Functional category of genes in which SNPs occur**	**Number of SNPs**	**Functional category of genes in which SNPs occur**	**Number of SNPs**	**Functional category of genes in which SNPs occur**	**Number of SNPs**
	Information pathways	2	Information pathways	2	Information pathways	0
	Conserved hypothetical proteins	11	Conserved hypothetical proteins	8	Conserved hypothetical proteins	3
	Cell wall and cell processes	0	Cell wall and cell processes	0	Cell wall and cell processes	0
	Intermediary metabolism and respiration	6	Intermediary metabolism and respiration	3	Intermediary metabolism and respiration	3
	Regulatory proteins	0	Regulatory proteins	0	Regulatory proteins	0
	Virulence, detoxification and adaptation	0	Virulence, detoxification and adaptation	0	Virulence, detoxification and adaptation	0
	Lipid metabolism	4	Lipid metabolism	2	Lipid metabolism	2
	PE/PPE protein families	0	PE/PPE protein families	0	PE/PPE protein families	0
	Insertion sequences and phages	3	Insertion sequences and phages	2	Insertion sequences and phages	1
	Hypothetical	5	Hypothetical	3	Hypothetical	2
	Intergenic	1	-		-	
	Total	32	Total	20	Total	11

Annotation of SNPs overlapping between SAWC 4498 and SAWC 5165 revealed that five common SNPs were intergenic, whilst 18 of the total number of 33 common SNPs led to non-synonymous changes. Ten of the overlapping SNPs were synonymous (Table 4.1.5). The reference strain of *M. tuberculosis* F11 used for alignment in this section is multi-drug resistant and contains the drug resistance conferring mutations; *katG*315 (AGC to ACC) and *rpoB*531 (TCG to TTG), with respect to the drug-susceptible reference strain *M. tuberculosis* H37Rv. The two clinical isolates studied in the present section are drug susceptible and are thus not expected to harbour these mutations. Due to the approach of this quality control exercise, it is expected that the two isolates should have SNPs at *katG*315 (ACC to AGC) and *rpoB*531 (TTG to TCG), since the drug resistant strain is used as a reference in this section. The nucleotide sequence corresponding to *katG*315 and *rpoB*531 were indeed AGC and TCG respectively, thus indicating that the strains are susceptible to the first line anti-tuberculosis drugs isoniazid and rifampicin and the validity of the quality of the analysis.

Table 4.1.5 Analysis of overlapping SNPs between SAWC 4498 and SAWC 5165

	All SNPs		**Non-synonymous SNPs**		**Synonymous SNPs**	
SAWC 4498/SAWC 5165 overlapping	**Functional category of genes in which SNPs occur**	**Number of SNPs**	**Functional category of genes in which SNPs occur**	**Number of SNPs**	**Functional category of genes in which SNPs occur**	**Number of SNPs**
	Information pathways	1	Information pathways	1	Information pathways	0
	Conserved hypothetical proteins	5	Conserved hypothetical proteins	4	Conserved hypothetical proteins	1
	Cell wall and cell processes	5	Cell wall and cell processes	3	Cell wall and cell processes	2
	Intermediary metabolism and respiration	5	Intermediary metabolism and respiration	3	Intermediary metabolism and respiration	2
	Regulatory proteins	2	Regulatory proteins	1	Regulatory proteins	1
	Virulence, detoxification and adaptation	1	Virulence, detoxification and adaptation	1	Virulence, detoxification and adaptation	0
	Lipid metabolism	2	Lipid metabolism	1	Lipid metabolism	1
	PE/PPE protein families	0	PE/PPE protein families	0	PE/PPE protein families	0
	Insertion sequences and phages	6	Insertion sequences and phages	4	Insertion sequences and phages	2
	Hypothetical	1	Hypothetical	0	Hypothetical	1
	Intergenic	5	-		-	
	Total	33	Total	18	Total	10

4.1.4.2. *M. TUBERCULOSIS* H37RV AS A REFERENCE GENOME

The pre-processing and quality control of raw sequencing data was repeated for SAWC 4498 and SAWC 5165 with *M. tuberculosis* H37Rv as the reference genome, as described in Section 4.1.1.2. The high quality trimmed reads were used as input in the NGS analysis pipeline as reported in Section 4.1.1.3, with the exception that *M. tuberculosis* H37Rv was used as the reference genome during the read alignment step, and consequent high confidence variants are reported with respect to *M. tuberculosis* H37Rv.

The multi-software approach described in Chapter 3, Section 3.8.6 was used to identify variants in SAWC 4498 and SAWC 5165, with respect to *M. tuberculosis* H37Rv. Figure 4.1.7.A and -B show the number of SNPs that are identical in position and base identity that overlap between the pipelines for SAWC 4498 and SAWC 5165, respectively. A total number of 789 high confidence SNPs were identified in SAWC 4498 compared to the *M. tuberculosis* H37Rv reference genome (Figure 4.1.7.A). The SNPs identified by only one or two of the three pipelines were not regarded as true variants. The *M. tuberculosis* SAWC 5165 isolate diverged from the *M. tuberculosis* H37Rv reference genome by 816 high confidence SNPs (Figure 4.1.6.B).

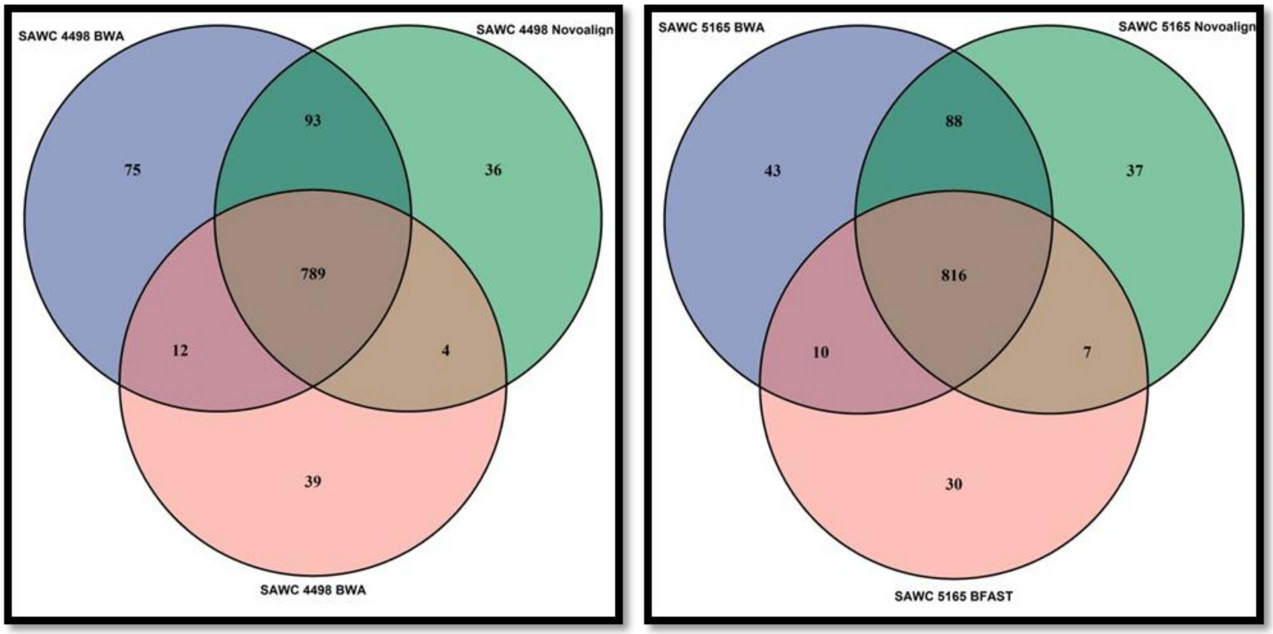


Figure 4.1.7. SNPs in SAWC 4498 and SAWC 5165. A. SNPs called with SAMTools and GATK using the alignments obtained from different mapping algorithms for SAWC 4498, employing *M. tuberculosis* H37Rv as a reference genome. B. SNPs called with SAMTools and GATK using the alignments obtained from different mapping algorithms for SAWC 5165, employing *M. tuberculosis* H37Rv as a reference genome.

The Venn diagrams shown in Figure 4.1.8.A and -B portrays the number of in/dels that are identical in start- and end position (length) for SAWC 4498 and SAWC 5165 respectively, in comparison with the *M. tuberculosis* H37Rv reference genome. A total number of 62 high confidence in/dels were identified in SAWC 4498 compared to the *M. tuberculosis* H37Rv reference genome (Figure 4.1.8.A). The in/dels identified by only one or two of the three pipelines were not regarded as true variants, upon further inspection it was shown that the alignment algorithms often did not agree on the start and end positions of the in/dels, despite realignment around in/dels being done. Mapping quality around the positions of these in/dels only identified by one or two of the pipelines was low. The *M. tuberculosis* SAWC 5165 isolate differed from the *M. tuberculosis* H37Rv reference genome by 64 high confidence in/dels (Figure 4.1.8.B).

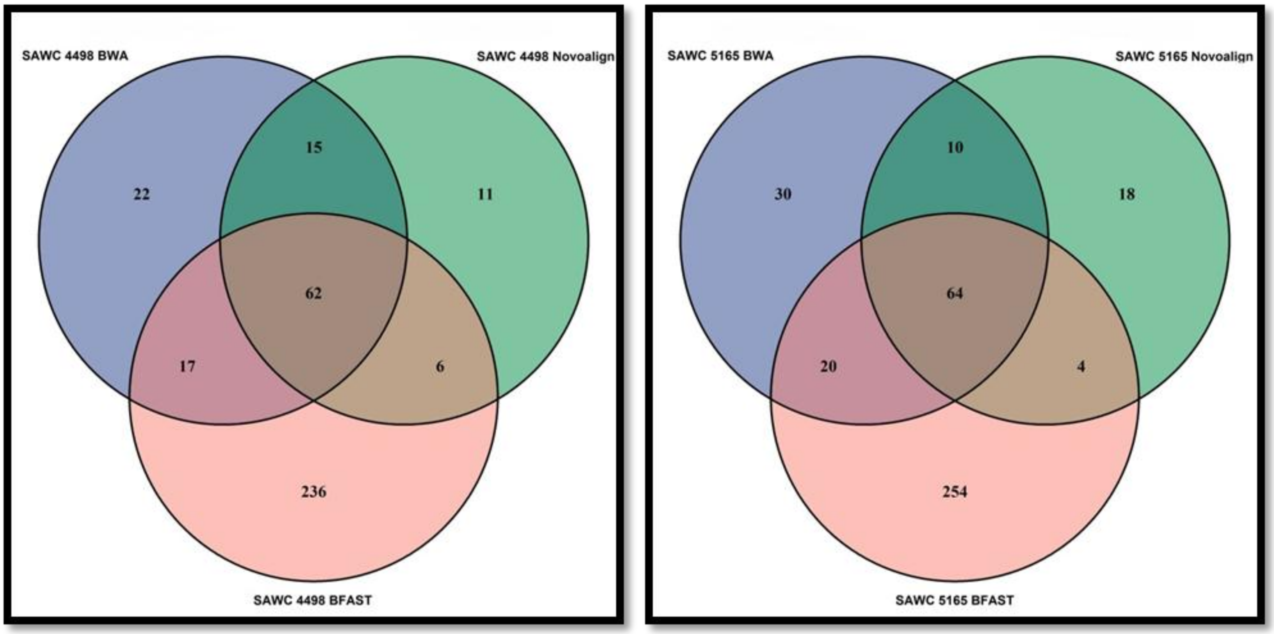


Figure 4.1.8. In/dels in SAWC 4498 and SAWC 5165. A. In/dels called with GATK using the alignments obtained from different mapping algorithms for SAWC 4498, employing *M. tuberculosis* H37Rv as a reference genome. B. In/dels called with GATK using the alignments obtained from different mapping algorithms for SAWC 5165, employing *M. tuberculosis* H37Rv as a reference genome.

The heterogeneity of the strains again highlights the diverse pool of strains present in the high TB incidence setting from which these strains were isolated. Figure 4.1.9.A shows a Venn diagram of the SNPs that overlap and are unique to SAWC 4498 and SAWC 5165 with respect to the *M. tuberculosis* H37Rv reference genome used in the analyses. SAWC 4498 and SAWC 5165 share 728 mutual SNPs and have 88 and 61 SNPs respectively that are unique to the strain (both in position and base identity). SAWC 4498 and SAWC 5165 share 56 in/dels with regards to *M. tuberculosis* H37Rv and have eight and six unique in/dels, respectively (Figure 4.1.9.B).

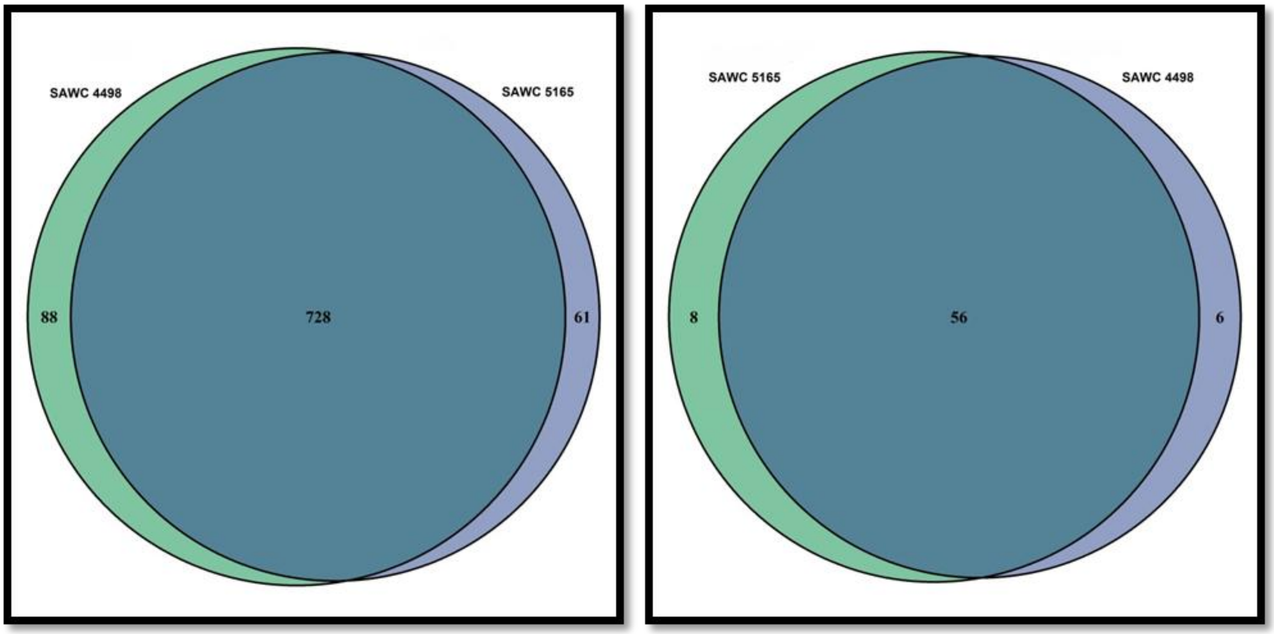


Figure 4.1.9. Overlapping variants between SAWC 4498 and SAWC 5165. A. SNPs that overlap in position and base identity between, or unique to, SAWC 4498 and SAWC 5165, employing *M. tuberculosis* H37Rv as a reference genome. B. In/dels that overlap in length, start- and end position and base identity between, or unique to, SAWC 4498 and SAWC 5165, employing *M. tuberculosis* H37Rv as a reference genome.

When the *M. tuberculosis* F11 strain was first sequenced and published, a 1094 bp deletion with respect to *M. tuberculosis* H37Rv was identified. This region contains the genes encoding *Rv1334* – *Rv1336* (Gagneux *et al*., 2006; Victor *et al*., 2004). The deletion named RD761 is thought to be characteristic to the *M. tuberculosis* F11 strain family. The approach followed in this study allowed for the identification of the deletion in SAWC 4498 and SAWC 5165, as shown in Figure 4.1.10 and Figure 4.1.11, respectively, confirming the identity of SAWC 4498 and SAWC 5165.

The alignments of a small genomic section done with the three different alignment algorithms; BFAST, BWA and Novoalign, for SAWC 4498, using *M. tuberculosis* H37Rv as reference genome, is shown in Figure 4.1.9.A, -B, and C, respectively. The section of the *M. tuberculosis* H37Rv genome shown includes *Rv1334* – *Rv1336*, the genes deleted in *M. tuberculosis* F11 strain family (Gagneux *et al*., 2006; Victor *et al*., 2004). In the BFAST alignment (Figure 4.1.10.A) there are no reads mapping to the genomic region from position 1 502 787 to 1 503 881, whilst there are spurious reads mapping to this region in the BWA- and Novoalign alignments (Figures 4.1.10.B and –C). The spurious reads are in all probability erroneously mapped to this region in these two alignments. The *M. tuberculosis* F11 clinical isolate SAWC 4498 visibly harbours RD761.

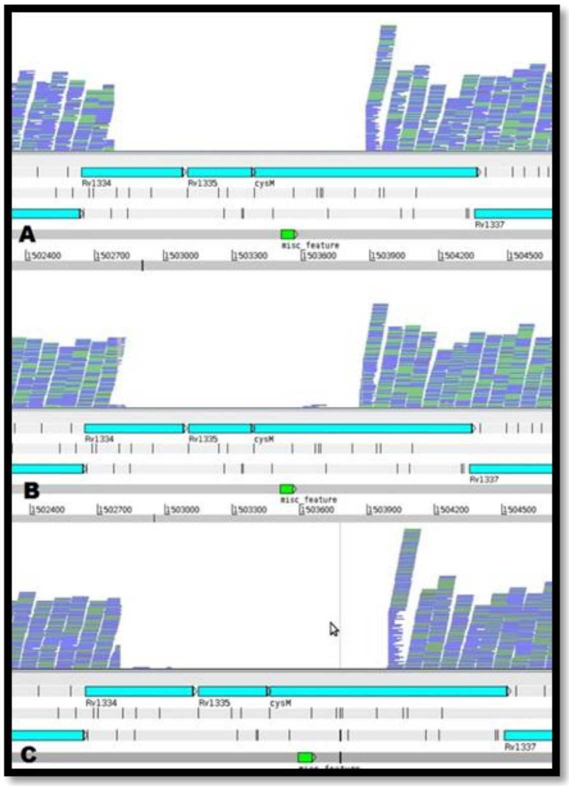


Figure 4.1.9. Alignment of reads across the genomic region containing *Rv1334 – Rv1336* in SAWC 4498. A. A screenshot of the visualisation of the BFAST alignment of the raw sequencing reads of SAWC 4498 to the *M. tuberculosis* H37Rv reference genome, indicating the presence of RD761. B. A screenshot of the visualisation of the BWA alignment of the raw sequencing reads of SAWC 4498 to the *M. tuberculosis* H37Rv reference genome, indicating the presence of RD761. C. A screenshot of the visualisation of the Novoalign alignment of the raw sequencing reads of SAWC 4498 to the *M. tuberculosis* H37Rv reference genome, indicating the presence of RD761.

The alignments across the genomic section spanning *Rv1334, Rv1335,* and *Rv1336*, in the genome of SAWC 5165 is shown in Figure 4.1.11.A, -B, and -C, respectively. The alignments produced by the three different alignment algorithms show the expected read coverage of the region upstream from the genomic position 1 502 787 and downstream to position 1 503 881, whilst there are spurious reads mapping to the genomic region between 1 502 787 and 1 503 881, spanning *Rv1334 – Rv1336*. The spurious reads are in all likelihood incorrectly mapped to this region. The *M. tuberculosis* F11 clinical isolate SAWC 5165 evidently harbours RD761.

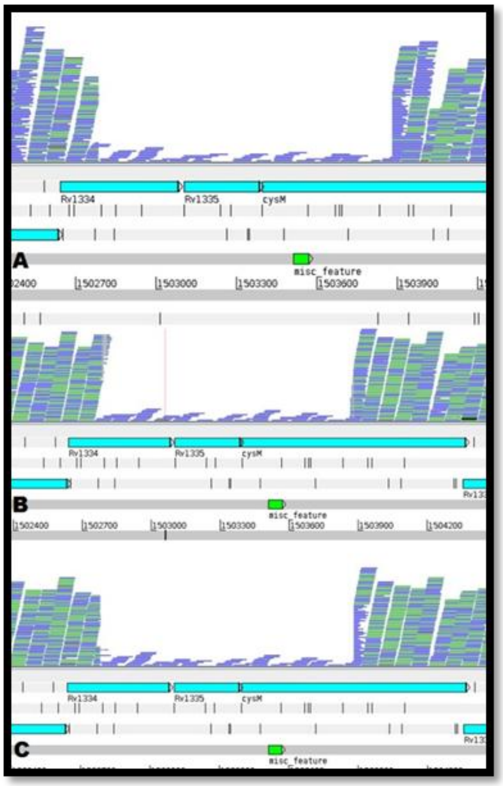


Figure 4.1.11. Alignment of reads across the genomic region containing *Rv1334 – Rv1336* in SAWC 5165. A. A screenshot of the visualisation of the BFAST alignment of the raw sequencing reads of SAWC 5165 to the *M. tuberculosis* H37Rv reference genome, indicating the presence of RD761. B. A screenshot of the visualisation of the BWA alignment of the raw sequencing reads of SAWC 5165 to the *M. tuberculosis* H37Rv reference genome, indicating the presence of RD761. C. A screenshot of the visualisation of the Novoalign alignment of the raw sequencing reads of SAWC 5165 to the *M. tuberculosis* H37Rv reference genome, indicating the presence of RD761.

The *M. tuberculosis* F11 genotype isolates analysed here also contain RD 149 (see Table 2.4.1), which was accurately detected in the WGS analyses of both SAWC 4498 and SAWC 5186 (data not shown).

4.1.5 SUB-CONCLUSIONS

In Section 4.1, the accuracy and quality of the customised NGS data analysis pipeline was evaluated and it was shown to be highly accurate. The pre-processed paired end reads mapped nearly equally well to both reference genomes employed. In general, only 0.26 – 0.49% more reads mapped to the *M. tuberculosis* F11 reference genome than to the *M. tuberculosis* H37Rv reference genome. Expectedly, considerably fewer SNPs and small in/dels were identified when mapping the reads to the *M. tuberculosis* F11 reference genome, than mapping to –H37Rv. This was expected, as *M. tuberculosis* H37Rv (a member of PGG3) is situated much further from the clinical isolates than *M. tuberculosis* F11 (a member of PGG2). All known genetic features of the *M. tuberculosis* F11 strain family were successfully identified with the analysis pipeline, highlighting the accuracy of the pipeline and providing more confidence in the ability of a multi-software approach to find distinct genetic variance in the isolates that were analysed in this section, as well as the rest of the isolates analysed in this study. However, the number of variations (SNPs and in/dels) between the alignments of the reads of the two isolates, SAWC 4498 and SAWC 5165, to the two reference genomes employed, did not agree. The read-alignment algorithms that were used in the multi-software approach are largely non-heuristic and thus aim to find the optimal alignment for each read. A combination of the repetitive nature of the *M. tuberculosis* genome and limited read length make some level of mapping error inevitable. This is especially true when aligning the sequencing reads of the *M. tuberculosis* F11 clinical isolates to the *M. tuberculosis* H37Rv reference genome, since *M. tuberculosis* H37Rv is phylogenetically further removed from the isolates than *M. tuberculosis* F11. Despite some discrepancies when using the different reference genomes as reference in the NGS data analysis pipeline, the known genetic features of the *M. tuberculosis* F11 clinical isolates was identified with high accuracy in the clinical isolates (using *M. tuberculosis* H37Rv as reference genome). The multi-software approach, using *M. tuberculosis* H37Rv, thus proves to be reliable for the identification of fixed genetic variation between different *M. tuberculosis* clinical isolates and strain families. This approach is particularly suitable for phylogenetic inference described in Section 4.2 and 4.3.

4.2. *M. TUBERCULOSIS* PHYLOGENY

4.2.1. INTRODUCTION

This section focusses on the reconstruction of the phylogeny of drug susceptible *M. tuberculosis* isolates from a number of *M. tuberculosis* strain families representative of the TB epidemic in Cape Town, South Africa. Fifty clinical isolates from different strain families, based on *IS6110* RFLP fingerprints, from the local setting that were identified as fully drug susceptible *M. tuberculosis* strains were included. These strains were subjected to whole genome sequencing on the Illumina HiSeq 2000 platform and the customised pipeline was used to align the reads obtained for these strains to the completed sequence of *M. tuberculosis* H37Rv as a reference (Chapter 3, Section 3.8.3). Ten genome sequence datasets of different published *M. tuberculosis* strains were included in the analysis (Table 3.2). The raw sequencing data for these strains were downloaded and were also subjected to analysis with the customised pipeline. *Mycobacterium canettii* was included as a closely related outgroup (Hershberg *et al*., 2008). High confidence variants for all isolates included in the analysis were taken into consideration and the strings of the concatenated variants in fasta format were aligned to each other with ClustalW (Larkin *et al.,* 2007). The approach followed is similar to previous phylogenetic studies which investigated the evolution of *M. tuberculosis* strains using genome wide SNPs (Comas *et al*., 2013; Gutacker *et al*., 2002; Namouchi *et al*., 2012). The resulting alignment was used to construct the phylogeny discussed in this section. The different cladograms obtained by different phylogeny software packages were compared and major nodes proved to be similar in the trees produced.

4.2.2. PHYLOGENY

A total number of 27 530 variant positions were identified in the 61 isolates (50 clinical isolates, 10 previously published sequences, *M. tuberculosis* H37Rv) that were included in the analyses. The concatenated sequence for every isolate included all possible SNP positions identified, if the isolate did not contain an SNP at a specific position, the reference (*M. tuberculosis* H37Rv) base at that position was recorded to ensure that the string of nucleotide sequences recorded for each isolate included all of the 27 530 bases being analysed.

4.2.2.1. *M. TUBERCULOSIS* GENERAL PHYLOGENY

A comprehensive phylogenetic reconstruction of the *M. tuberculosis* clinical isolates and selected published *Mycobacterium sp*. based on genome wide SNPs is shown in Figure 4.2.1. In agreement with previous studies, *Mycobacterium canettii* (indicated as M_canetti in Figure 4.2.1) proves to be ancestral to the evolution of the human adapted *Mycobacterium* species – *Mycobacterium africanum* (indicated as Maf in Figure 4.2.1) and *M. tuberculosis*. Likewise, *M. africanum* is ancestral to *M. tuberculosis*. It is evident from the phylogeny that the *M. africanum* lineage five isolates (5444 and Maf_lin5) cluster together and can be distinguished from *M. africanum* lineage six (MAf_lin6). Furthermore, the ancestral human adapted *M. tuberculosis* strains, the East African Indian (EAI) strain (SAWC 1659) and the recently reported *M. tuberculosis* lineage seven strain (Mtb_l7), clearly precedes their modern descendants (Blouin *et al*., 2012; Flores *et al*., 2007; Tsolaki *et al*., 2004). These ancestral strains (*M. canettii*, *M. africanum*, *M. tuberculosis* EAI and *M. tuberculosis* lineage 7) all have the TBD1 region present, in contrast with their modern descendants in which this region is deleted, see Figure 4.2.1 (Brosch *et al*., 2002).

The three principal genetic groups based on SNPs in the *katG* and *gyrA* genes noticeably cluster in the phylogeny summarised in the cladogram depicted in Figure 4.2.1. The clade forming PGG1 is comprised of CAS- (SAWC 995, SAWC 3385, SAWC 4370, SAWC 3740), typical Beijing- (SAWC 2701, SAWC 1116, SAWC 4570, SAWC 1125, SAWC 4437), and atypical Beijing (SAWC 6519, SAWC 1453) strains. The clade containing all Beijing strains (typical and atypical) is nested within the PGG1 clade and the CAS clade forms its sister clade.

The evolution of PGG2 is visibly more complex than that of PGG1. The combined clade forming PGG2 and PGG3 is a sister clade to PGG1. PGG2 is comprised of one clade, made up of the Low copy clade (LCC)- (SAWC 2336, SAWC 4336, SAWC 4972, SAWC 3200), pre-Haarlem- (SAWC 2026), Haarlem-like- (SAWC 4046, SAWC 4978) and Haarlem strains (Erdmann, SAWC 3448, SAWC 2282, SAWC 7367, SAWC 6680), and an evolutionary grade that includes a possible intermediary LCC strain (SAWC 3933), Quebec strain (SAWC 1595), the LAM strains (SAWC 4498, SAWC 5165, SAWC 1955, SAWC 3517, LAM 1503, SAWC 2904, SAWC 2511, SAWC 5276, SAWC 5260, SAWC 2262, SAWC 3656, SAWC 3388, SAWC 3100, SAWC 3651, KZN 605, SAWC 2576, KZN 4207, KZN 1435), and a group of strains with similar but undesignated spoligotype names that belong to *IS6110* RFLP family 32 (SAWC 1123, SAWC 2005). These two strains are confirmed members of PGG2 based on the *gyrA*95 ACC sequence and seem to be the precursors to PGG3. From this phylogenetic reconstruction, it is suggested that the Erdmann strain (Miyoshi-Akiyama *et al*., 2012) could be a Haarlem precursor. The Quebec strain (SAWC 1595) is basal to the clade including the LAM family of strains, F32 strains and PGG3. The evolution of the LAM family of strains is shown in Figure 4.2.1 and -4.2.2 and will be discussed in Section 4.2.2.2. The complexity of the evolutionary relationship among PGG2 strains found in Cape Town, South Africa is highlighted by the phylogenetic reconstruction shown in Figure 4.2.1. The predominant lineages within PGG2 have evolved independently from a common progenitor.

PGG3 is comprised of a number of strains that form a clade that is nested within the major clade that spans over PGG2 and PGG3. PGG3 has thus evolved from an ancestral PGG2 member. PGG3 also includes the *M. tuberculosis* H37Rv laboratory strain (Mtb_H37Rv) that is employed as the reference strain in the current study's bioinformatics analysis. PGG3 represents the most modern group of *M. tuberculosis* strains and includes T-strains (SAWC 5330, SAWC 3839, SAWC 1397, SAWC 478, SAWC 337, SAWC 4472) as well as strains with undesignated spoligotypes (SAWC 5440, SAWC 1543, SAWC 4302).

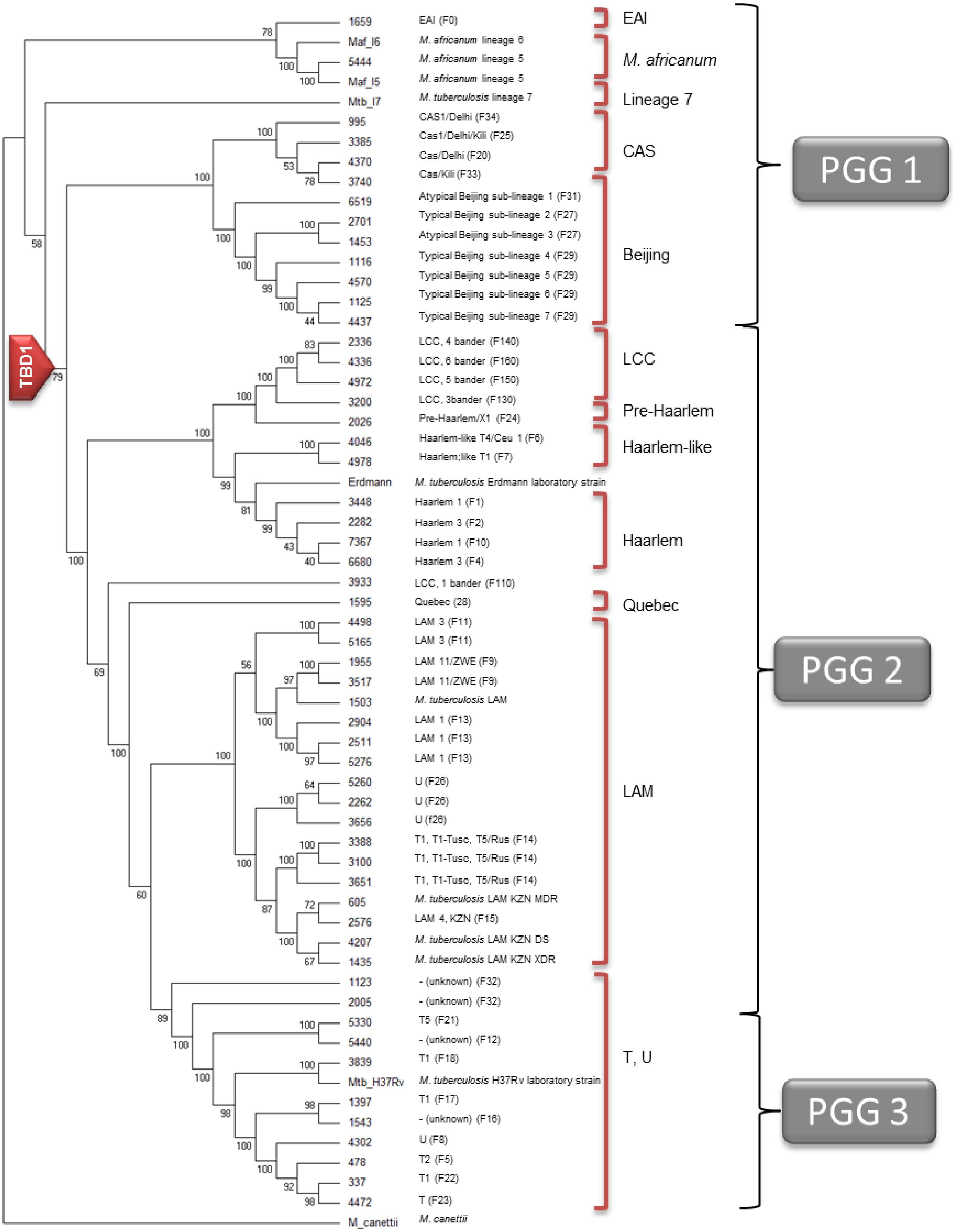

1659 EAI (F0)
Maf_I6 M. africanum lineage 6
5444 M. africanum lineage 5
Maf_I5 M. africanum lineage 5
Mtb_I7 M. tuberculosis lineage 7
995 CAS1/Delhi (F34)
3385 Cas1/Delhi/Kili (F25)
4370 Cas/Delhi (F20)
3740 Cas/Kili (F33)
6519 Atypical Beijing sub-lineage 1 (F31)
2701 Typical Beijing sub-lineage 2 (F27)
1453 Atypical Beijing sub-lineage 3 (F27)
1116 Typical Beijing sub-lineage 4 (F29)
4570 Typical Beijing sub-lineage 5 (F29)
1125 Typical Beijing sub-lineage 6 (F29)
4437 Typical Beijing sub-lineage 7 (F29)
2336 LCC, 4 bander (F140)
4336 LCC, 6 bander (F160)
4972 LCC, 5 bander (F150)
3200 LCC, 3bander (F130)
2026 Pre-Haarlem/X1 (F24)
4046 Haarlem-like T4/Ceu 1 (F8)
4978 Haarlem;like T1 (F7)
Erdmann M. tuberculosis Erdmann laboratory strain
3448 Haarlem 1 (F1)
2282 Haarlem 3 (F2)
7367 Haarlem 1 (F10)
6680 Haarlem 3 (F4)
3933 LCC, 1 bander (F110)
1595 Quebec (28)
4498 LAM 3 (F11)
5165 LAM 3 (F11)
1955 LAM 11/ZWE (F9)
3517 LAM 11/ZWE (F9)
1503 M. tuberculosis LAM
2904 LAM 1 (F13)
2511 LAM 1 (F13)
5276 LAM 1 (F13)
5260 U (F26)
2262 U (F26)
3656 U (f26)
3388 T1, T1-Tusc, T5/Rus (F14)
3100 T1, T1-Tusc, T5/Rus (F14)
3651 T1, T1-Tusc, T5/Rus (F14)
605 M. tuberculosis LAM KZN MDR
2576 LAM 4, KZN (F15)
4207 M. tuberculosis LAM KZN DS
1435 M. tuberculosis LAM KZN XDR
1123 - (unknown) (F32)
2005 - (unknown) (F32)
5330 T5 (F21)
5440 - (unknown) (F12)
3839 T1 (F18)
Mtb_H37Rv M. tuberculosis H37Rv laboratory strain
1397 T1 (F17)
1543 - (unknown) (F16)
4302 U (F8)
478 T2 (F5)
337 T1 (F22)
4472 T (F23)
M_canettii M. canettii
EAI
M. africanum
Lineage 7
CAS
Beijing
LCC
Pre-Haarlem
Haarlem-like
Haarlem
Quebec
LAM
T, U
PGG 1
PGG 2
PGG 3
TBD1
78
100
100
100
53
78
100
100
100
100
99
100
44
58
83
100
100
100
100
100
99
81
99
43
40
79
100
100
56
100
97
100
100
97
64
100
69
100
100
100
100
72
87
100
67
60
89
100
100
100
100
98
98
100
100
92
98

Figure 4.2.1. Genome-wide SNP-based phylogeny of *M. tuberculosis* clinical isolates and selected published genomes of *Mycobacterium* species. The evolutionary history was inferred by using the Maximum Likelihood method based on the General Time Reversible model (Nei and Kumar, 2000). The bootstrap consensus tree inferred from 1000 replicates is taken to represent the evolutionary history of the taxa analysed. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (1000 replicates) is shown next to the branches (Felsenstein, 1989). Initial tree(s) for the heuristic search were obtained by applying the Neighbor-Joining method to a matrix of pairwise distances estimated using the Maximum Composite Likelihood (MCL) approach. The analysis involved 62 nucleotide sequences. All positions containing gaps and missing data were eliminated. There were a total of 27 530 positions in the final dataset. Evolutionary analyses were conducted in MEGA5 (Tamura *et al*., 2011).

4.2.2.2. *M. TUBERCULOSIS* LAM PHYLOGENY

The comprehensive phylogenetic reconstruction of the LAM subgroup of *M. tuberculosis* is shown in Figure 4.2.2. The position of the LAM strains in the overall evolution of *M. tuberculosis* can be seen in Figure 4.2.1. The LAM group of strains is evolutionarily complex and consists of six *IS6110* RFLP families, namely F11, F9, F13, F26, F14 and F15, forming the six LAM sub-lineages. Strains belonging to the six *IS6110* RFLP families cluster together without exception.

Within the clade that forms the LAM family of strains are two major sister clades. One of these clades consists of the LAM3 (F11) strains, the LAM11/ZWE (F9) strains, including the previously published LAM 1503 strain, and the LAM1 strains (F13). Within this clade, the LAM11/ZWE (F9)- and LAM1 (F13) strains cluster together to form a sister clade to the LAM3 (F11) clade. The second clade is comprised of the strains with unknown (U) spoligotypes (F26), forming a sister clade to the T1, T1-Tusc, T5/Rus (F14) strains and the KZN (LAM4, F15) strains.

From the phylogenetic reconstruction (Figure 4.2.2), the LAM4 KZN (F15) clinical isolate (SAWC 2576) included in this study clusters with the LAM4 KZN strains sequenced by the Broad Institute. The clinical isolate (SAWC 3517) clusters with the KZN 605 MDR isolate to form clade nested within the LAM4 KZN clade. Within this clade, the KZN drug sensitive isolate and KZN XDR isolate cluster together.

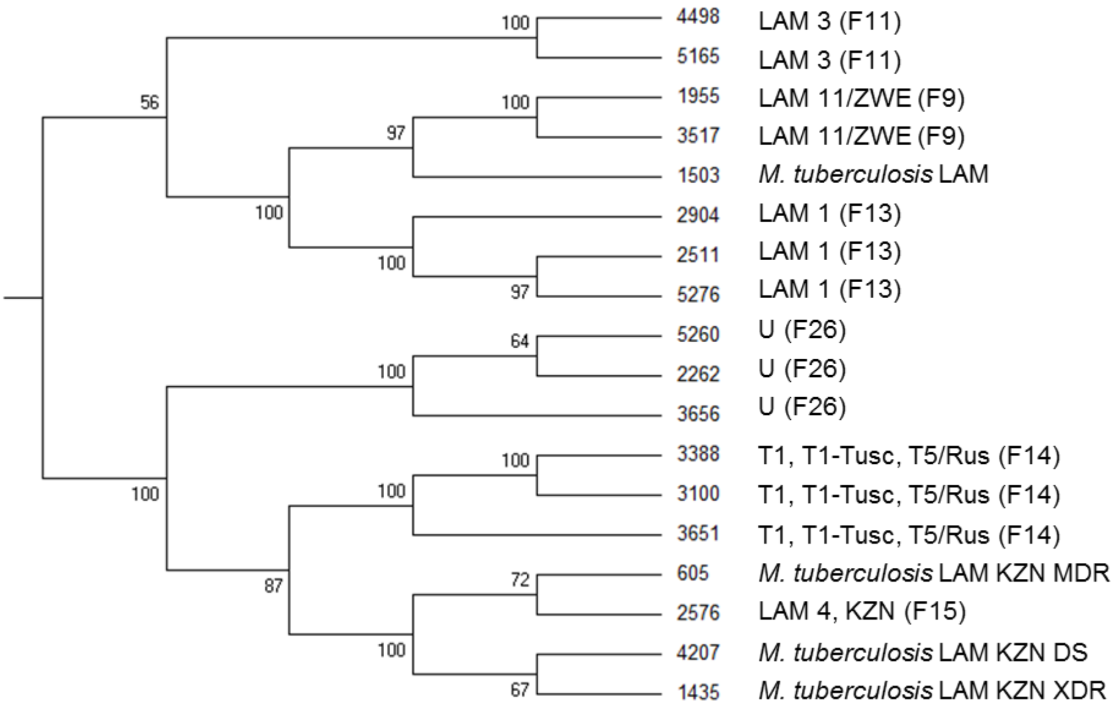


Figure 4.2.2. Genome wide SNP-based phylogeny of the LAM family of strains of *M. tuberculosis*. The evolutionary history was inferred in the same manner as described for Figure 4.2.1. This phylogenetic tree is a sub-tree of the general *M. tuberculosis* and *M. africanum* tree shown in Figure 4.2.1.

4.2.3. SUB-CONCLUSIONS

In this part of the study, the unprecedented resolution provided by whole genome sequencing for investigating strain differentiation and evolution was exemplified by the robust phylogenetic data obtained. Previous studies have employed subsets of SNPs, sequences of subsets of genes and other informative genetic markers to infer phylogenetic relationships among the different strains of *M. tuberculosis* (Baker *et al*., 2004, Filliol *et al*., 2006, Gutacker *et al*., 2006). SNPs contribute to a large proportion of mycobacterial genome variation and have been proved to be able to delineate closely related isolates from each other (Gutacker *et al*., 2002). The present study employed genome wide SNPs of all isolates included in the analysis to reconstruct the phylogeny of selected *M. tuberculosis* and *M. africanum* strains. The same general groupings were found when different phylogenetic inference software packages were used (see Section 3.5.6.11). The data showed a clonal population structure of *M. tuberculosis* consisting of distinct lineages. *IS6110* RFLP families and spoligotypes were specifically associated with defined SNP clusters, thereby confirming the robustness of the results. The general phylogenetic structure shows congruence with former genotypic groupings of *M. tuberculosis* (Baker *et al*., 2004, Filliol *et al*., 2006, Gutacker *et al*., 2006). The general lineages of *M. tuberculosis* as defined previously by both SNPs and RDs could be resolved by this phylogenetic reconstruction of the evolution of *M. tuberculosis* (Baker *et al*., 2004; Gagneux *et al*., 2006). This SNP-based phylogeny also mimics the phylogeny constructed using large deletion events, taking into account that in *M. tuberculosis* genetic material is more readily lost than gained (Pym and Brosch, 2000). This similarity is emphasised by the mapping of the RDs and IS*6110* insertion sites as depicted in Figure 4.3.11. It is evident that in this study setting tuberculosis is caused by an assortment of strains imported from around the globe. This genome-wide SNP-based phylogenetic reconstruction of the evolution of *M. tuberculosis* offers novel insights into the unique global representation of the *M. tuberculosis* isolates included in the analysis. Comprehensive phylogenetic studies such as the present one lay the foundation for studies that examine the relationship between genotype and disease phenotype in large, mixed-strain sample sets.

4.3. FUNCTIONAL ANALYSIS OF GENETIC VARIATION IN LAM ISOLATES

4.3.1. INTRODUCTION

This section examines the variants that are unique to all the LAM isolates included in the present study, and variants unique to the LAM sub-lineages. Figure 4.3.1 shows the evolutionary relationships among the LAM isolates included in the phylogenetic study, the different sub-lineages are indicated in grey boxes and named according to the *IS6110* RFLP families. These *IS6110* RFLP family names will be used to refer to certain LAM sub-lineages or groups of sub-lineages that can be differentiated by specific nucleotide variants.

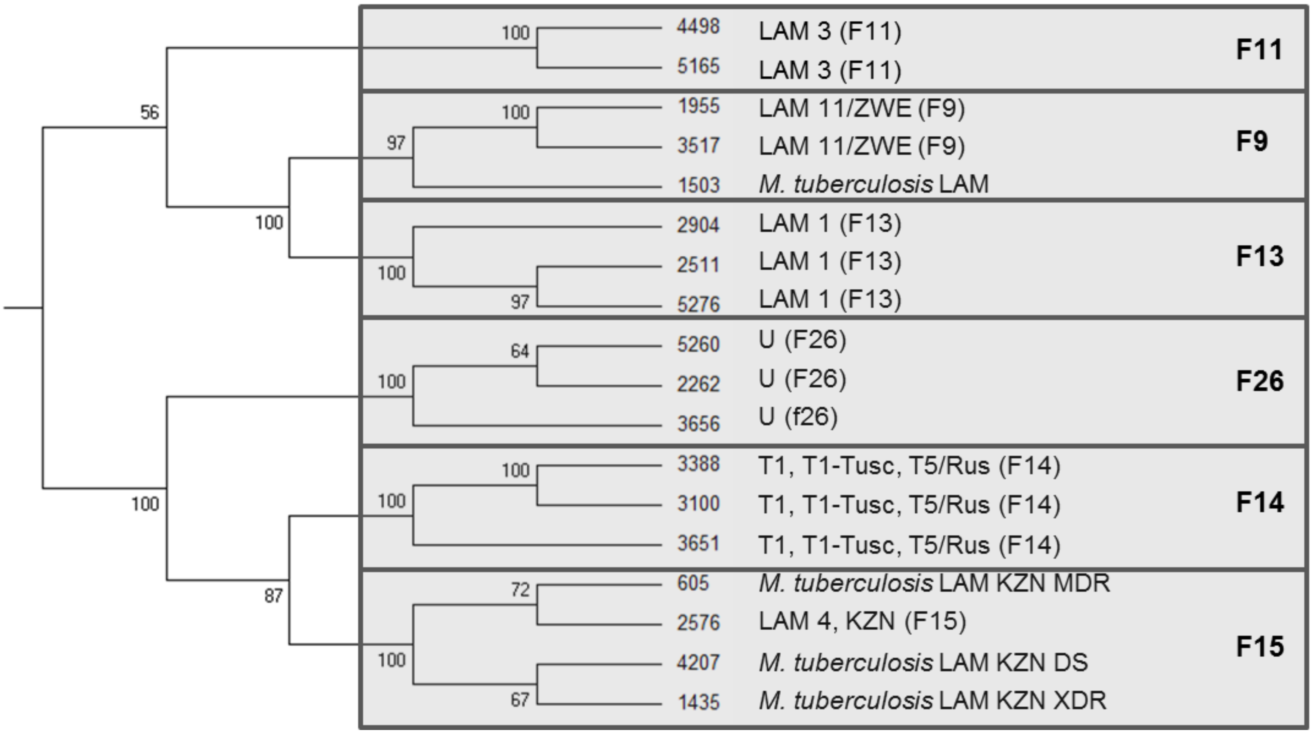


Figure 4.3.1. Genome wide SNP-based phylogeny of the LAM family of strains of *M. tuberculosis* indicating sub-lineages of the LAM genotype. The different LAM sub-lineages are separated by grey boxes and named according to the different IS*6110* RFLP families.

4.3.2. *M. TUBERCULOSIS* LAM SNP ANALYSIS

The high confidence SNPs identified in a total of 56 isolates included in the present study were analysed and comparisons were made between different groups of strains. *M. africanum* isolates and the *M. canetti* isolate were excluded when doing the SNP analysis. SNPs in the LAM group of strains (n=18) were compared to all non-LAM strains included in the study to identify SNPs that are uniquely identified in all of the LAM strains included, or sub-groups of LAM strains included, but are not present in any of the non-LAM strains (n=38) included in the analysis. This analysis was not restricted to the clinical isolates sequenced for the purpose of this study, but included the publically available *M. tuberculosis* data. All SNPs in specific genes are comprehensively listed in Appendix C.

4.3.2.1. LAM SPECIFIC SNPS

As shown in Table 4.3.1, a total of 55 SNPs were shown to be uniquely present in all LAM isolates included in the analysis (n=18), and can thus be considered as LAM-specific SNPs. Of these 55 SNPs, nine are found in non-coding regions (intergenic), and of the remaining 46 SNPs, 15 were synonymous and 31 were non-synonymous. The large proportion of non-synonymous SNPs observed, is consistent with previous findings and suggests that *M. tuberculosis* has not been under high purifying selective pressure (Hershberg *et al*., 2008). It is hypothesised that clonal organisms such as *M. tuberculosis* undergo random mutation events. In the absence of selective pressures, synonymous- and non-synonymous changes are expected to be equally prevalent. When purifying selection is taking place, synonymous nucleotide changes are overrepresented. An overrepresentation of non-synonymous changes indicates the presence of positive selection (Hershberg *et al*., 2008). However, it has also been shown that large numbers of non-synonymous changes may be due to the close relatedness of the isolates under investigation. Somewhat deleterious non-synonymous genomic changes that are not fixed within populations may not yet have been removed by purifying selection.

Table 4.3.1. Functional distribution of SNPs common to all LAM isolates

Functional category	Number of SNPs	Non-synonymous	Synonymous
Information pathways	7	3	4
Conserved hypothetical proteins	10	6	4
Cell wall and cell processes	10	9	1
Intermediary metabolism and respiration	7	3	4
Regulatory proteins	1	0	1
Virulence, detoxification and adaptation	2	2	0
Lipid metabolism	5	4	1
Insertion sequences and phages	2	2	0
PE/PPE	2	2	0
Intergenic	9	-	-
Total	55	31	15

Figure 4.3.2 shows the functional distribution of genes containing the SNPs common to all LAM isolates included in this study, but absent from all the non-LAM isolates included. Due to the high number of protein coding sequences classified as “conserved hypothetical” or “unknown”, it is expected that a high proportion of genome wide SNPs would be assigned to this category upon

annotation and functional assignment (Camus *et al*., 2002; Cole *et al*., 1998). Of the total number of 55 SNPs identified in all LAM isolates that were absent from their non-LAM counterparts, 10 SNPs occur in genes associated with the mycobacterial cell wall and cell processes. Proteins in this category include membrane associated proteins as well as secreted and trans-membrane proteins (Camus *et al*., 2002). Other functional categories that were highly represented among the genes harbouring LAM-specific SNPs are "intermediary metabolism and respiration" and "information pathways". Four of the SNPs in genes that are functionally categorised in "information pathways" were found in DNA replication, recombination and repair genes (3R-system). Three of these SNPs occur in the highly mutable genes *recC, dnaG and ligB* (Mestre *et al*., 2011; Dos Vultos *et al*., 2008). Genes involved in the 3R-system are said to play a significant role in the evolution of exceedingly clonal organisms such as *M. tuberculosis* (Dos Vultos *et al*., 2008). Other SNPs that are LAM-specific are found in the functional categories "lipid metabolism", "insertion sequences and phages", "PE/PPE", "virulence, detoxification and adaptation", and "regulatory proteins", in order of descending prevalence. These functional categories are however very broad and cellular processes are inherently linked. It is clear from the changes observed within the LAM lineage that these strains are adapting to their human hosts in order to ensure its survival and possibly increase its dissemination, as is the case with the Beijing family (Hanekom *et al*., 2011).

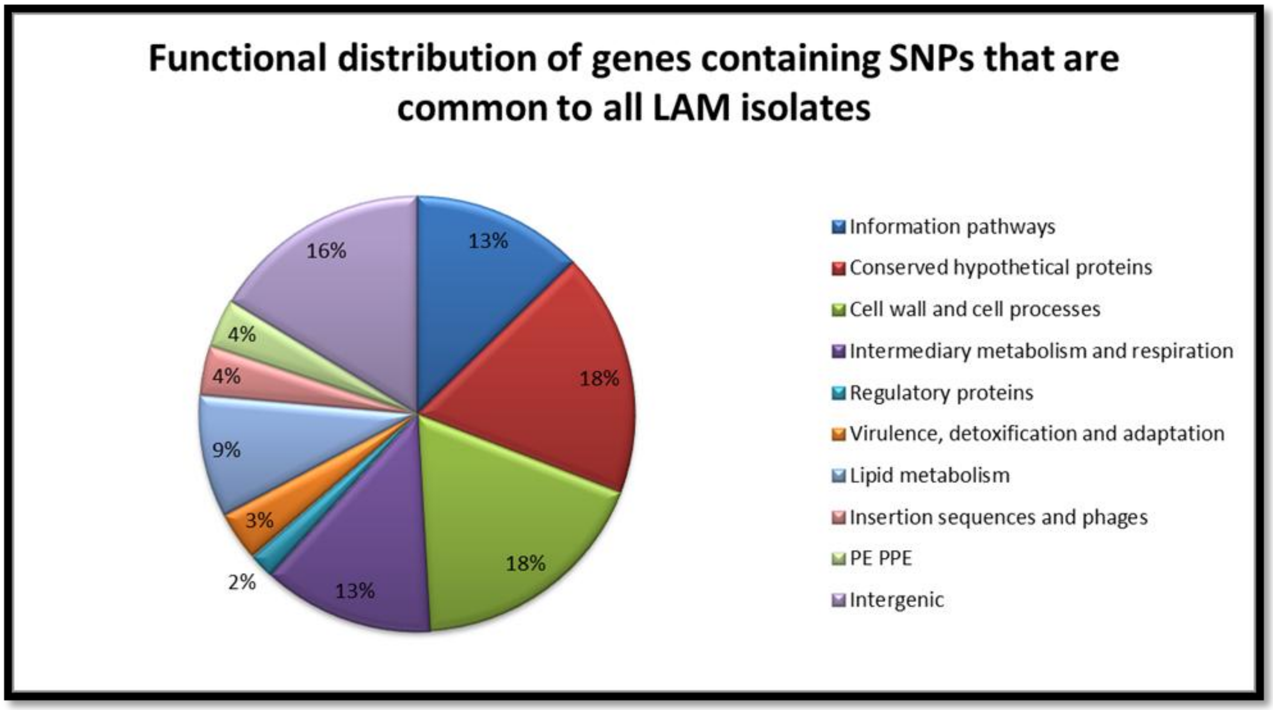


Figure 4.3.2. Functional distribution of genes in which LAM-specific SNPs occur. All SNPs that are common to 18 LAM isolates included in this analysis were functionally annotated. The SNPs listed are not found in any of the non-LAM isolates included in the analysis (n=38).

4.3.2.2. SNPS UNIQUE TO CLADES WITHIN THE LAM GENOTYPE

The topology of the phylogenetic tree shown in Figure 4.3.1 is determined by SNPs that occur uniquely in certain clades. These SNPs define the nodes and can be used to distinguish certain sub-lineages of the LAM family of strains from each other. SNPs that are responsible for the branching points in the phylogenetic hypothesis depicted in Figure 4.3.1 are summarised in Table 4.3.2.

Table 4.3.2. Summary of SNPs unique to clades of the LAM genotype

	F11		F9, F13		F9		F13		F26, F14, F15		F26		F14, F15		F14		F15	
Functional category	**Non-syn**	**Syn**	**Non-syn**	**Syn**	**Non-syn**	**Syn**	**Non-syn**	**Syn**	**Non-syn**	**Syn**	**Non-syn**	**Syn**	**Non-syn**	**Syn**	**Non-syn**	**Syn**	**Non-syn**	**Syn**
Information pathways	5	5	1	2	0	2	1	0	2	0	4	0	1	0	6	0	3	3
Conserved hypothetical proteins	14	10	9	5	2	3	4	3	4	0	11	11	1	1	14	6	8	4
Cell wall and cell processes	19	6	8	5	8	2	1	5	6	3	11	8	0	2	10	8	8	2
Intermediary metabolism and respiration	25	10	5	4	6	2	3	1	5	4	16	17	1	0	8	7	5	3
Regulatory proteins	5	1	0	0	0	1	0	0	1	1	2	0	1	0	3	0	0	0
Virulence, detoxification and adaptation	1	2	0	2	1	1	2	0	0	0	3	3	0	0	2	2	1	0
Lipid metabolism	11	5	1	2	2	2	0	1	3	2	6	2	2	1	5	3	7	0
Insertion sequences and phages	2	4	0	0	0	0	1	0	0	0	0	1	0	0	1	1	0	0
PE/PPE	9	5	0	0	3	0	1	1	1	0	8	4	1	0	4	3	3	1
Intergenic	25		4		0		6		5		12		2		7		9	
Total: syn + non-syn + intergenic	164		48		35		30		37		119		13		90		57	

Non-syn= non-synonymous, Syn=synonymous

SNPS UNIQUE TO LAM3 (F11)

The LAM3 (F11) genotype of *M. tuberculosis* is a prominent cause of TB in the Western Cape province of South Africa. This genotype shares many genetic characteristics with other members of the LAM family. This clade can be distinguished by 164 SNPs, of which 25 are in non-coding regions. Of the remaining SNPs, 91 are non-synonymous and 48 are synonymous (Table 4.3.2).

The functional distribution of the genes in which SNPs uniquely identified in the LAM3 (F11) isolates occur is shown in Figure 4.3.3. A large number of SNPs are located in genes coding for proteins involved in the intermediary metabolism and respiration, in contrast with other clades where the majority of the single nucleotide variation occur in genes involved in cell wall and cell processes or are in conserved hypothetical genes. Other highly represented functional groups include “cell wall and cell processes” and SNPs in conserved hypothetical genes. Ten percent of the variation is found in genes coding for proteins involved in the lipid metabolism.

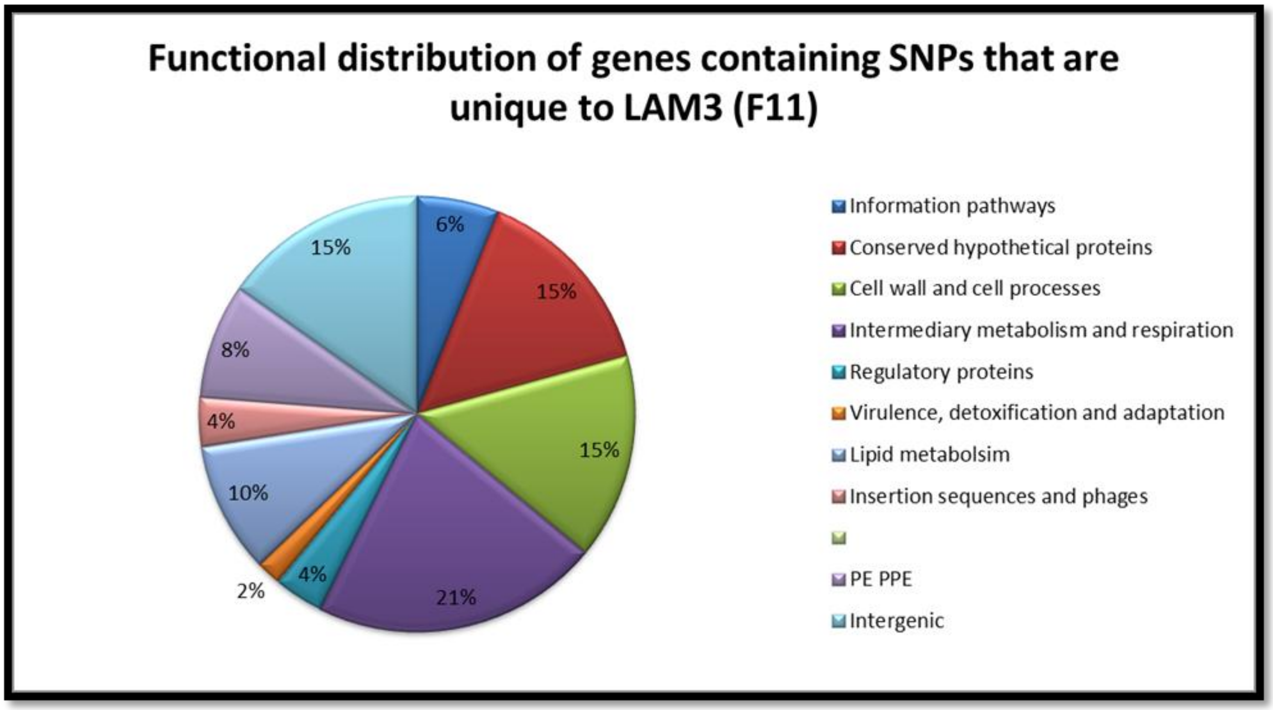


Figure 4.3.3. Functional distribution of genes harbouring SNPs common to LAM3 F11 (n=2). All SNPs that are common to LAM3 (F11) strains were functionally annotated. The SNPs listed are not found in any of the other LAM (LAM4 KZN (F15), T1, T1-Tusc, T5/Rus (F14), LAM U (F26), LAM11/ZWE (F9), LAM1 (F13)) or non-LAM isolates included in the analysis.

SNPS UNIQUE TO LAM11/ZWE (F9) AND LAM1 (F13)

In this study, 48 SNPs were identified to be unique to the clade consisting of LAM11/ZWE (F9) and LAM1 (F13), and were thus not identified in any of the other LAM- or non-LAM isolates included in the present study. Of the total of 48 SNPs uniquely identified in this subgroup of LAM isolates, only 4 are found in intergenic regions. Twenty four of the SNPs identified were non-synonymous and 20 were synonymous (Table 4.3.2).

The graphical presentation of the functional distribution of the genes containing SNPs distinguishing this group from the rest of the *M. tuberculosis* isolates included in this study is shown in Figure 4.3.4. A large proportion of the SNPs are in genes classified as "conserved hypothetical". Other functional categories that are highly represented include "cell wall and cell processes" and "intermediary metabolism and respiration".

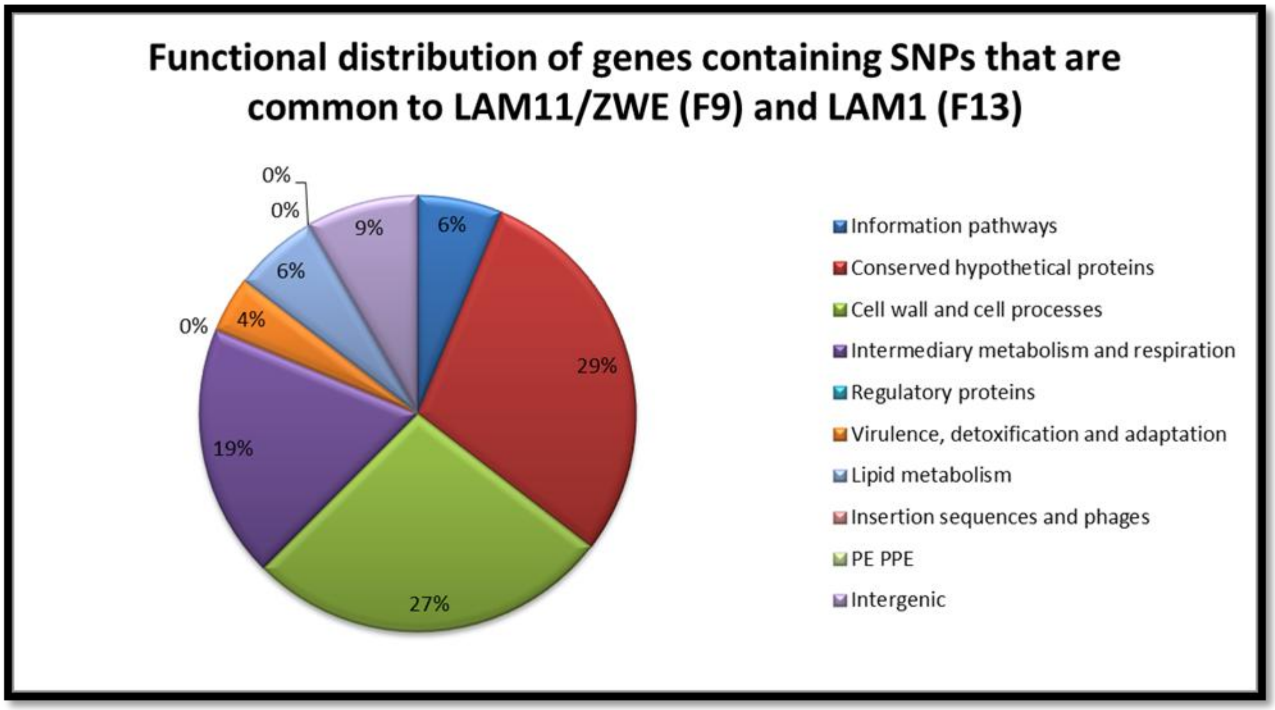


Figure 4.3.4. Functional distribution of genes containing SNPs common to LAM11/ZWE (F9) and LAM1 (F13) (n=6). All SNPs that are common to LAM11/ZWE (F9)- and LAM1 (F13) strains were functionally annotated. The SNPs listed are not found in any of the other LAM (LAM4 KZN (F15), T1, T1-Tusc, T5/Rus (F14), LAM U (F26), LAM3 (F11)) or non-LAM isolates included in the analysis.

SNPS UNIQUE TO LAM11/ZWE (F9)

A total number of 35 SNPs were uniquely identified in the LAM 11/ZWE isolates included in this analysis (see Table 4.3.2). These SNPs were not found in any of the other isolates included in this study, besides the LAM11/ZWE (F9) isolates. Interestingly, none of these were intergenic, 22 were non-synonymous and the remaining 13 were synonymous.

The functional distribution of genes harbouring SNPs that are common to all LAM 11/ZWE isolates included in this study is shown in Figure 4.3.5. The majority of the SNPs that define this clade are found in genes coding for proteins involved in the cell wall, cell processes, intermediary metabolism and respiration. Other highly represented functional categories include “lipid metabolism” and “PE/PPE” gene families.

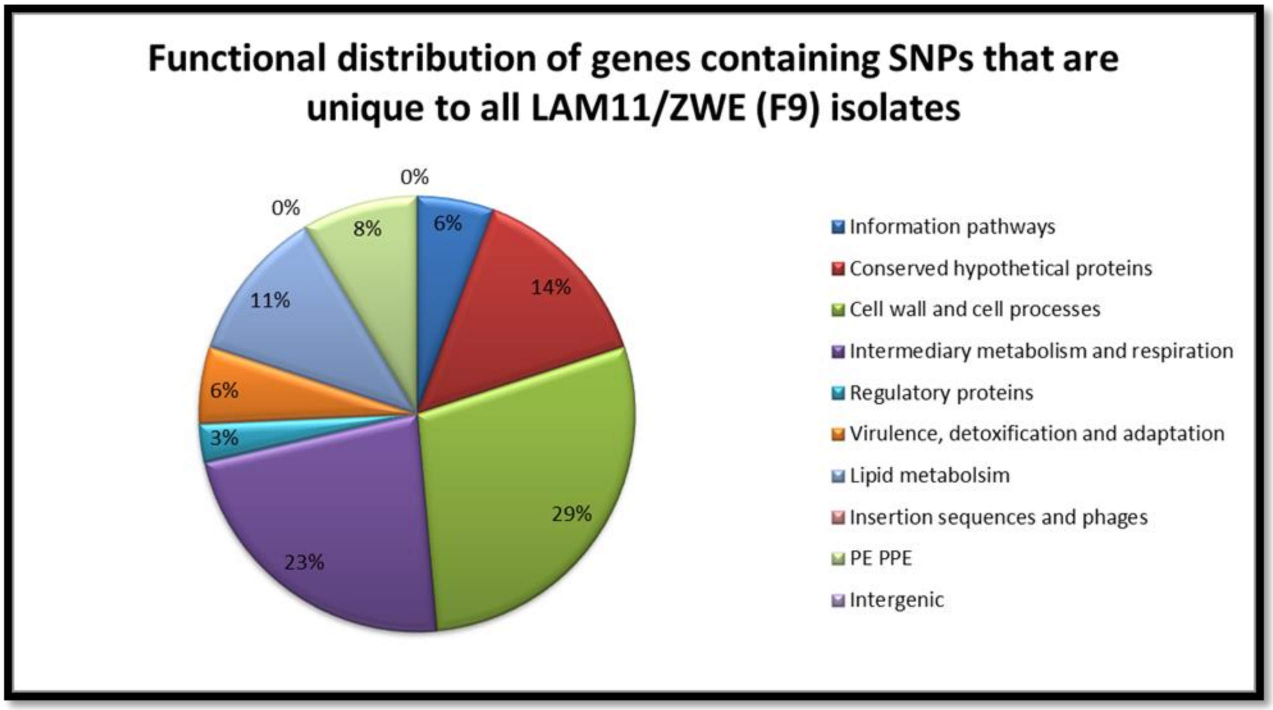


Figure 4.3.5. Functional distribution of genes containing SNPs common to LAM11/ZWE (F9) (n=3). All SNPs that are common to LAM11/ZWE (F9) strains were functionally annotated. The SNPs listed are not found in any of the other LAM (LAM4 KZN (F15), T1, T1-Tusc, T5/Rus (F14), LAM U (F26), LAM1 (F13), LAM3 (F11)) or non-LAM isolates included in the analysis.

SNPS UNIQUE TO LAM1 (F13)

A total number of 30 SNPs were identified to be unique to the LAM1 (F13) clade, of which six are intergenic, 13 cause amino acid substitutions and are thus non-synonymous, and 11 are synonymous (Table 4.3.2).

Figure 4.3.6 shows the functional distribution of the genes in which the SNPs uniquely identified in the LAM1 (F13) isolates occur. A large proportion of SNPs were identified in conserved hypothetical genes. Other changes occur mostly in genes coding for proteins involved in the cell wall, cell processes, intermediary metabolism and respiration.

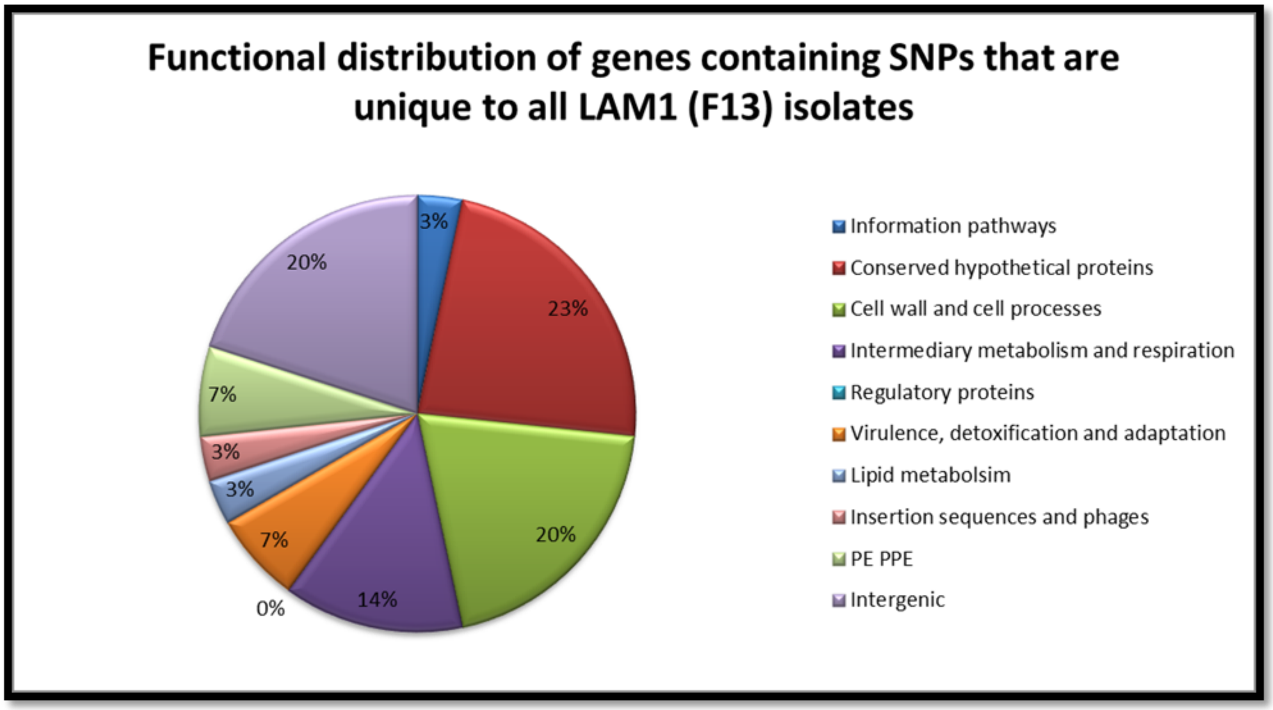


Figure 4.3.6. Functional distribution of genes containing SNPs common to LAM1 (F13) (n=3). All SNPs that are common to LAM1 (F13) strains were functionally annotated. The SNPs listed are not found in any of the other LAM (LAM4 KZN (F15), T1, T1-Tusc, T5/Rus (F14), LAM U (F26), LAM11/ZWE (F9), LAM3 (F11)) or non-LAM isolates included in the analysis.

SNPS UNIQUE TO LAM U (F26) T1, T1-TUSC, T5/RUS (F14), LAM4 KZN (F15)

Table 4.3.2 summarises the SNPs found in LAM U (F26) T1, T1-Tusc, T5/Rus (F14), LAM4 KZN (F15), but not in any of the other LAM isolates (F11, F9 or F13) or non-LAM isolates included in the analysis. Of the total number of 37 SNPs uniquely identified to this sub-group of LAM isolates, 22 are non-synonymous, 10 are synonymous and five occur in intergenic regions.

The functional distribution of the genes in which the 37 uniquely identified SNPs in this clade occur, are shown in Figure 4.3.7. The functional categories “cell wall and cell processes” and “intermediary metabolism and respiration” were equally represented, each containing 9 SNPs. Other highly represented functional groups include genes encoding for proteins involved in lipid metabolism and the “conserved hypothetical proteins”. The high frequency of non-synonymous changes seen in the genes encoding for proteins involved in the cell wall and cell processes, suggests a gradual change in the bacteria’s first barrier of defence against the host immune system. It is likely that, in response to this change, the cellular processes inherently linked to the pathways and systems involved in the cell wall and other cell processes such as the lipid metabolism, intermediary metabolism and general respiration need to compensate and adjust.

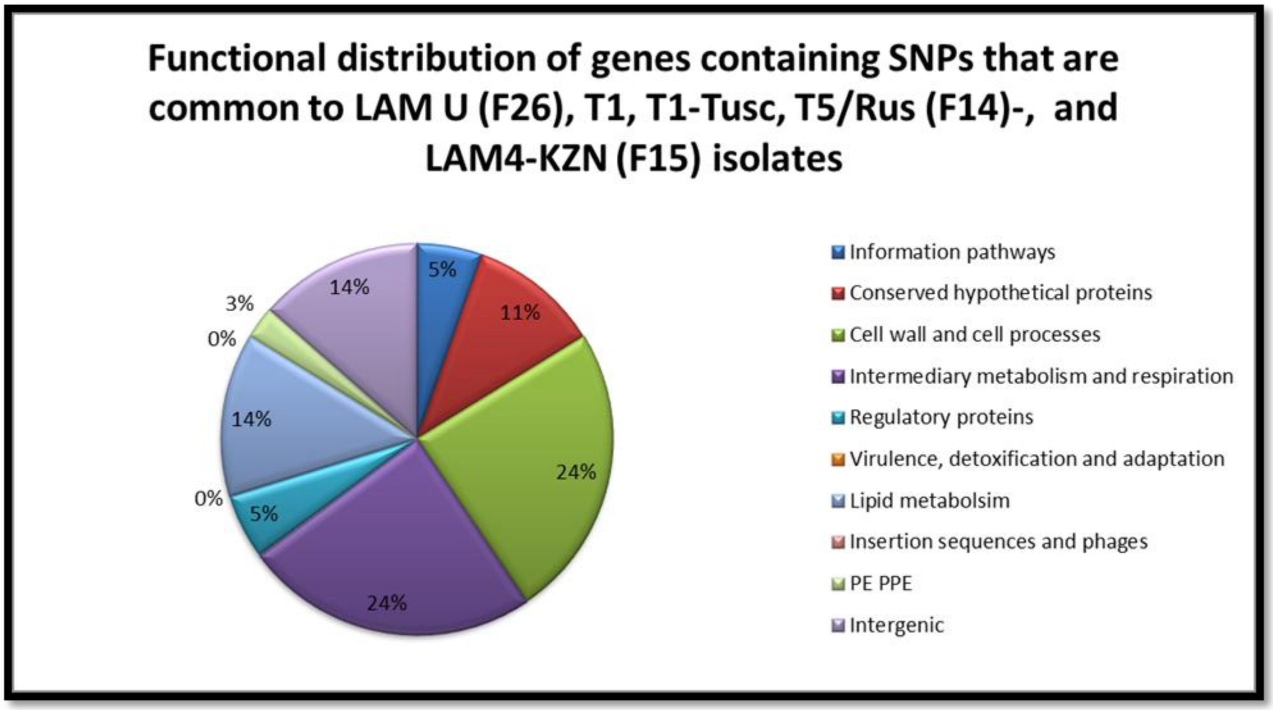


Figure 4.3.7. Functional distribution of genes containing SNPs common to LAM U (F26)-, T1, T1-Tusc, T5/Rus (F14)-, and LAM4 KZN (F15) (n=10). All SNPs that are common to LAM U (F26)-, T1, T1-Tusc, T5/Rus (F14)-, and LAM4 KZN (F15) strains were functionally annotated. The SNPs listed are not found in any of the other LAM (LAM11/ZWE (F9), LAM1 (F13), LAM3 (F11) or non-LAM isolates included in the analysis.

SNPS UNIQUE TO LAM U (F26)

The LAM U (F26) isolates included in this study were shown to have 119 unique SNPs in common, which were not found in any of the other LAM- or non-LAM isolates investigated here. Twelve of these SNPs are in non-coding regions, whilst 61 cause amino acid substitutions and 46 are synonymous (Table 4.3.2). LAM U (F26) isolates share a large number of SNPs that distinguishes them from the rest of the *M. tuberculosis* isolates included in this study.

Figure 4.3.8 shows the functional distribution of the uniquely identified SNPs, in isolates belonging to LAM U (F26), within genes in the form of a pie chart. Interestingly, the functional category; ‘intermediary metabolism and respiration” shows the most variation, with 28% of the SNPs present in genes involved in this functional group. Other functional categories that are highly represented are the diverse conserved hypothetical proteins and proteins involved in the cell wall and cell processes. This group of strains also shows variation in the PE/PPE family of genes. As mentioned previously, abundant variation in these genes has been reported in literature but the reliability of the variation identified with NGS technologies remains questionable. Interestingly, in contrast with the SNPs that distinguish other subgroups of the LAM genotype, only four SNPs are found in genes coding for proteins involved in the information pathways. Only one of these, namely *dinP*, is involved in the 3R system.

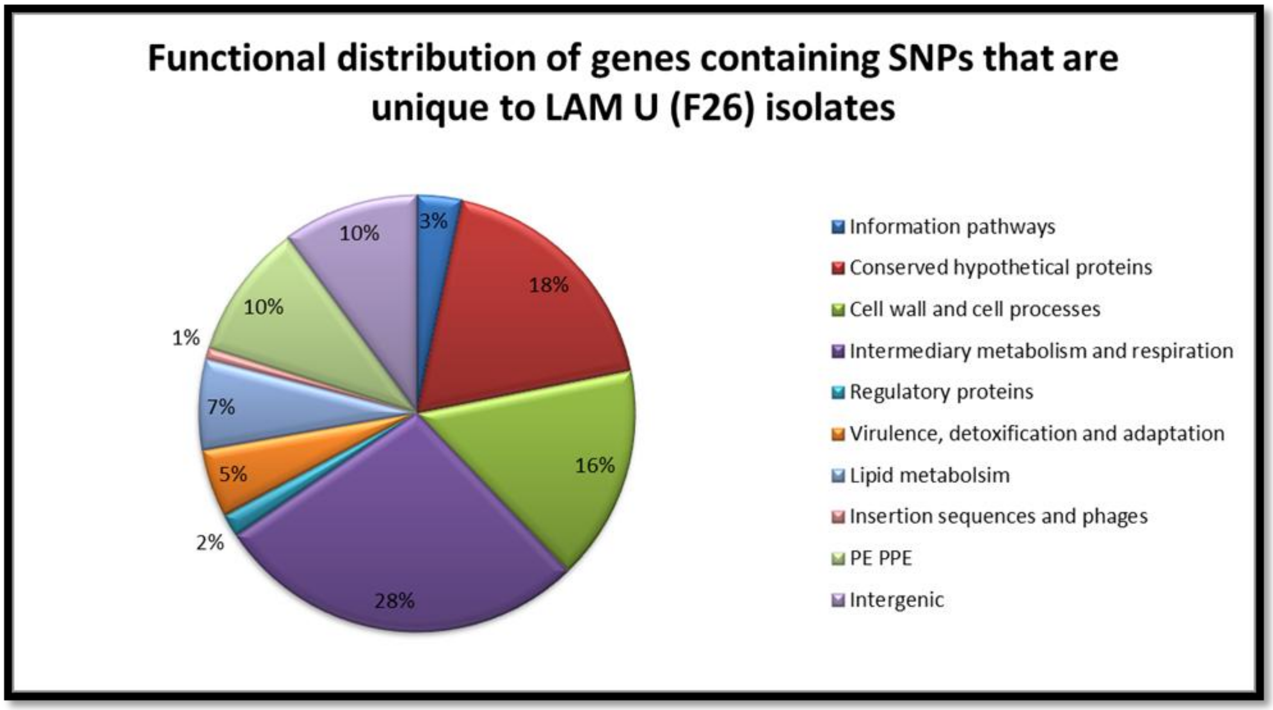


Figure 4.3.8. Functional distribution of genes containing SNPs common to LAM U (F26) (n=3). All SNPs that are common to LAM U (F26) strains were functionally annotated. The SNPs listed are not found in any of the other LAM (LAM4 KZN (F15), T1, T1-Tusc, T5/Rus (F14), LAM11/ZWE (F9), LAM1 (F13), LAM3 (F11)) or non-LAM isolates included in the analysis.

SNPS UNIQUE TO T1, T1-TUSC, T5/RUS (F14) AND LAM4 KZN (F15)

Twelve SNPs were uniquely identified in the clade consisting of T1, T1-Tusc, T5/Rus (F14) and LAM4 KZN (F15) strains, and were not found in any of the other LAM- or non-LAM isolates included in this study. Thirteen SNPs were unique to this clade and the functional categories of the genes in which these occur are shown in Figure 4.3.9. Of these unique SNPs, two occur in non-coding regions. Four SNPs are synonymous and the remaining seven cause non-synonymous changes. Three SNPs were found in genes coding for proteins involved in the lipid metabolism and two SNPs were found in genes coding for proteins in the functional category; “cell wall and cell processes”. One SNP was found in *Rv3263*, encoding a DNA methylase, which is one of the proteins involved in the 3R system.

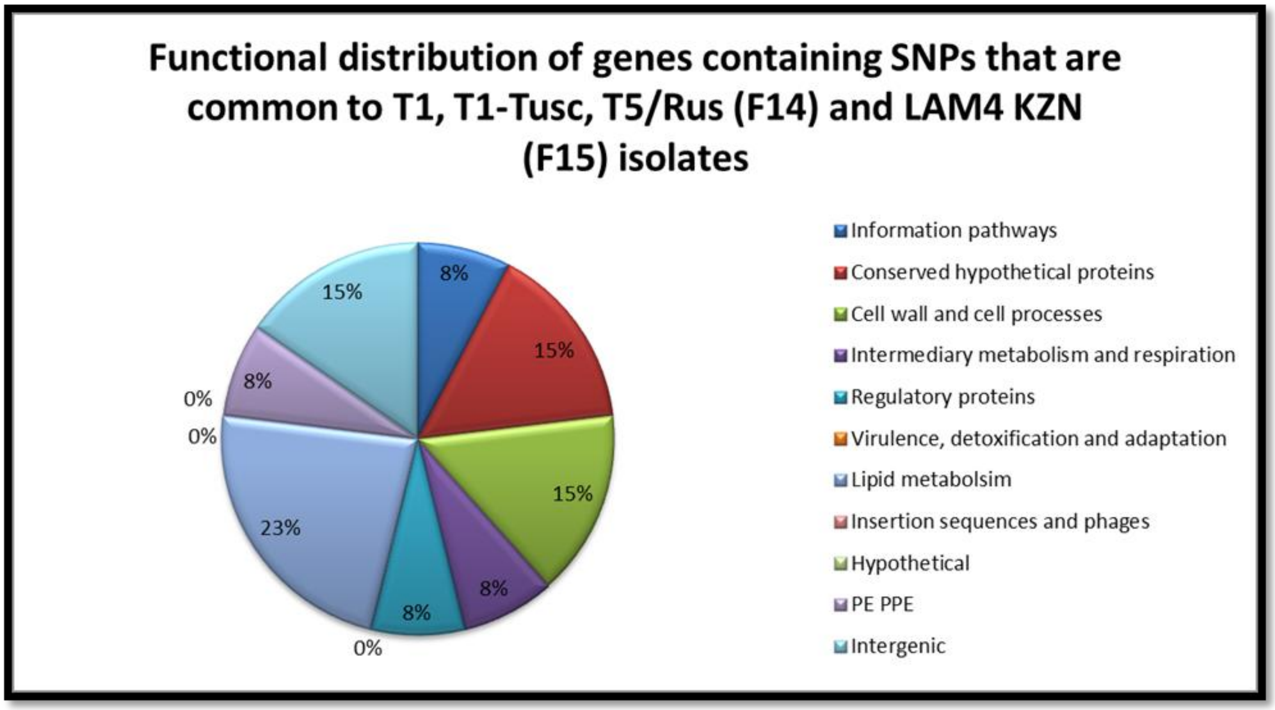


Figure 4.3.9. Functional distribution of genes containing SNPs common to T1, T1-Tusc, T5/Rus (F14) and LAM4 KZN (F15) (n=7). All SNPs that are common to T1, T1-Tusc, T5/Rus (F14)- and LAM4 KZN (F15) strains were functionally annotated. The SNPs listed are not found in any of the other, LAM U (F26), LAM11/ZWE (F9), LAM1 (F13), LAM3 (F11) or non-LAM isolates included in the analysis.

SNPS UNIQUE TO T1, T1-TUSC, T5/RUS (F14)

A total number of 90 SNPs were uniquely identified in the T1, T1-Tusc, T5/Rus (F14) isolates, of which seven occur in non-coding regions, 53 cause amino acid substitutions (non-synonymous), and 30 are synonymous (see Table 4.3.2). Again, the high frequency of non-synonymous SNPs points to the low influence that natural selection has on the evolution of *M. tuberculosis*

The graphical presentation of the functional distribution of the genes, in which the uniquely identified SNPs for this clade occur, is shown in Figure 4.3.10. A large proportion of the SNPs are in genes coding for proteins involved in the cell wall, other cell processes, intermediary metabolism and respiration. SNPs in conserved hypothetical genes make up 20% of the total number of SNPs. Four of the SNPs in genes involved in the information pathways occur in genes of the 3R system, three of which are in the hyper mutable genes; *polA*, *nei*, and *dnaQ*. This sub-lineage of the LAM genotype clearly underwent a large number of evolutionary changes that distinguishes it from the rest of the LAM isolates included in this study, as well as from the non-LAM isolates included.

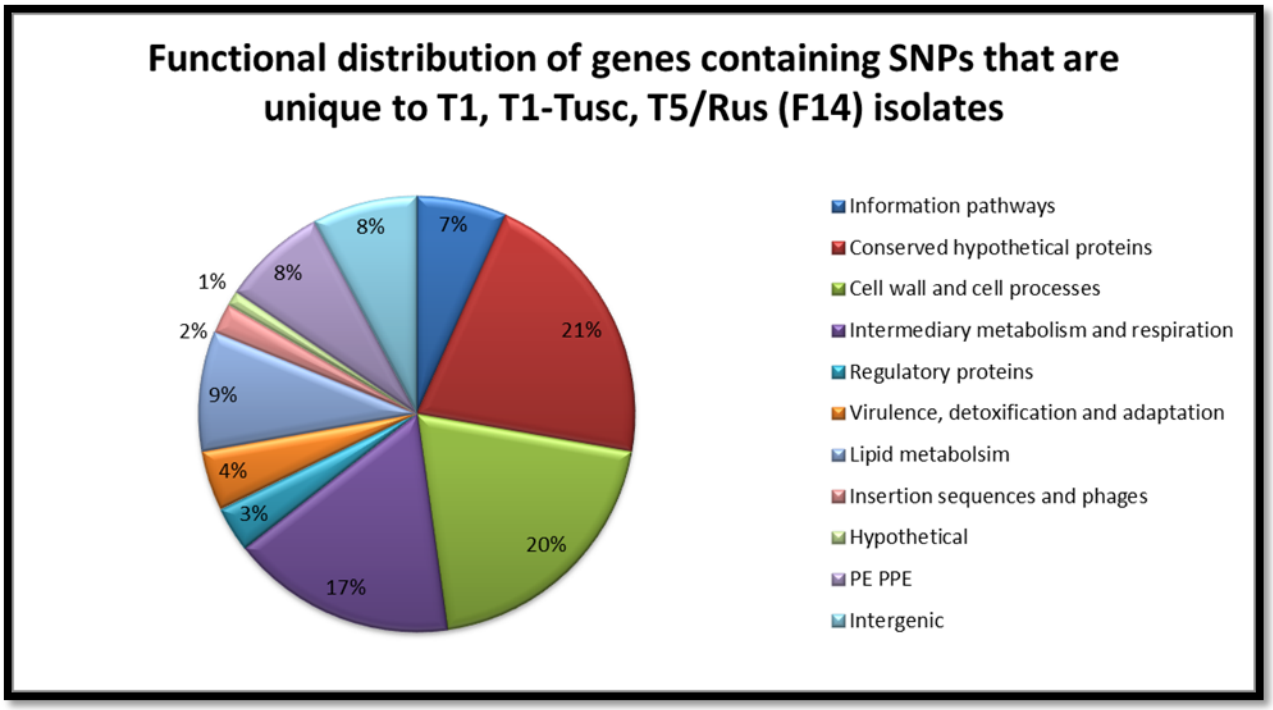


Figure 4.3.10. Functional distribution of genes containing SNPs common to T1, T1-Tusc, T5/Rus (F14) (n=3). All SNPs that are common to T1, T1-Tusc, T5/Rus (F14) strains were functionally annotated. The SNPs listed are not found in any of the other LAM (LAM4 KZN (F15), LAM U (F26), LAM11/ZWE (F13), LAM1 (F13), LAM3 (F11)) or non-LAM isolates included in the analysis.

SNPS UNIQUE TO LAM4 KZN (F15)

The LAM4 KZN isolates included in this analysis share 57 SNPs, 34 of which cause non-synonymous changes and 14 SNPs do not lead to amino acid changes. Of the total number of 57 SNPs uniquely identified in this clade, nine proved to be in intergenic regions (see Table 4.3.2).

The functional category distribution of the genes containing SNPs uniquely identified in all isolates of the LAM4 KZN clade is graphically presented in Figure 4.3.11. The functional categories that are highly represented include the genes coding for proteins involved in cell wall and cell processes and intermediary metabolism and respiration. Other highly represented categories are the intergenic SNPs and the SNPs located in conserved hypothetical genes. Conserved hypothetical proteins could potentially be involved in any of the mentioned functional categories, as their respective functions are not yet known. Furthermore, 10% of SNPs occur in genes coding for proteins involved in information pathways. Three of these genes are involved in the 3R system, one of which is the hyper-mutable *dnaQ*. The SNPs in the genes of the 3R system are non-synonymous and thus have the potential to alter protein function. Three SNPs unique to this clade are identified within the genes coding for the PE/PPE family of proteins. Studies have shown that PE and PPE genes have high sequence variation. However, the repetitive nature of these genes makes them problematic for NGS analysis and it is debatable whether or not to trust variants called by NGS analysis techniques in these areas, further verification is needed to make conclusions regarding all nucleotide variants identified within these repetitive sections of the genome. Future NGS technologies aspire to better resolve genomic regions with a highly repetitive nature.

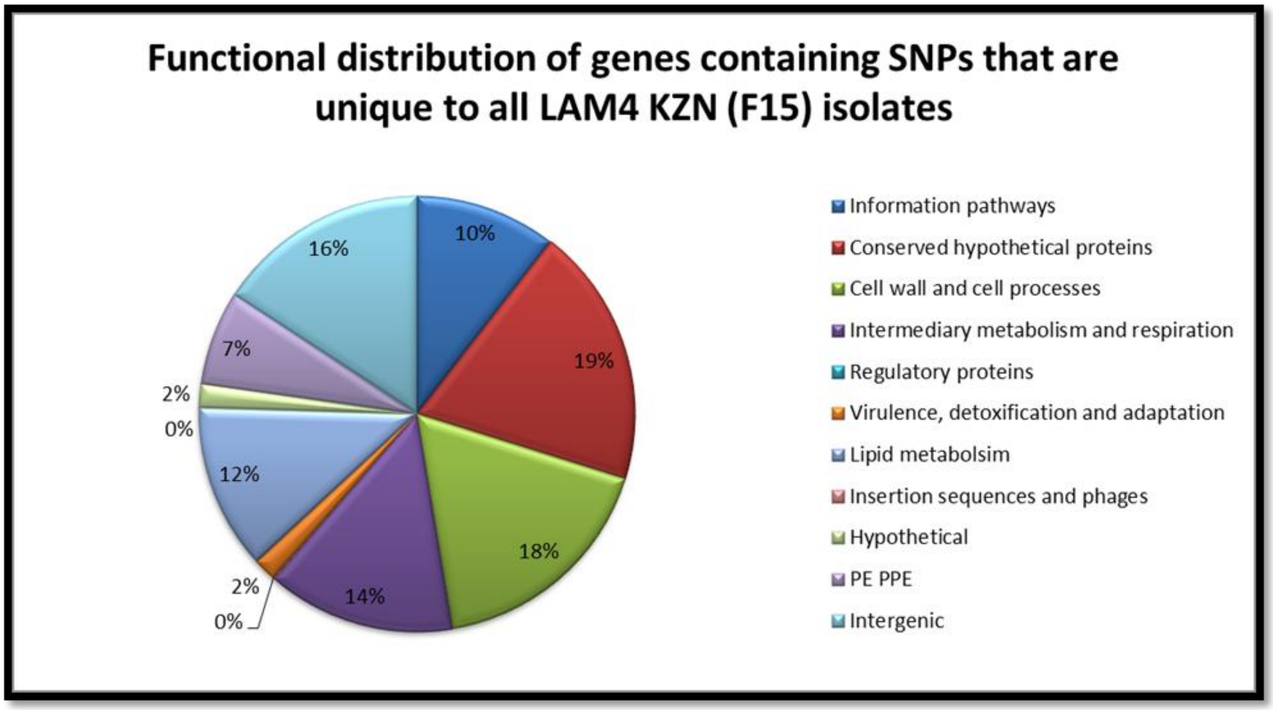


Figure 4.3.11. Functional distribution of genes containing SNPs common to LAM4 KZN (F15) (n=4). All SNPs that are common to LAM4 KZN (F15) strains were functionally annotated. The SNPs listed are not found in any of the other LAM (T1, T1-Tusc, T5/Rus (F14), LAM U (F26), LAM11/ZWE (F13), LAM1 (F13), LAM3 (F11)) or non-LAM isolates included in the analysis.

SNPS UNIQUE TO 17 OF 18 LAM ISOLATES ANALYSED

The SNPs uniquely identified in all the LAM isolates included in this study, excluding any one of the LAM4 KZN isolates included (n=4), are summarised in Table 4.3.3. The SNPs are thus present in 17 LAM isolates, but absent in any one isolate in clade consisting of LAM4 KZN isolates. A total number of 23 SNPs were identified in this group of strains consisting of 17 LAM isolates. Of these, six are in non-coding regions and 8 SNPs cause amino acid changes (non-synonymous), whilst 9 are synonymous. The fact that these single nucleotide changes are present in all but one of the LAM isolates included in the analysis may be due to sequencing errors or more likely due to erroneous mapping or SNP calling during the NGS data analysis pipeline. Alternatively, it could suggest that certain SNPs might have been fixed in the LAM family at one point in time but that LAM4 KZN isolates are systematically reverting some of the single nucleotide evolutionary events due to speculated selective pressure experienced by the bacteria within the host. A total number of 28 SNPs were identified that are present in 17 out of 18 LAM isolates included, 23 of these were present in all LAM isolates except one of 5 LAM4 KZN isolates (described above), whilst the other five SNPs occur in 17

LAM isolates except one random isolate, not forming a particular pattern of exclusion, and not limited to the LAM4 KZN (F15) clade. These were not included in Table 4.3.3.

Table 4.3.3. SNPs common to all isolates, excluding any one of 5 LAM4 KZN isolates

Functional category	Number of SNPs	Non-synonymous	Synonymous
Information pathways	1	0	1
Conserved hypothetical proteins	7	4	3
Cell wall and cell processes	6	3	3
Intermediary metabolism and respiration	2	1	0
Regulatory proteins	0	0	0
Virulence, detoxification and adaptation	1	0	1
Lipid metabolism	0	0	1
Insertion sequences and phages	0	0	0
PE/PPE	0	0	0
Intergenic	6	-	-
Total	*23*	*8*	*9*

4.3.3. *M. TUBERCULOSIS* LAM LSP ANALYSIS

In order to identify possible LSPs (more commonly known as RDs), the depth of coverage for all of the alignments for each of the LAM isolates analysed was considered and the alignments across the genomic regions that had zero coverage (no reads mapping to that region) were inspected visually. This technique is time consuming and it should be noted that this approach is dependent on the subjective interpretation of the alignment, the nature of the genomic region in which the LSP occurs (mapping in repetitive or high GC regions are often erroneous), and the coverage at which the target genome has been sequenced. Several previously reported LSPs were identified in subsets of the LAM isolates analysed. A number of these LSPs were confirmed by PCR in a small subset of LAM and non-LAM *M. tuberculosis* clinical isolates. PCR confirmation of a selected subset of the deletions provided a level of confidence to which LSPs can be called by merely considering NGS data and the appropriate analysis approach, as described in Chapter 3, Section 3.5.6.13. The presence or absence of known- and newly identified LSPs detected in analysed the *M. tuberculosis* LAM isolates are summarised in Table 4.3.4. Interestingly, RD149 has been identified in all of the LAM isolates, RD149 is however not LAM-specific as it has been identified in other lineages of *M. tuberculosis*, e.g. the Beijing genotype (Tsolaki *et al*., 2004).

Table 4.3.4. The presence or absence of LSPs in the LAM genotype isolates analysed

LAM family	Strain (SAWC)	RD115	RD149	RD150	RD152	RD174	RD^{Rio}	RD761	RD plcD*	RD Rv0145*	RD Rv2277*
F13	2511	0	1	0	0	1	1	0	1	0	0
	2904	0	1	0	0	1	1	0	1	0	0
	5276	0	1	0	0	1	1	0	1	0	0
F9	LAM1503	0	1	0	1	1	1	0	0	0	0
	1955	0	1	0	1	1	1	0	0	0	0
	3517	0	1	0	1	1	1	0	0	0	0
F26	2262	1	1	1	0	0	0	0	0	0	1
	3656	1	1	1	0	0	0	0	0	0	1
	5260	1	1	1	0	0	0	0	0	0	1
F14	3100	1	1	0	0	0	0	0	0	1	0
	3388	1	1	0	0	0	0	0	0	1	0
	3651	1	1	0	0	0	0	0	0	1	0
F15	KZN605	1	1	0	0	0	0	0	0	0	0
	KZN1435	1	1	0	0	0	0	0	0	0	0
	KZN4207	1	1	0	0	0	0	0	0	0	0
	2576	1	1	0	0	0	0	0	0	0	0
F11	4498	0	1	0	0	0	0	1	0	0	0
	5165	0	1	0	0	0	0	1	0	0	0

A zero (0) and a red coloured cell in the field of a specific LSP indicate the absence of that LSP in the corresponding isolate, the genomic region is thus present. A one (1) and a green coloured cell in the field of a specific LSP indicate the presence of that LSP in the corresponding isolate, the region is thus deleted. An RD description marked with an asterisk indicates that the particular RD was identified in the present study, Other RDs were previously reported.

4.3.4. SUMMARY OF THE *M. TUBERCULOSIS* LAM GENOTYPE

A previous study has identified specific IS*6110* insertion sites in members of PGG2, as well as the loss of direct variable repeat (DVR) spacer regions across lineages of PGG2 (Streicher, 2007). Figure 4.3.11 summarises these lineage defining events and combines it with the findings of the present study. The clades, as defined in Section 4.3.1, are indicated with the corresponding IS*6110* RFLP families. The SNPs defining the different sub-lineages of the LAM genotype can thus be extrapolated to the lineages established in Figure 4.3.12. The presence of the identified large sequence polymorphisms in the form of RDs are shown on the relationship plot.

The LAM lineage was found to consist of 6 strain families (F9, F11, F13, F14, F15, F26), that share common IS*6110* insertions in *Rv1754, Rv2352, Rv0835:Rv0836* (Streicher, 2007). According to the international designation, F11, F13, F15 and F9 correspond to LAM3, LAM1, LAM4 and LAM11/ZWE, respectively, based on their characteristic spoligotype patterns and share the characteristic deletion of DVR 21-24 on their spoligotype profile (Brudey *et al*., 2006; Filliol *et al*., 2002).

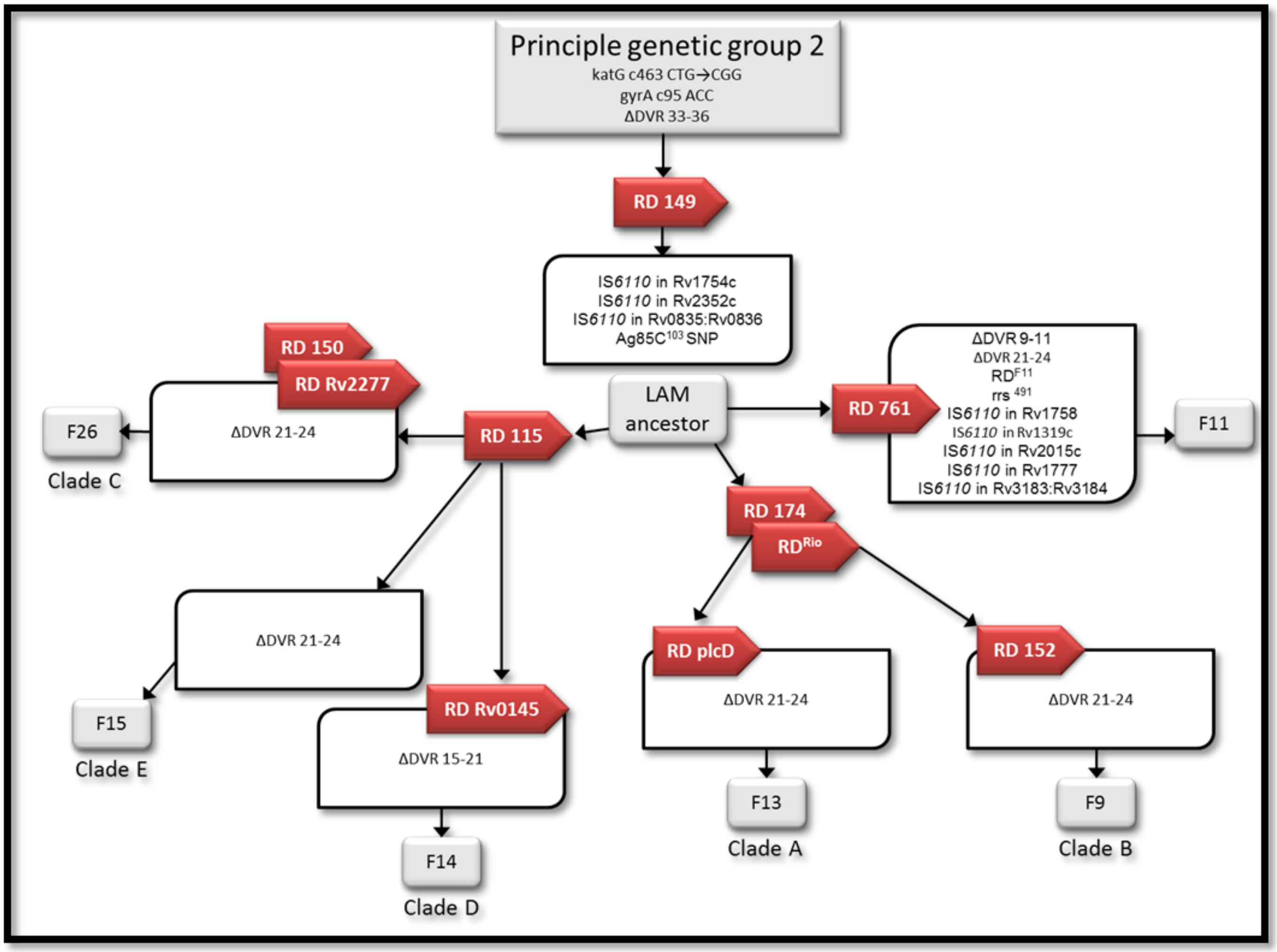


Figure 4.3.12. Relationship plot of the *M. tuberculosis* LAM genotype. LSPs (RDs) are indicated as red arrow blocks on the branches of the RFLP families in which they are present. *IS6110* insertion sites, specific SNPs, DVR regions lost, and the clades (defined in Section 3) to which each *IS6110* RFLP family corresponds are indicated.

4.3.5. NON-LAM EVOLUTION

The SNPs identified in all the non-LAM *M. tuberculosis* isolates included in this analysis (thus excluding *M. canetti* and *M. africanum* isolates) were annotated and the genes in which the SNPs occur were assigned to functional categories (Figure 4.3.13). Thirteen percent of the SNPs were in intergenic regions. The functional categories most highly represented include "intermediary metabolism and respiration" and "cell wall and cell processes". A large number of SNPs were found in conserved hypothetical genes, making it challenging to infer the physiological consequences thereof. Other functional categories represented, in descending order of representation include "lipid metabolism", "PE/PPE protein families", "information pathways", "regulatory proteins", "virulence, detoxification and adaptation", and "insertion sequences and phages". This pattern of evolution is similar to that seen in the LAM isolates analysed in Section 4.3.2. It can thus be said that the currently circulating strains of different lineages of *M. tuberculosis* are evolving in the same direction. The bacteria are making adjustments to the cell wall, which is the outer barrier that is in direct contact with the host and host-immune response. Lipid metabolism is inherently linked with the cell wall and other cell processes, hence changes in this functional category are expected to accompany cell wall changes. It is anticipated that respiration processes and intermediary metabolism need to adapt in response to the changes taking place in the cell wall. As information pathways underlie all cellular processes, they also need to adjust to support the other changes that the bacteria undergo. It is however unclear whether the adjustments in the cell wall and related processes mediate the rest of the evolutionary changes that are observed, or vice versa. The unidirectional evolution shown in this study strongly supports the hypothesis of host-pathogen co-evolution (Gagneux, 2012; Gagneux *et al.*, 2006).

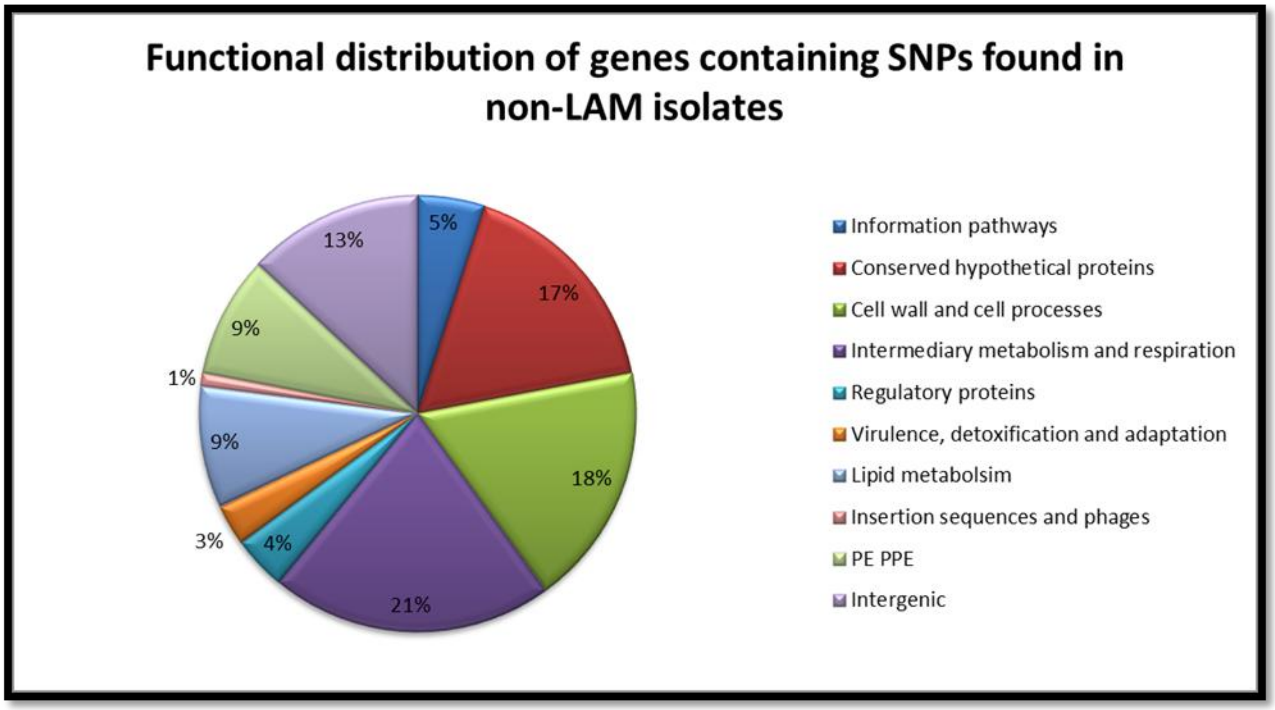


Figure 4.3.13. Functional distribution of genes containing SNPs in non-LAM isolates. All SNPs that were identified in the analysed non-LAM *M. tuberculosis* isolates were functionally annotated.

4.3.6. SUB-CONCLUSIONS

The SNP variation summarised in this section (4.3) indicates that the genes involved in the cell wall, cell processes, intermediary metabolism and respiration of the LAM genotype are primarily influenced during evolution. A similar distribution of SNP variation was observed when the non-redundant SNPs of the non-LAM isolates were analysed and functionally annotated. A large proportion of the SNP variants cause non-synonymous changes, which is consistent with what was previously observed in *M. tuberculosis* (Baker *et al*., 2004). A set of 55 SNPs were identified to be unique to the LAM isolates included in this analysis. Different clades within the LAM lineage are defined by subsets of SNPs and RDs that are responsible for the topology of the phylogenetic tree presented in Section 4.2.

4.4. RD^{RIO} VS NON-RD^{RIO}

4.4.1. INTRODUCTION

In order to investigate the functional changes that occurred during the evolution of the RD^{Rio} vs. non-RD^{Rio} LAM strains in more detail, two of the LAM isolates included in the phylogenetic and genomic analysis were selected. These isolates were cultured, whole cell lysate proteins were extracted from these cultures, and the proteins were analysed by mass spectrometry (as described in Chapter 3, Section 3.8 – 3.14). These clinical isolates (SAWC 3517 and SAWC 3651) were genotyped using the internationally standardised methods of spoligotyping and *IS6110* RFLP, (Table 4.4.1), and were shown to harbour all of the LAM specific genetic markers. SAWC 3517 is a member of *IS6110* family 9 and is characterised by ΔDVR 21-24, 27-30, 33-36. In contrast, SAWC 3651 presents with an *IS6110* RFLP pattern characteristic of F14 and ΔDVR 15-23 and 33-36 are deleted.

Table 4.4.1. Clinical isolates used for proteomic analysis

***IS6110* family**	**SAWC number**	**Common name**	**Spoligotype**
9	3517	LAM 11/ZWE	■■■■■■■■■■■■■■■■■■■■□□□□■■□□□□■■□□□□■■■■■■■
14	3651	T1, T1-Tusc, T5/Rus	■■■■■■■■■■■■■■□□□□□□□□□■■■■■■■■■□□□□■■■■■■■

The phylogenetic positions of SAWC 3517 and SAWC 3651 are indicated on the cladogram (Figure 4.4.1). Isolate SAWC 3517 harbours the 35 SNPs that were found to be unique to the LAM11/ZWE (F9) strians, while isolate SAWC 3651 has 90 SNPs uniquely associated with the T1, T1-Tusc, T5/Rus (F14) strains (Section 4.3).

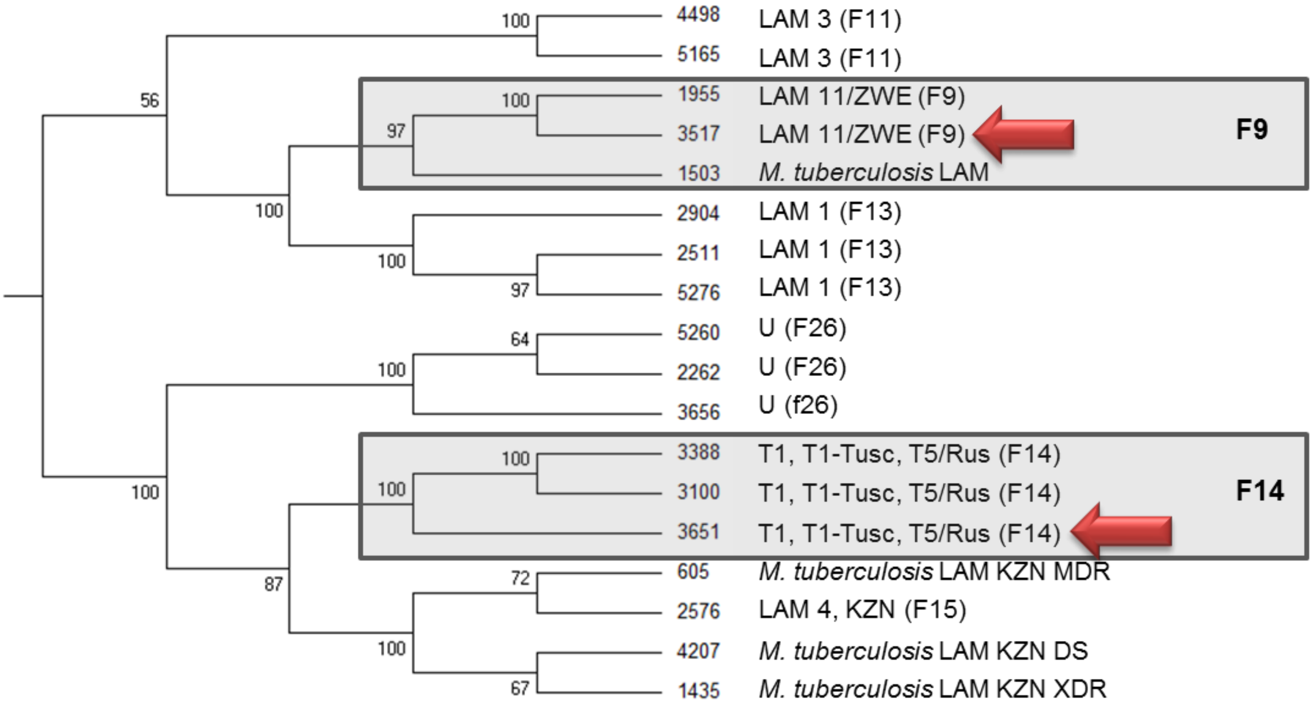


Figure 4.4.1. ***M. tuberculosis*** **LAM genotype phylogeny, showing the position of SAWC 3517 and SAWC 3651**

Isolates SAWC 3517 and SAWC 3651 possess additional genomic features that distinguish them from each other and other lineages of *M. tuberculosis*. The large genomic deletions present in the *M. tuberculosis* LAM genotype isolates analysed are summarised in Table 4.3.8. Both SAWC 3517 and SAWC 3651 have a number of genomic regions deleted relative to the *M. tuberculosis* H37Rv genome sequence. Both isolates have RD149 deleted, RD174, RD^{Rio}, and RD152 are deleted in SAWC 3517, while RD115 and a newly identified gene deletion (*Rv0145*) are deleted in SAWC 3651. A summary of the genes deleted in each of the RDs deleted in SAWC 3517 and SAWC 3651 are shown in Table 4.4.2.

Table 4.4.2. Description of genes in RDs present in SAWC 3517 and/or 3651

RD	Size (bp)	Genes involved	Gene name	Gene description	Functional category	Deleted in SAWC 3517/3651
RD in pks15/1	7	*Rv2946c*	*pks1*	polyketide synthase	Lipid metabolism	3517/3651
		Rv2947c	*pks15*	polyketide synthase	Lipid metabolism	
RD^{Rio}	26 314	*Rv3345*	*PE_PGRS49*	PE-PGRS family protein	PE/PPE families	3517
		Rv3346	*PE_PGRS50*	PE-PGRS family protein	PE/PPE families	
		Rv3347		conserved hypothetical protein	Cell wall and cell processes	
		Rv3348		transposase	Insertion sequences and phages	
		Rv3349		transposase	Insertion sequences and phages	
		Rv3350		transposase	Insertion sequences and phages	
		Rv3351	*PPE56*	PPE family protein	PE/PPE families	
		Rv3352		conserved hypothetical protein	Conserved hypothetical	
		Rv3353		oxidoreductase	Intermediary metabolism and respiration	
		Rv3354		conserved hypothetical protein	Conserved hypothetical	
		Rv3355c		conserved hypothetical protein	Cell wall and cell processes	
RD174	3650	*Rv1992c*	*ctpG*	metal cation transporting P-type ATPase	Cell wall and cell processes	3517
		Rv1993	*ctpG*	metal cation transporting P-type ATPase	Cell wall and cell processes	
		Rv1994		conserved hypothetical protein	Conserved hypothetical	
		Rv1995		hypothetical protein	Conserved hypothetical	
		Rv1996		conserved hypothetical protein	Virulence, detoxification and adaptation	
		Rv1997	*ctpF*	metal cation transporting P-type ATPase	Cell wall and cell processes	
RD115	2607	*Rv0376c*		conserved hypothetical protein	Conserved hypothetical	3651
		Rv0377		transcriptional regulator, lysR-family	Regulatory	
		Rv0378		conserved glycine rich protein	Conserved hypothetical	
RD149	9248	*Rv1572c*		conserved hypothetical protein	Insertion sequences and phages	3517/3651
		Rv1573		phiRv1 phage protein	Insertion sequences and phages	
		Rv1574		phiRv1 phage protein	Insertion sequences and phages	
		Rv1575		phiRv1 phage protein	Insertion sequences and phages	
		Rv1576		phiRv1 phage protein	Insertion sequences and phages	
		Rv1577		phiRv1 phage protein	Insertion sequences and phages	
		Rv1578c		phiRv1 phage protein	Insertion sequences and	

Table 4.4.2. Description of genes in RDs present in SAWC 3517 and/or 3651 continues

					phages	
RD	Size (bp)	Genes involved	Gene name	Gene description	Functional category	Deleted in SAWC 3517/3651
RD0145	BND	*Rv0145*		Possible adenosylmethionine dependant methyltransferase	Lipid metabolism	Rv3651

BND = the exact boundaries not determined

4.4.2. GENOMIC ANALYSIS

An extensive genomic comparison of SAWC 3517 and SAWC 3651 was done to exploit other genetic features that may underlie the phenotype of each strain. Single nucleotide variants, small genomic in/dels and regions of difference (large deletions) were identified in SAWC 3517 and SAWC 3651 with the NGS data analysis pipeline, described in Chapter 3, Section 3.5.6, and is presented in this section.

The SNPs identified in SAWC 3517 and SAWC 3651 are summarised in Table 4.4.3. In total, 819 high confidence SNPs were identified in SAWC 3517, relative to the reference genome *M. tuberculosis* H37Rv. Of these, 282 were synonymous, 446 were non-synonymous and 91 were intergenic. Similarly, 831 high confidence SNPs were identified in SAWC 3651 with respect to the reference genome. Of the total, 282 SNPs were synonymous, 457 were non-synonymous and 92 were intergenic. These SNPs were annotated and functional categories were assigned to the genes in which the SNPs occur. The functional category distribution for the genes in which SNPs found in SAWC 3517 and SAWC 3651 occur, are also shown in Table 4.4.3.

Table 4.4.3. Summary of SNPs identified in SAWC 3517 and SAWC 3651

	3517			3651		
Functional group	**All SNPs**	**Syn SNPs**	**Non-syn SNPs**	**All SNPs**	**Syn SNPs**	**Non-syn SNPs**
Information pathways	50	23	27	52	18	34
Conserved hypothetical proteins	149	58	91	145	52	93
Cell wall and cell processes	159	52	107	155	56	99
Intermediary metabolism and respiration	134	49	85	139	54	85
Regulatory proteins	29	13	16	33	15	18
Virulence, detoxification and adaptation	30	17	13	27	15	12
Lipid metabolism	77	37	40	83	37	46
PE/PPE protein families	90	32	58	91	34	57
Insertion sequences and phages	10	1	9	14	1	13
Hypothetical	1	0	1	2	0	2
Intergenic	91	-	-	92	-	-
Total	819	282	446	831	282	457

The SNPs and small in/dels identified in SAWC 3517 and SAWC 3651 were then compared to each other. These comparisons are shown in the Venn diagrams is Figure 4.4.2.A and Figure 4.4.2.B, respectively. SAWC 3517 and SAWC 3651 share 572 SNPs and 51 small in/dels with respect to *M.*

tuberculosis H37Rv. SAWC 3517 has 247 SNPs that were not identified in SAWC 3651, and SAWC 3651 has 259 SNPs that were not found in SAWC 3517. Fourteen distinct small in/dels were shown to be unique to each strain analysed in this section. Of the total number of SNPs not shared between the two isolates, 84 SNPs were identified in SAWC 3517 that were not identified in any other isolate included in the phylogeny analysis described in Section 4.2, whilst 48 SNPs were unique to SAWC 3651.

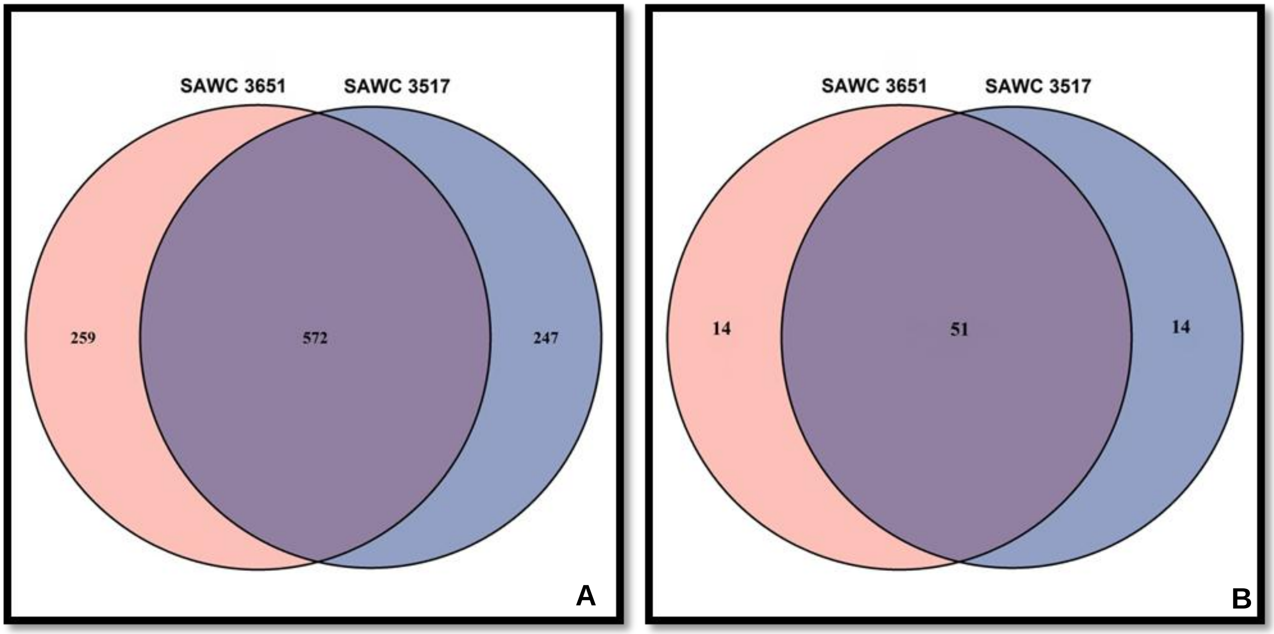


Figure 4.4.2. Overlapping variants between SAWC 4498 and SAWC 5165. A. SNPs that overlap in position and base identity between, or unique to, SAWC 3517 and SAWC 3651, employing *M. tuberculosis* H37Rv as a reference genome. B. In/dels that overlap in length, start- and end position and base identity between, or unique to, SAWC 3517 and SAWC 3651, employing *M. tuberculosis* H37Rv as a reference genome.

The SNPs identified in only one of the two isolates under investigation in this section, and SNPs identified in both of the isolates were functionally annotated. The SNPs (247 in SAWC 3517 and 259 in SAWC 3651) occurred in genes involved in the cell wall and cell processes and intermediary metabolism and respiration. A large number of the SNPs occurred in conserved hypothetical genes. The functional distribution for non-overlapping SNPs is shown in Figure 4.4.3.A and –B, respectively for SAWC 3517 and SAWC 3651. The SNPs that overlap between SAWC 3517 and SAWC 3651 are shown in Figure 4.4.3.C, and the functional distribution is similar to that of the non-overlapping/unique SNPs.

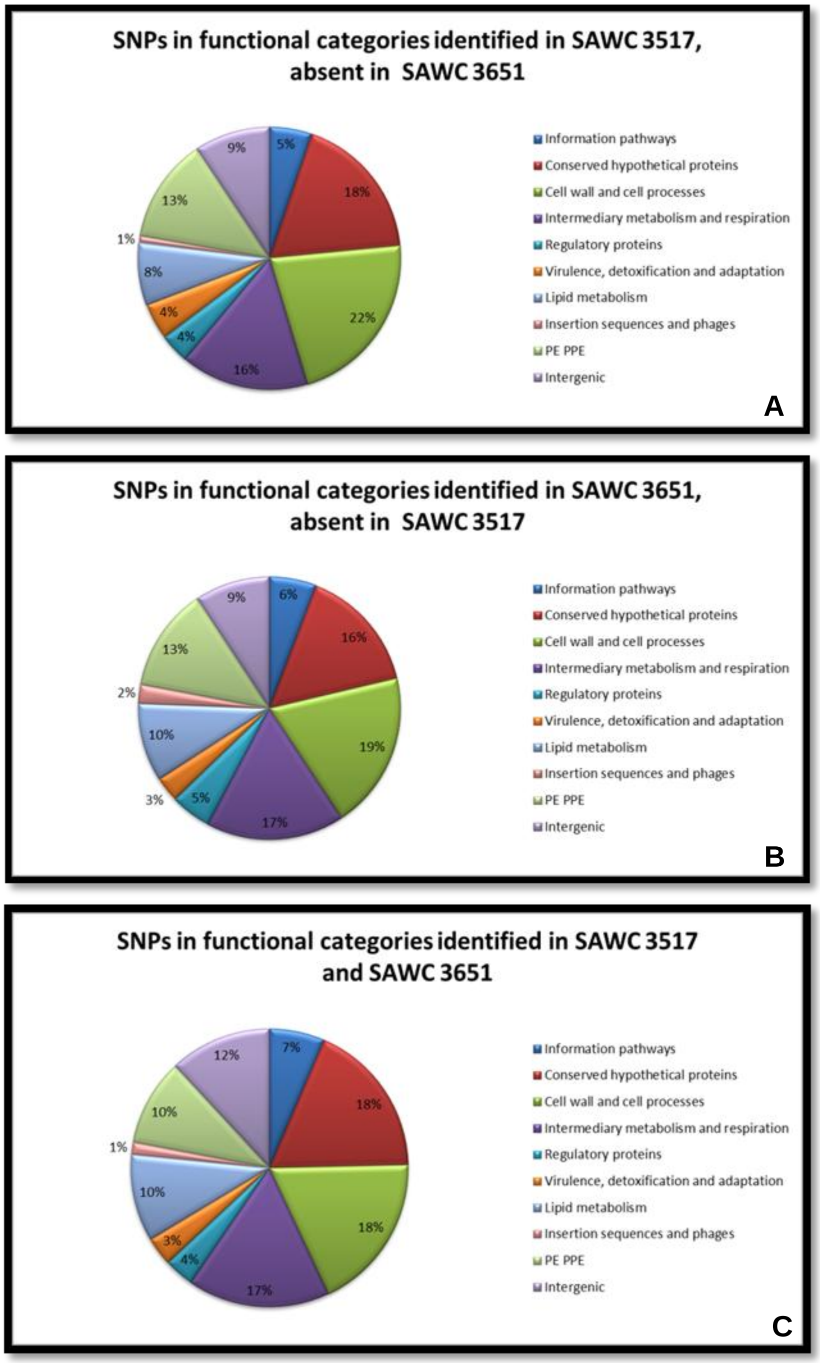


Figure 4.4.3. Functional distribution of genes containing SNPs identified in SAWC 3517 or SAWC 3651. A. SNPs identified only in SAWC 3517, and not in SAWC 3651 were functionally annotated. B. SNPs identified only in SAWC 3651 and not in SAWC 3517 were functionally annotated. C. Functionally annotated SNPs identified in SAWC 3517 and SAWC 3651.

The SNPs identified in the two isolates under investigation in this section are summarised in Table 4.4.4. SNPs identified in only one of either SAWC 3517 or SAWC 3651 and SNPs that were identified in both isolates were functionally annotated, and the functional distribution of the SNPs is shown in the table.

Table 4.4.4. SNPs identified in SAWC 3517 and SAWC 3651

	3517			3517/3651			3651		
Functional category	**Number of SNPs**	**Non-syn**	**Syn**	**Number of SNPs**	**Non-syn**	**Syn**	**Number of SNPs**	**Non-syn**	**Syn**
Information pathways	13	3	10	37	1	2	15	11	4
Conserved hypothetical proteins	45	29	16	104	9	5	41	30	11
Cell wall and cell processes	54	37	17	105	8	5	50	29	21
Intermediary metabolism and respiration	39	25	14	95	5	4	44	25	19
Regulatory proteins	9	5	4	20	0	0	13	7	6
Virulence, detoxification and adaptation	11	4	7	19	0	2	8	3	5
Lipid metabolsim	19	11	8	58	1	2	25	17	8
PE/PPE protein families	32	23	9	8	0	0	33	22	11
Insertion sequences and phages	2	1	1	58	0	0	6	5	1
Intergenic	23	-	-	68	-	-	24	-	-
Total	**247**	**138**	**86**	**572**	**24**	**20**	**259**	**149**	**86**

Non-syn= non-synonymous, Syn=synonymous

Small in/dels that were identified in SAWC 3517 but that were not found in SAWC 3651 are listed in Table 4.4.5. Small deletions and insertions can have detrimental effects on transcription and translation of the genes that they occur in, leading to possibly dysfunctional proteins. In/dels of which the sizes are not multiples of three have particularly negative effects on the gene products, as the reading frame of the gene is altered. The in/dels identified only in SAWC 3517 range from one to 13 base pairs and are in genes of varying function, or in intergenic regions. Three of the in/dels identified in SAWC 3517 and not in SAWC 3651 were located in PE/PPE gene families.

Table 4.4.5. Small in/dels identified in SAWC 3517, not identified in SAWC 3651

	Gene	Gene name	Gene description	Functional category	Size (bp)	Insertion/Deletion	Position in gene (bp)	Stop position (AA)	New stop position (AA)
SAWC 3517	*Rv0050*	*ponA1*	Bi-functional penicillin-binding protein	Cell wall and cell processes	3	Insertion	1891	679	680
	Intergenic	-	Us. of *Rv0302*	-	1	Deletion	-	-	-
	Rv0305c	*PPE6*	PPE family protein	PE/PPE families	1	Deletion	2430	964	0
	Rv0425c	*ctpH*	Metal cation transporting P-type ATPase	Cell wall and cell processes	12	Deletion	4491	1540	1536
	Intergenic	-	Us. of *Rv1028A, Rv1027c*	-	13	Deletion	-	-	-
	Rv1034c	-	Transposase	Insertion sequences and phages	4	Deletion	147	130	53
	Rv1291c	-	Conserved secreted protein	Cell wall and cell processes	1	Insertion	138	112	98
	Rv1446c	*opcA*	oxpp cycle protein	Intermediary metabolism and respiration	1	Insertion	19	304	58
	Rv1450c	*PE_PGRS27*	PE-PGRS family protein	PE/PPE families	1	Deletion	2279	1330	867
	Rv2352c	*PPE38*	PPE family protein	PE/PPE families	6	Deletion	482	392	390
	Rv2476c	*gdh*	NAD-dependent glutamate dehydrogenase	Intermediary metabolism and respiration	1	Deletion	2885	1625	966
	Rv2715		Hydrolase	Intermediary metabolism and respiration	1	Insertion	4	342	48
	Intergenic	-	Us. of *Rv3474, Rv3473c*	-	3	Deletion	-	-	-
	Intergenic	(position3964463)	-	-	1	Deletion	-	-	-

US= upstream, AA= amino acid

In/dels that were identified in SAWC 3651 and not found in SAWC 3517 are summarised in Table 4.4.6. Fourteen in/dels were found to be unique to SAWC 3651, and range in size from one to 15 base pairs. Three of the in/dels were in intergenic regions whilst six were in PE/PPE family genes. As discussed previously, the resolution of NGS platforms in repetitive regions, such as PE/PPE genes, is questionable. However, the in/dels presented in Table 4.4.4 were identified by GATK in three different alignments with independent mapping algorithms, that provides a high confidence for the presence of these variants.

Table 4.4.6. Small in/dels identified in SAWC 3651, not identified in SAWC 3517

	Gene	**Gene name**	**Gene description**	**Functional category**	**Size (bp)**	**Insertion/Deletion**	**Position in gene (bp)**	**Stop position (AA)**	**New stop position (AA)**
SAWC 3651	*Rv0032*	*bioF2*	8-amino-7-oxononanoate synthase	Intermediary metabolism and respiration	1	Insertion	2176	772	749
	Rv0124	*PE_PGRS2*	PE-PGRS family protein	PE/PPE families	1	Insertion	591	488	368
	Rv0304c	*PPE5*	PPE family protein	PE/PPE families	15	Deletion	825	2205	2200
	Intergenic	*(position 582458)*	-	-	8	Deletion	-	-	-
	Rv0604	*lpqO*	Lipoprotein	Cell wall and cell processes	2	Insertion	758	317	256
	Intergenic	-	Us. of *Rv0785*	-	1	Insertion	-	-	-
	Rv0878c	*PPE13*	PPE family protein	PE/PPE families	1	Insertion	1307	444	0
	Intergenic	-	Us. of *Rv0914c*	-	1	Insertion	-	-	-
	Rv1396c	*PE_PGRS25*	PE-PGRS family protein	PE/PPE families	1	Deletion	1178	577	415
	Rv1796	*mycP5*	Proline rich membrane-anchored mycosin	Intermediary metabolism and respiration	3	Deletion	523	586	585
	Rv2353c	*PPE39*	PPE family protein	PE/PPE families	1	Deletion	168	355	98
	Rv2396	*PE_PGRS41*	PE-PGRS family protein	PE/PPE families	1	Insertion	723	362	253
	Rv2415c	-	Conserved hypothetical protein	Conserved hypothetical	3	Insertion	709	298	299
	Rv3337	-	Conserved hypothetical protein	Conserved hypothetical	1	Insertion	255	129	0

US= upstream, AA= amino acid

4.4.2. PROTEOMIC ANALYSIS

The clinical isolates representative of RD^{Rio} (SAWC 3517) and non-RD^{Rio} (SAWC 3651) LAM lineages were cultured and whole cell lysate proteins were extracted when the cultures reached mid-log growth phase. *M. tuberculosis* H37Rv was cultured in parallel with the clinical isolates and whole cell lysate proteins were also extracted from this reference strain. Technical and biological replicates showed good correlation and variation between replicates was minimal, heatmaps of protein profiles are shown in Appendix F. Quantitative mass spectrometry was done to identify the relative abundance of proteins present in these three strains. *M. tuberculosis* H37Rv was used as the baseline of expressed proteins to determine the relative abundance of the corresponding proteins in SAWC 3517 and SAWC 3651, respectively. Proteins were considered to be up or down-regulated when a fold change of two or more was observed for a single protein, with a p-value of ≤ 0.05.

Table 4.4.7 summarises the number of genomic differences and differentially abundant proteins in SAWC 3517 and SAWC 3651. SAWC 3517 and SAWC 3651 have similar numbers of SNPs with regards to the reference strain, *M. tuberculosis* H37Rv. Both strains contain 65 in/dels, of which 14 are unique to each. A total of 24 genes are affected by the RDs present in SAWC 3517 and 11 genes are affected by RDs present in SAWC 3651. The difference between the SNPs, in/dels and large genomic deletions (RDs) found in each strain is discussed in Section 4.3. A total of 2249 proteins were identified in SAWC 3517 and 2161 proteins were identified in SAWC 3651. The proteins detected in only one of either *M. tuberculosis* H37Rv and either SAWC 3517 or SAWC 3651, were not included in the differential expression analysis. Differential analysis was done if a protein was present in all 4 replicates of one strain analysed and present in all 4 of the replicates of the strain that it was compared to. In SAWC 3517, 186 proteins were up-regulated and 126 were down-regulated, with respect to *M. tuberculosis* H37Rv. Eighty five proteins were up-regulated, and 157 proteins were down-regulated in SAWC 3651 (see Table 4.4.7).

Table 4.4.7. Summary of SNPs and regulated proteins in SAWC 3517 and SAWC 3651

	SNPs				In/dels		Proteins	
Strain	All SNPs	Syn. SNPs	Non-syn. SNPs	Intergenic SNPs	Small in/dels	Deleted/truncated genes	Up-regulated	Down-regulated
SAWC 3517	819	282	446	91	65	24	186	126
SAWC 3651	831	282	457	91	65	11	85	157

Figure 4.4.4 shows the up- and down-regulated proteins, in functional categories, for SAWC 3517 and SAWC 3651. Interestingly, there was no congruence between the differentially abundant proteins

identified in SAWC 3517 and SAWC 3651. I.e. none of the proteins overly abundant in one strain were also found to be overly abundant in the other and vice versa for under-abundant proteins. A large number of proteins that showed the same intensity levels as that observed in the *M. tuberculosis* H37Rv reference strain, was however detected in SAWC 3517 and SAWC 3651. Very few hypothetical proteins, transcripts of insertion sequences and phages and PE/PPE proteins were identified to be differentially regulated in SAWC 3517 and SAWC 3651. The epidemiological success of the RD^{Rio} LAM strains has been hypothesised to be partly ascribed to the deletion of certain PE/PPE genes. This seems contradictory, as other studies have shown that the deletion of PE/PPE genes contributes to high virulence in certain strains (Lazzarini *et al*., 2008; Manabe *et al*., 2003; Zhang *et al*., 2007).

From Figure 4.4.4 it is evident that the number of up-regulated proteins involved in the lipid metabolism was very similar for SAWC 3517 and SAWC 3651. However, nearly double the number of proteins involved in the lipid metabolism was shown to be down-regulated in SAWC 3517 as compared to SAWC 3651. The same distribution was seen for proteins involved in virulence, detoxification and adaptation. Approximately twice as many regulatory proteins were up-regulated in SAWC 3517 when compared to SAWC 3651. In contrast, the two strains show nearly equal numbers of proteins in this functional category that were down-regulated. In SAWC 3517, more than 3 times as many proteins involved in the intermediary metabolism and respiration were up-regulated, compared to SAWC 3651, whilst more proteins in this functional category were down-regulated in SAWC 3651 than in SAWC 3517. A comparable distribution of the number of differentially regulated proteins involved in the cell wall and cell processes was shown. Proteins with unknown function (conserved hypothetical proteins) were shown to be up- and down-regulated in SAWC 3651 and SAWC 3517. Since very little is known about the roles of these proteins in the physiology of the pathogen, one cannot make any conclusions as to what this difference in regulation of these conserved hypothetical proteins between the two strains might mean. More than twice the number of proteins involved in information pathways was shown to be up-regulated in SAWC 3517, in comparison with SAWC 3651. In contrast, twice as many proteins in this functional category were down-regulated in SAWC 3651, compared to SAWC 3517.

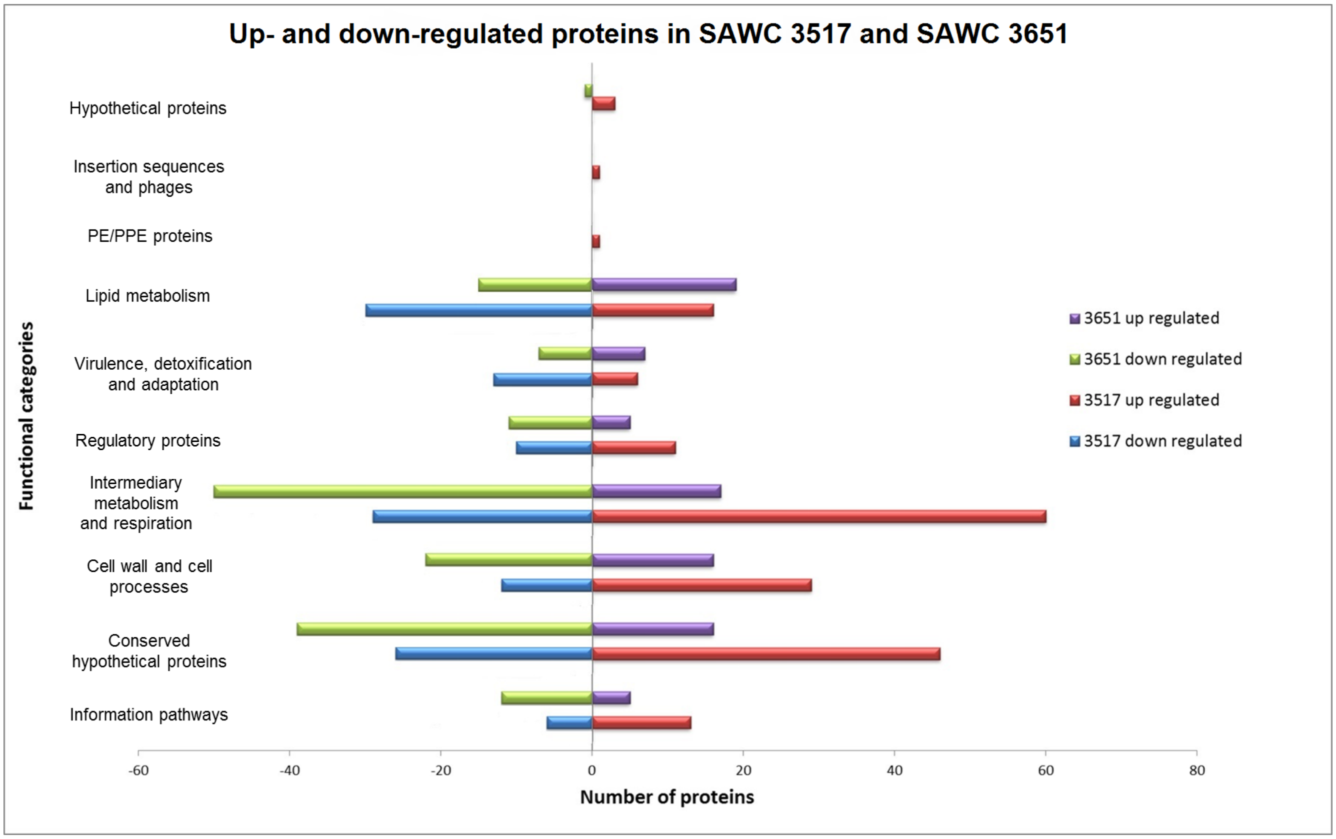


Figure 4.4.4. Up and down-regulated proteins in SAWC 3517 and SAWC 3651.

None of the proteins were detected if the corresponding genes were deleted (RD149 and RD115 in SAWC 3651, RD 149, RD174, RD^{Rio} in SAWC 3517). Note that when a protein is detected in one strain, but not in the other, one cannot confer that that protein is not at all present in the second strain, as the detection of very low abundance proteins is limited by the label-free quantitative proteomics method that is used. Hence, the protein might be present, but only in very low abundance. Nevertheless, not detecting proteins of which the corresponding genes are deleted is reassuring. The majority of proteins that were observed in only one of the strains analysed were mostly identifications based on the presence of that protein in one or two of the replicates. Those proteins were consequently not used for the quantitative comparison. Therefore no comparison was made here as to the presence of proteins in one strain compared to the absence in another. Only up- and down-regulated proteins based on the relative abundances of the identified proteins are compared. However, the fact that none of the proteins involved in the deletions present in these two strains were detected, serves as a secondary confirmation of the genomic deletions.

The up- and down-regulated proteins identified in SAWC 3517 and SAWC 3651 were mapped to *M. tuberculosis* biochemical pathways with the open-access online tool, iTUBY, a version of interactive pathways explorer (iPATH) 2 (Letunic *et al*., 2008; Yamada *et al*., 2011). The primary map used by iTUBY summarises the metabolism in biological systems as annotated in the literature to date.

Analysis of proteins with iTUBY allows for a general overview of protein functions and their interrelatedness.

The proteins involved in the RDs present in either SAWC 3517 or SAWC 3651, were mapped to the metabolic pathways that they are involved in, using iTUBY (see Figure 4.4.5). Note that some conserved hypothetical proteins cannot be mapped to biochemical pathways as no possible function for the proteins has been determined. If a possible function has been assigned to a conserved hypothetical protein based on sequence similarity, the protein will be mapped by iTUBY. The proteins involved in RD115 are not present in SAWC 3651 and their absence may have an effect on other proteins within the same pathways. The proteins involved in RD115 are shown in Figure 4.4.5.(i). The figure legend gives more details as to the sub-categories in which the proteins play a role. Likewise, proteins involved in RD174 and RD^{Rio} are shown in Figure 4.4.5.(ii) and 4.4.5.(iii), respectively. The genes encoding these proteins are deleted in SAWC 3517. Each grey block consists of a number of grey ovals that in turn, each represents a gene (Rv number) that is broadly categorised according to function, but that has not necessarily been linked to a specific pathway. These categories are indicated as A, B, and C in Figure 4.4.5.i, -ii, and –iii.

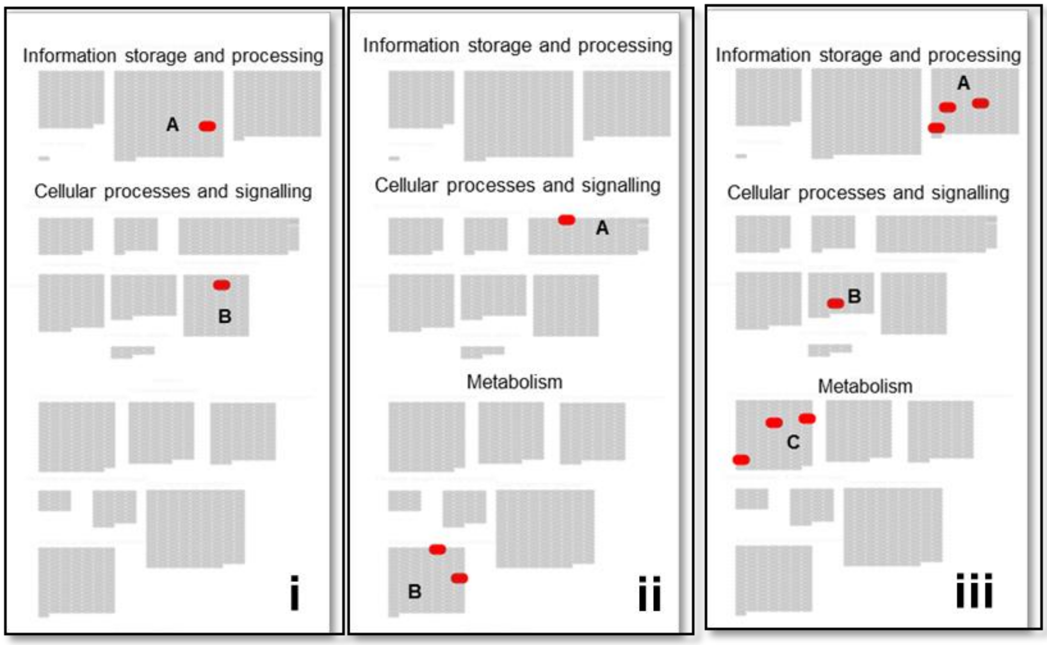

Information storage and processing
A
Cellular processes and signalling
B
i
Information storage and processing
Cellular processes and signalling
A
Metabolism
B
ii
Information storage and processing
A
Cellular processes and signalling
B
Metabolism
C
iii

Figure 4.4.5. Proteins involved in RDs in *M. tuberculosis* pathways. (i). RD115, A: Transcription, B: Posttranslational modification. (ii). RD174, A: Signal transduction mechanisms, B: Inorganic ion transport and metabolism. (iii). RD^{Rio}, A: Replication, recombination, repair, B: cell motility, C: Energy production and conversion.

Proteins that were identified to be down-regulated in SAWC 3517 were mapped to metabolic pathways in the same manner as above, using iTUBY. This map is shown in Figure 4.4.6. Compared to other bacteria, *M. tuberculosis* contains few two-component systems. Nevertheless, these systems play an important role in the early intracellular survival of the pathogen and are also implicated in aspects of mycobacterial virulence (Tucker *et al*., 2007). Proteins involved in two-component systems responsible for the maintenance of turgor pressure, initial hypoxic response and proteins involved in mineral and organic ion transport were down-regulated in SAWC 3517. The two component system, KdpD/KdpE, down-regulated here, has been shown to be an important adaptive mechanism employed during host infection, enhancing virulence of the bacteria and promoting survival of the pathogen, besides its role as a potassium regulator (Freeman *et al*., 2013).

Ether lipids can constitute major parts of the cell membranes of many bacteria, ether lipids may also act as antioxidants, but their role in mycobacteria is largely unknown. Proteins involved in ether-lipid metabolism, as well as various general energy-related metabolic pathways were down-regulated in SAWC 3517. These include pyruvate metabolism, the tri-carboxylic acid (TCA) cycle, oxidative phosphorylation and general amino acid metabolisms. Several proteins involved in information storage and processing (see legend of Figure 4.4.6) and proteins involved in purine- and pyrimidine metabolism were down-regulated in the non-RD^{Rio} representative of *M. tuberculosis*. Other down-regulated proteins in this strain include various conserved proteins with unknown function (conserved hypothetical proteins) and proteins involved in the porphyrin metabolism of *M. tuberculosis* – possibly playing a role in the iron homeostasis of the bacterium and heme degradation in the host (Nambu *et al*., 2013).

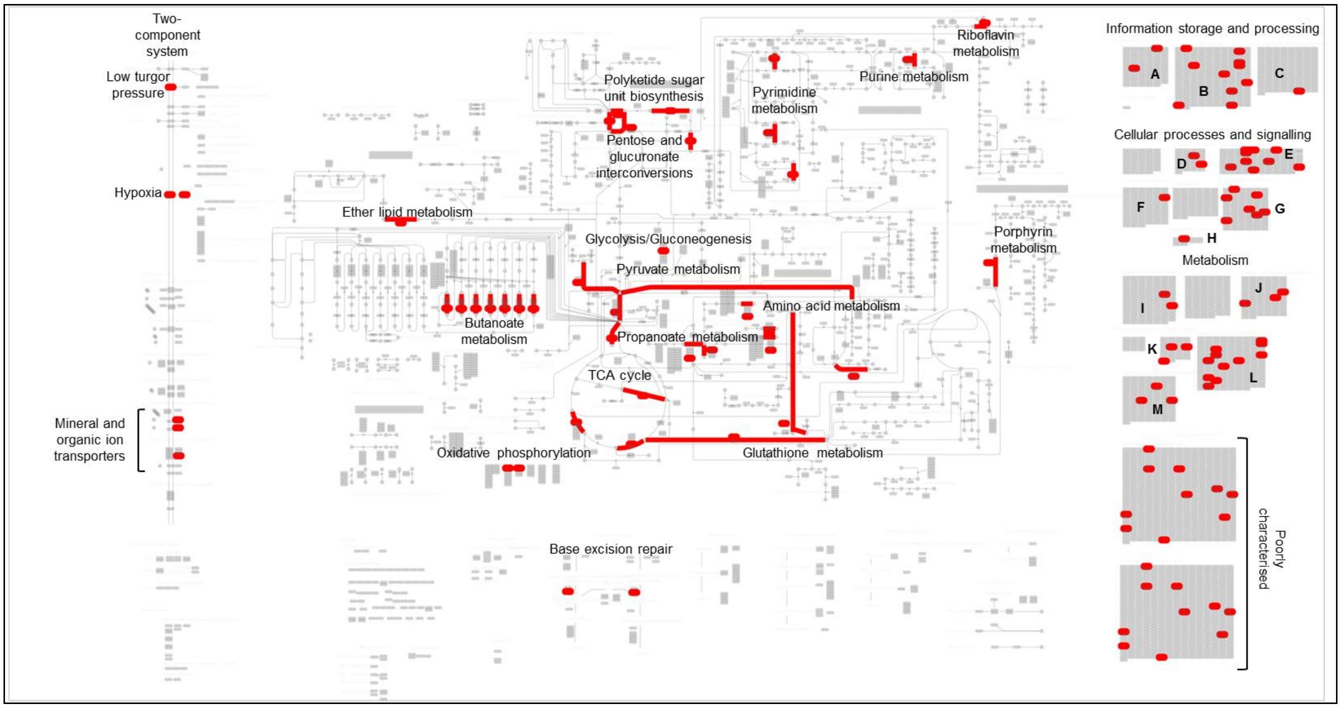


Figure 4.4.6. Down-regulated proteins in SAWC 3517. A: Translation, ribosomal structure, B: Transcription, C: Replication, recombination, repair, D: Defence mechanisms, E: Signal transduction mechanisms, F: Cell wall/membrane, G: Posttranslational modification, H: Intracellular trafficking, I: Energy production and conversion,

J: Amino acid transport and metabolism, K: Coenzyme transport, L: Lipid transport and metabolism, M: Inorganic ion transport and metabolism.

The proteins that were identified to be up-regulated in SAWC 3517 are shown in Figure 4.4.7. The up- and down regulation of proteins can be caused by genomic alterations such as SNPs, small in/dels and RDs, or external stressors. Since the strains in question here were all cultured under the same conditions, one could assume that these external factors are the same for all strains and that the regulated proteins seen here are mainly due to genomic differences between the strains. Proteins can also be up- or down-regulated in response to one another, e.g. the up-regulation of one protein due to an underlying genomic feature, leads to the up- or down-regulation of another protein in a related pathway, setting in place a cascade of events. Proteins involved in an array of biochemical pathways were up-regulated in SAWC 3517. Pathways affected are labelled on the figure and the sub-categories of general pathways (information storage and processing, cellular processes and signalling, and metabolism) are explained in the legend of Figure 4.4.7.

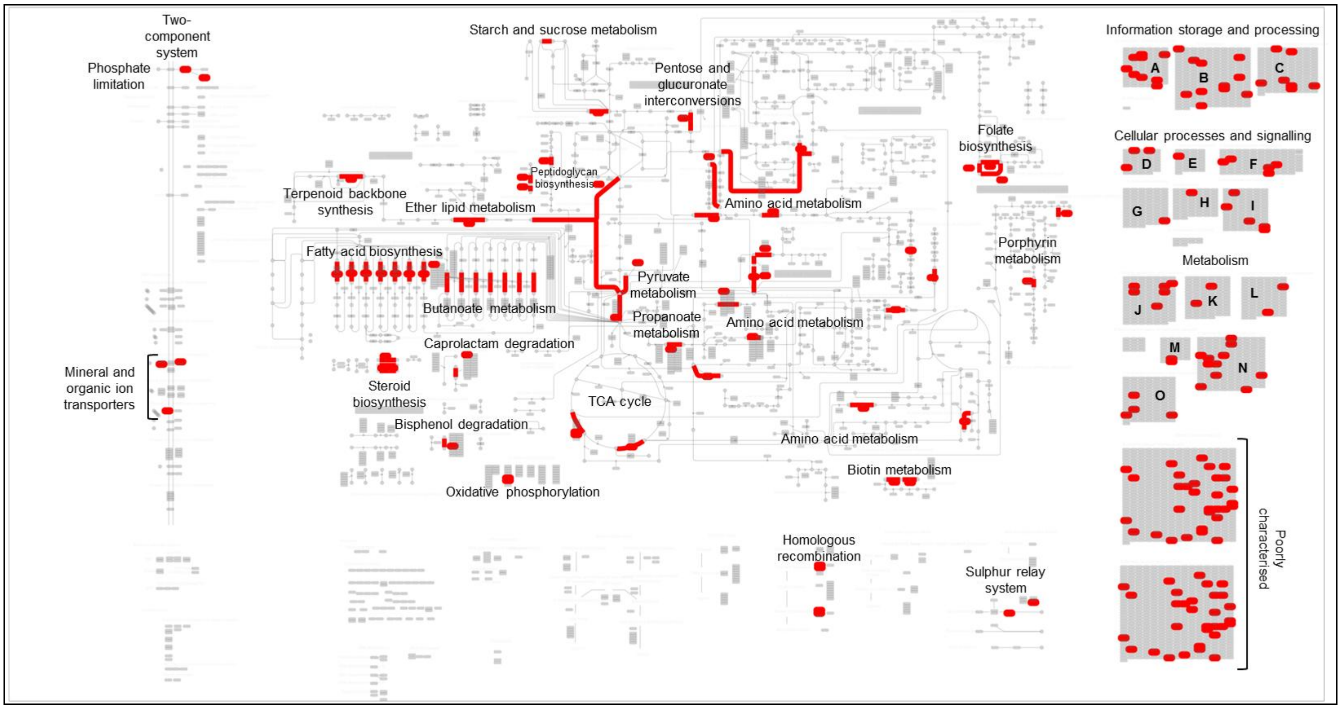


Figure 4.4.7. Up-regulated proteins in SAWC 3517. A: Translation, ribosomal structure, B: Transcription, C: Replication, recombination and repair, D: Cell cycle control, cell division, E: Defence mechanisms, F: Signal transduction mechanisms, G: Cell wall/membrane, H: Cell motility, I: Posttranslational modification, J: Energy production and conversion, K: Carbohydrate transport, L: Amino acid transport and metabolism, M: Coenzyme transport, N: Lipid transport and metabolism, O: Inorganic ion transport and metabolism.

Proteins that were down-regulated in SAWC 3651 were also mapped to biochemical pathways and are shown in Figure 4.4.8. Several of the down-regulated proteins have unknown functions. The general biochemical pathways are labelled on the figure. Sub-categories of the information storage and processing, cellular processes and signalling and metabolism are labelled A – M and are explained in the legend of Figure 4.4.8.

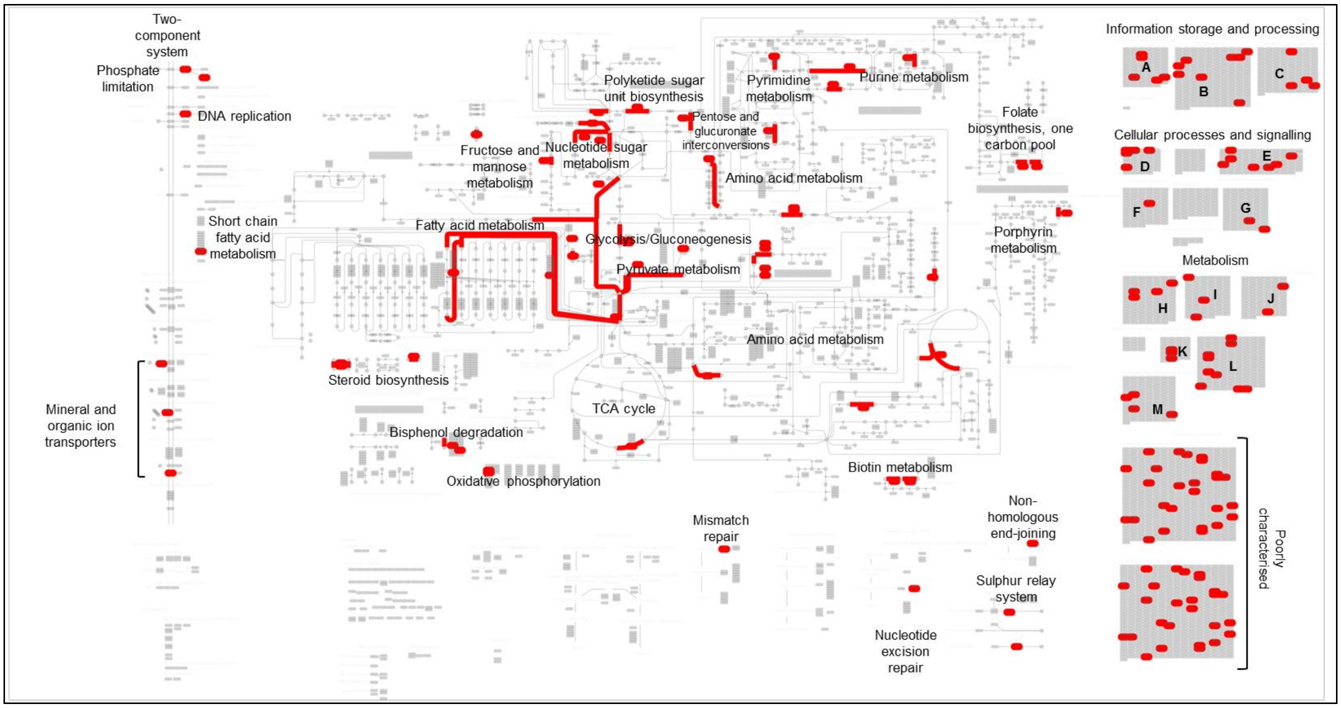

Two-component system
Phosphate limitation
DNA replication
Short chain fatty acid metabolism
Mineral and organic ion transporters
Polyketide sugar unit biosynthesis
Pyrimidine metabolism
Purine metabolism
Fructose and mannose metabolism
Nucleotide sugar metabolism
Pentose and glucuronate interconversions
Amino acid metabolism
Folate biosynthesis, one carbon pool
Fatty acid metabolism
Glycolysis/Gluconeogenesis
Pyruvate metabolism
Porphyrin metabolism
Amino acid metabolism
Steroid biosynthesis
TCA cycle
Bisphenol degradation
Oxidative phosphorylation
Biotin metabolism
Mismatch repair
Non-homologous end-joining
Sulphur relay system
Nucleotide excision repair
Information storage and processing
A
B
C
Cellular processes and signalling
D
E
F
G
Metabolism
H
I
J
K
L
M
Poorly characterised

Figure 4.4.8. Down-regulated proteins in SAWC 3651. A: Translation, ribosomal structure, B: Transcription, C: Replication, recombination, repair, D: Cell cycle control, cell division, E: Signal transduction mechanisms, F: Cell wall/membrane, G: Posttranslational modification, H: Energy production and conversion, I: Carbohydrate transport, J: Amino acid transport and metabolism, K: Coenzyme transport, L: Lipid transport and metabolism, M: Inorganic ion transport and metabolism.

Proteins identified in SAWC 3651 that were shown to be up-regulated in comparison with *M. tuberculosis* H37Rv were mapped to the corresponding biochemical pathways using iTUBY. This map is shown in Figure 4.4.9 and the general metabolic pathways are labelled on the figure. The sub-categories of the information storage and processing, cellular processes and signalling and metabolism are named A – M and are described in the legend of the figure. Fewer proteins were up-regulated in SAWC 3651 with respect to with *M. tuberculosis* H37Rv, than was up-regulated in SAWC 3517 with respect to *M. tuberculosis* H37Rv.

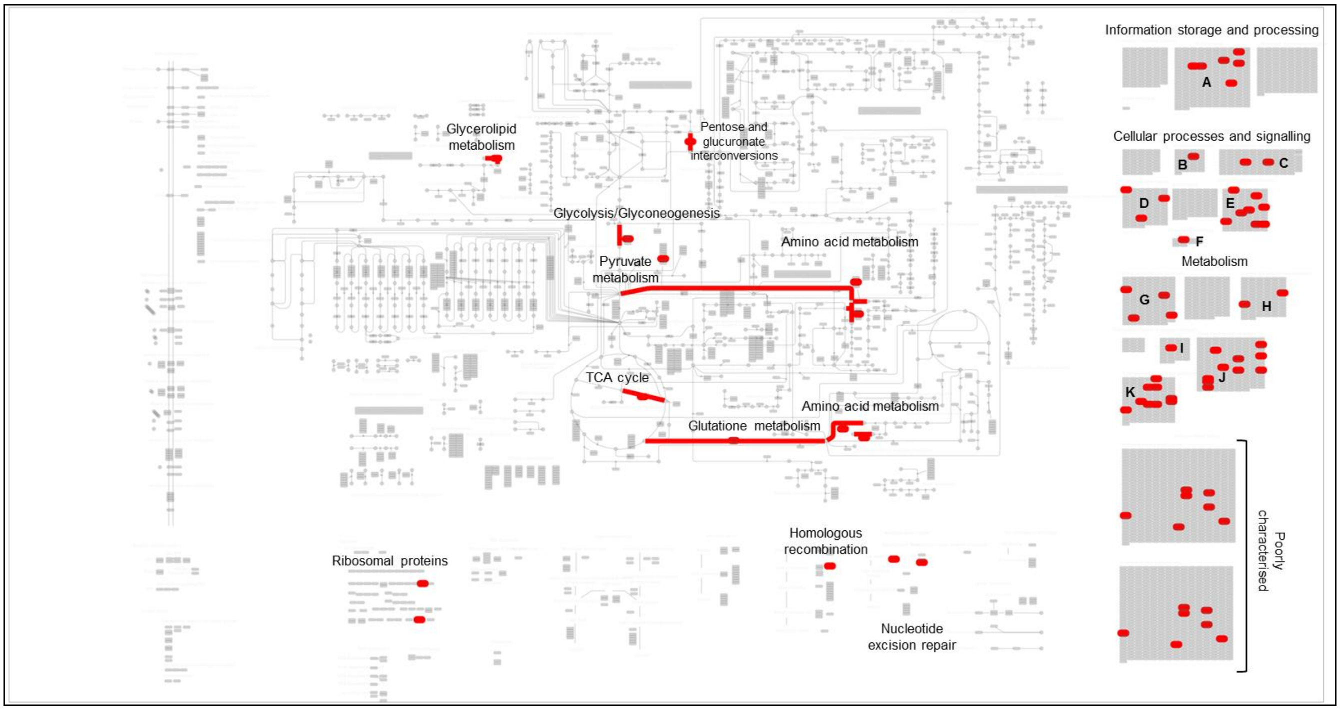


Figure 4.4.9. Up-regulated proteins in SAWC 3651. A: Transcription, B: Defence mechanisms, C: Signal transduction mechanisms, D: Cell wall/membrane, E: Posttranslational modification, F: Intracellular trafficking, G: Energy production and conversion, H: Amino acid transport and metabolism, I: Coenzyme transport, J: Lipid transport and metabolism, K: Inorganic ion transport and metabolism.

As expected, the association between SNPs and functional groups seen in SAWC 3517 and SAWC 3651 with respect to the *M. tuberculosis* H37Rv reference genome, correlates with the differentially abundant proteins identified in the same strain to a certain extent, see Figure 4.4.10. However, some functional categories that are highly represented in the SNP analysis showed very few to no corresponding proteins to be differentially regulated. This was largely restricted to the PE/PPE family proteins in both SAWC 3517 and SAWC 3651. A larger percentage of proteins involved in the lipid metabolism were shown to be down-regulated in SAWC 3517, and up-regulated in SAWC 3651, compared to the percentage of SNPs corresponding to this functional category in both strains. The functional category, “intermediary metabolism and respiration” show similar proportions of SNPs in both SAWC 3517 and SAWC 3651, whilst 32% of the up-regulated proteins in SAWC 3517 belong to this category and 23% in this category were down-regulated. The opposite is seen for the proteins involved in intermediary metabolism and respiration in SAWC 3651.

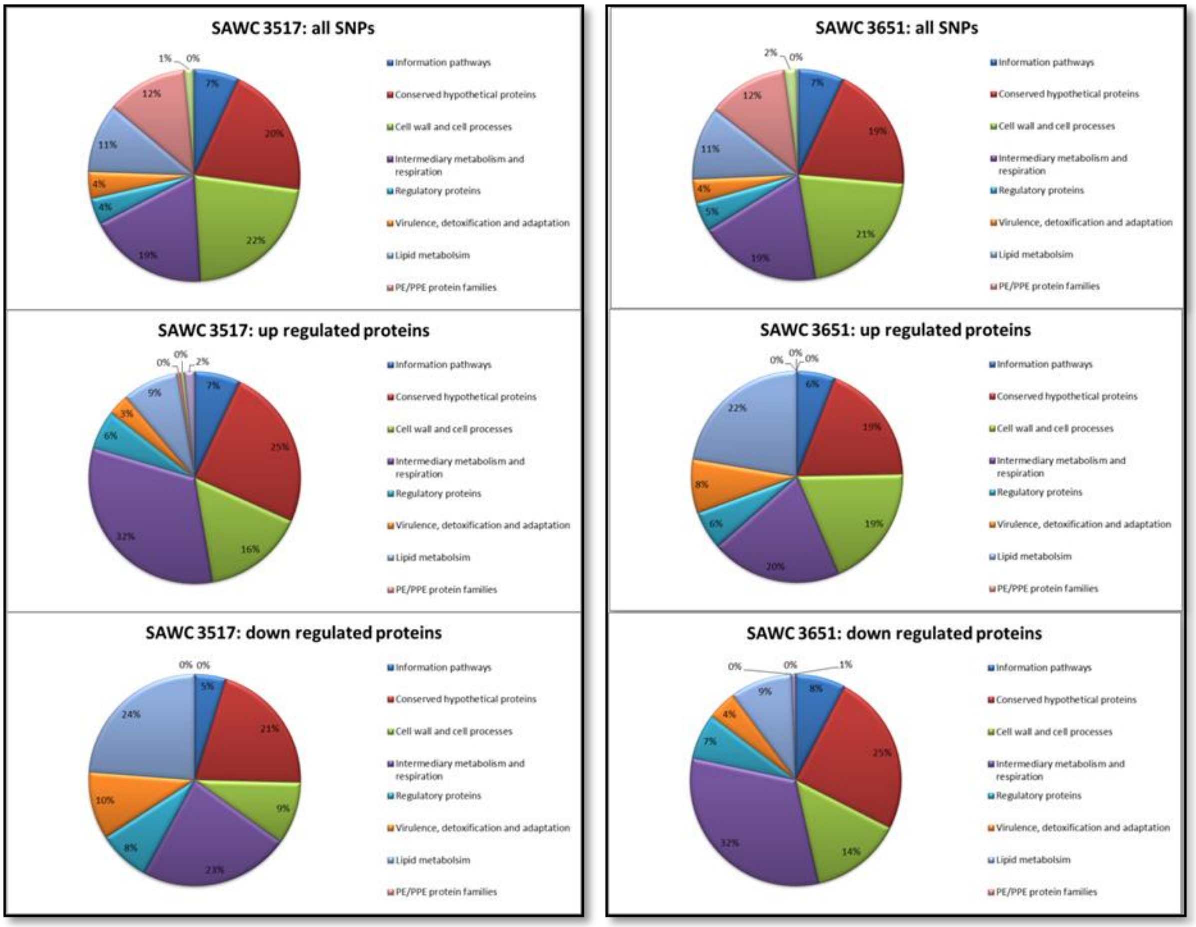


Figure 4.4.10. Summary of SNPs, up- and down-regulated proteins from SAWC 3517 and SAWC 3651. A. SAWC 3517. B. SAWC 3651

The proteins identified in SAWC 3517 that were quantified as more or less abundantly expressed than the corresponding protein in *M. tuberculosis* H37Rv were considered and the ten proteins with the highest fold of up- and down regulation are summarised in Table 4.4.8. Proteins that are up-regulated in SAWC 3517 but down-regulated in SAWC 3651 are highlighted in purple, whilst proteins that are down-regulated in SAWC 3517 but up-regulated in SAWC 3651 are highlighted in green. This colouring scheme is also used in Table 4.4.9. A complete list of all proteins identified to be up- and down-regulated in one isolate and down-regulated in the other, and vice versa, is given in Appendix D.

Table 4.4.8.Decription of 10 highly up- and down-regulated proteins in SAWC 3517

	Protein	Name	Function	Functional group	Fold regulated	p-value
SAWC 3517 up-regulated	Rv1883c	Rv1883c	Conserved hypothetical protein	Conserved hypothetical	158.1	0.000194
	Rv0787	Rv0787	Hypothetical protein	Conserved hypothetical	147.5	0.000199
	Rv1604	ImpA	Probable inositol-monophosphatase (imp)	Intermediary metabolism and respiration	53.3	0.000522
	Rv3323c	MoaX	Probable MoaD-MoaE fusion protein	Intermediary metabolism and respiration	46.7	0.000519
	Rv3322c	Rv3322c	Possible methyl transferase	Intermediary metabolism and respiration	43.0	1.42E-05
	Rv1318c	Rv1318c	Possible adenylate cyclase (ATP pyrophosphate-lyase, adenylyl cyclase)	Intermediary metabolism and respiration	33.9	0.000207
	Rv3874	EsxB	10 kDa culture filtrate antigen (LHP, cfp10)	Cell wall and cell processes	32.8	0.00514
	Rv1644	TsnR	Possible 23s rRNA methyl transferase	Information pathways	29.0	0.026018
	Rv3327	Rv3327	Probable transposase fusion protein	Insertion seqs and phages	23.7	0.01342
	Rv2821c	Rv2821c	Conserved hypothetical protein	Conserved hypothetical	20.0	0.000399
SAWC 3517 down-regulated	Rv2258c	Rv2258c	Possible transcriptional regulatory protein	Regulatory	387.1	0.000613
	Rv0670	End	Probable endonuclease iv (endodeoxyribonuclease iv, apurinase)	Information pathways	72.9	0.00382
	Rv2428	AhpC	Alkyl hydroperoxide reductase C protein	Virulence, detoxification and adaptation	70.4	0.011916
	Rv1177	FdxC	Probable ferredoxin	Intermediary metabolism and respiration	24.6	0.01942
	Rv2429	AhpD	Alkyl hydroperoxide reductase D protein	Virulence, detoxification and adaptation	21.9	0.00793
	Rv0034	Rv0034	Conserved hypothetical protein	Conserved hypothetical	20.6	0.014389
	Rv1687c	Rv1687c	Probable conserved ATP-binding proteins, ABS transporter	Cell wall and cell processes	17.1	0.00866
	Rv2930	FadD26	Fatty-acid-AMP ligase (synthetase, synthase)	lipid metabolism	15.6	0.025218
	Rv2476c	Gdh	Probable NAD-dependent glutamate dehydrogenase (NAD-dependent glutamic dehydrogenase)	Intermediary metabolism and respiration	14.4	0.000931
	Rv1445c	DevB	Probable 6-phosphogluconolactonase (6PGL)	Intermediary metabolism and respiration	14.0	0.000217

The regulated proteins identified in SAWC 3651 with respect to *M. tuberculosis* H37Rv were considered and the ten proteins with the highest fold of up- and down regulation are summarised in Table 4.4.9. Proteins that are up-regulated in SAWC 3651 but down-regulated in SAWC 3517 are highlighted in green, whilst proteins that are down-regulated in SAWC 3651 and up-regulated in SAWC 3517 are highlighted in purple.

Table 4.4.9.Decription of 10 highly up- and down-regulated proteins in SAWC 3651

	Protein	Name	Function	Functional group	Fold regulated	p-value
SAWC 3651 up-regulated	Rv0379	SecE2	Possible protein transport protein	Cell wall and cell processes	181.4	0.009525
	Rv3472	Rv3472	Conserved hypothetical protein	Conserved hypothetical	65.2	0.039015
	Rv3854c	EthA	Mono-oxygenase	Intermediary metabolism and respiration	38.8	0.04152
	Rv0045c	Rv0045c	Possible hydrolase	lipid metabolism	17.5	0.020074
	Rv2923c	Rv2923c	Conserved hypothetical protein	Conserved hypothetical	16.3	0.002493
	Rv1687c	Rv1687c	Probable conserved ATP-binding protein ABC transporter	Cell wall and cell processes	15.0	1.36E-06
	Rv0360c	Rv0360c	Conserved hypothetical protein	Conserved hypothetical	13.7	8.94E-05
	Rv1187	RocA	Probable pyrroline-5-carboxylate dehydrogenase	Intermediary metabolism and respiration	13.0	0.004345
	Rv2428	AhpC	Alkyl hydroperoxide reductase C	Virulence, detoxification and adaptation	10.9	0.005588
	Rv2429	AhpD	Alkyl hydroperoxide reductase D	Virulence, detoxification and adaptation	10.4	0.010209
SAWC 3651 down-regulated	Rv1883c	Rv1883c	Conserved hypothetical protein	Conserved hypothetical	209.8	0.005276
	Rv3323c	MoaX	Probable MoaD-MoaE fusion protein	Intermediary metabolism and respiration	102.0	0.036147
	Rv3322c	Rv3322c	Possible methyl transferase	Intermediary metabolism and respiration	89.8	0.052514
	Rv1318c	Rv1318c	Possible adenylate cyclase (ATP pyrophosphate-lyase, adenylyl cyclase)	Intermediary metabolism and respiration	49.3	0.052388
	Rv2702	PpgK	Polyphosphate glucokinase (Polyphosphate-glucose phosphotransferase)	Intermediary metabolism and respiration	38.7	0.014812
	Rv1844c	Gnd1	Probable 6-phosphogluconate dehydrogenase	Intermediary metabolism and respiration	32.4	0.001226
	Rv0958	Rv0958	Possible magnesium chelates	Intermediary metabolism and respiration	15.5	0.000211
	Rv1698	Rv1698	Outer membrane protein	Cell wall and cell processes	15.4	0.003256
	Rv3213c	Rv3213c	Possible SOJ/para-related protein	Cell wall and cell processes	14.8	5.12E-05
	Rv1154c	Rv1154c	Hypothetical protein	Conserved hypothetical	13.6	0.043581

All the proteins that were identified in SAWC 3517 and SAWC 3651 and that were shown to be differentially abundant were compared. It was shown that 68 proteins were up-regulated in SAWC 3517 (with respect to *M. tuberculosis* H37Rv) and down-regulated in SAWC 3651 (also with respect to *M. tuberculosis* H37Rv). A total number of 40 proteins were up-regulated in SAWC 3651 and down-regulated in SAWC 3517. The functional classification of these proteins are summarised in Table 4.4.10 and a more comprehensive list of these proteins are included in Appendix D. The remaining proteins were either differentially abundant in one strain and identified in similar abundance as the corresponding protein in *M. tuberculosis* H37Rv, or not only identified in one of the respective isolates.

Table 4.4.10. Functional distribution of proteins that are up-regulated in one isolate and down-regulated in the other

Functional category	Up in SAWC 3517, down in SAWC 3651	Up in SAWC 3651, down in SAWC 3517
Information pathways	5	1
Conserved hypothetical proteins	16	6
Cell wall and cell processes	9	5
Intermediary metabolism and respiration	24	7
Regulatory proteins	4	3
Virulence, detoxification and adaptation	3	4
Lipid metabolism	6	14
PE/PPE protein families	0	0
Insertion sequences and phages	0	0
Hypothetical	1	0
Total	68	40

From this analysis, two operons, as well as the ESX genes identified and that are known to be involved in virulence, were selected with the view to elucidate possible mechanisms of gene regulation. Some of the differentially abundant proteins mentioned in this discussion are not listed in Table 4.4.8 or 4.4.9, but are included in Appendix D.

MOLYBDENUM COFACTOR BIOSYNTHESIS ASSOCIATED PROTEINS

Proteins involved in molybdenum cofactor (MoCo) biosynthesis and genes in the same genomic region as MoCo biosynthesis genes were identified in over-abundance in SAWC 3517. These proteins include MoeW (Rv2338c), MoaA (Rv3119), Rv3322c, MoaX (Rv3323c) and MoaC3 (Rv3324c). Molybdenum uptake has been implicated in mycobacterial virulence, likely through the function of one or a combination of molybdenum cofactor (MoCo) dependent proteins (Williams *et al.*, 2011). MoCo allows molybdenum to be bound by enzymes in order to catalyse redox reactions in carbon-, nitrogen- and sulphur metabolism. The biosynthesis of MoCo is mediated by various enzymes, including

Rv3323c, Rv2338c, Rv3119, Rv3324c (Williams *et al*., 2011; Zhang and Gladyshev, 2008). Unfortunately, very little is known about the regulation of the MoCo biosynthetic pathway in *M. tuberculosis*. Despite the fact that several homologs of the MoCo biosynthetic pathway genes are found in *M. tuberculosis*, certain genes involved in this pathway were shown to be essential for the *in vitro* growth of *M. tuberculosis* (Lamichhane *et al*., 2003; Sassetti *et al*., 2003). *M. tuberculosis* MoCo biosynthesis protein X *(*MoaX- a MoaD and MoaE fusion protein) mutants were shown to have a reduced ability to infect macrophages (Rosas-Magallanes *et al*., 2007). Furthermore, MoaX have been shown to provide additional MoaD and MoaE activity under certain conditions, and is implicated in *de novo* MoCo biosynthesis *in vitro* (Williams *et al*., 2011). The over-abundance of proteins involved in the initial steps of MoCo biosynthesis in SAWC 3517, could contribute significantly to the hypothesised increased virulence and transmissability of RD^{Rio}-LAM strains (Lazzarini *et al*., 2007, 2008). This could, in part, explain the epidemiological success of the RD^{Rio} lineage globally. In contrast, Rv0438c (MoeA2), Rv3322c and Rv3323 (MoaX) were shown to be present in low abundance in SAWC 3651 in comparison with *M. tuberculosis* H37Rv. This suggests that the regulation of the MoCo biosynthetic pathway lies within the genomic differences between SAWC 3517 and SAWC 3651. Despite extensive mining of the literature and searching numerous databases for indications of protein interactions or regulatory mechanisms of the MoCo biosynthesis pathway, no evident explanations were found that may aid in the comprehension of the phenomenon observed between the two *M. tuberculosis* clinical isolates representative of the RD^{Rio}-LAM and a non-RD^{Rio}-LAM lineages analysed in this section.

OXIDATIVE STRESS RESPONSE PROTEINS

Analysis of the second operon, including Rv2428 and Rv2429, showed that these proteins were down-regulated in SAWC 3517 and up-regulated in SAWC 3651. Certain mutations in Rv2428 have been shown to confer Isoniazid resistance (Sandgren *et al*., 2009). Rv2428 has been predicted to be in the RelA/Rv2583c regulon, while Rv2429 is predicted to be in the SenX3/ Rv0490-RegX3/Rv0491 regulon (Dahl *et al*., 2003; Parish *et al*., 2003). Rv0491 were shown to be over-abundant in SAWC 3517, and under-abundant in SAWC 3651. Rv2428 and Rv2429 are involved in oxidative stress response and work together to establish an NADH-dependent peroxidase and peroxynitrite reductase that provides cellular protection. The mechanism governing the expression of this operon’s transcription and translation remains to be determined.

ESX-1 SECRETION SYSTEM PROTEINS

The *M. tuberculosis* ESX-1 secretion system acts to secrete virulence factors across the cell envelope (Bitter *et al*., 2009). Attenuated *M. bovis* BCG vaccine strains have a deletion of a 9.5 kb genomic region spanning nine open reading frames of genes primarily involved in the ESX-1 secretion system.

This genomic region is termed RD1 and is present in all virulent *M. tuberculosis* strains and has thus been implicated in virulence and pathogenesis of *M. tuberculosis* (Daugelat *et al*., 2003; Lewis *et al*., 2003). The ESX-1-secretion associated protein, EccA1 has been shown to influence mycolic acid synthesis and the mycolic acid composition of the cell envelope, which may correlate with the SNP changes which were significantly associated with cell wall processes, and therefore the influence of evolution on the cell wall. The deletion of, or mutations in, EccA1 were shown to reduce the mycolic acid composition of the envelope and cell wall integrity and also reduced virulence and intracellular growth (Joshi *et al*., 2012). Substrates of the ESX-1 secretion system are mutually dependent on each other for secretion (Fortune *et al*., 2005).

Several ESX-1 associated proteins were shown to be over abundantly expressed in SAWC 3517, compared to *M. tuberculosis* H37Rv. Predicted ESX-1-secretion-associated proteins EspE, EspF, EspG1and EspK (Rv3864, Rv3865, Rv3866 and Rv3879c, respectively) were present in high abundance in the RD^{Rio}-LAM representative. In addition, EccA (Rv3868) and EccD1 (Rv3877) were also overly abundant. EccA is a conserved component of the ESX-1 secretion system, whilst EccD1 is predicted to be a transmembrane protein mediating the transport of virulence factors across the mycobacterial cell membrane (Abdallah *et al*., 2007; Bitter *et al*., 2009). Furthermore, EsxB (Rv3874), also known as the mycobacterial antigen CFP10, is correspondingly overly abundant in SAWC 3517. This suggests that the general ESX-1 secretion system is up-regulated in the RD^{Rio}-LAM representative, even though some of the proteins encoded for by ESX-1 secretion genes are not detected. The cellular localisation of some of the proteins involved in this secretion system (membrane bound proteins) could restrict their detection by means of protein mass spectrometry of the whole cell lysate proteins of the cultured isolate. The high abundance of ESX-1 associated proteins correlate with the hyper-virulence of RD^{Rio}-LAM strains that is implied by the epidemiological dominance that this lineage presents with (Gibson *et al*., 2008; Lazzarini *et al*., 2008, 2007). In contrast, EspG, EccA1 and EspK (Rv3866, Rv3868 and Rv3879, respectively) were shown to be present in low abundance in SAWC 3651, in comparison with *M. tuberculosis* H37Rv.

4.4.3. SUB-CONCLUSIONS

It is evident from on-going whole genome and proteomic studies on *M. tuberculosis*, including the present study, that the genomes of circulating *M. tuberculosis* strains are actively evolving. These genomic changes cause significant changes in abundance of a subset of cellular proteins which could impact on the physiology of the strain.

The immense differences in the proteomes of seemingly closely related strains of *M. tuberculosis* LAM genotype isolates are shown in Section 4.4. It is clear that SNPs, small in/dels and the absence or truncation of genes have a significant effect on the expression of not only the corresponding proteins but also on other proteins that are functionally related or that may be involved in the same pathways. The causal relationship of these genomic changes on the proteome is not obvious. It is thus very challenging to attribute proteomic abundance changes to specific genomic events with this whole proteome approach. Despite these challenges we were able to show that various virulence associated proteins are up-regulated in the RD^{Rio}-LAM lineage, however this needs further validation with additional functional and experimental work.

From the data shown in this section, it is apparent that the proteome of SAWC 3517 (the representative of the RD^{Rio} LAM lineage) is severely affected by the underlying genomic changes that it harbours, these include the large genomic deletions, RD149, RD174 and RD^{Rio} (the largest genomic deletion described to date) (Lazzarini *et al*., 2008). The functional classification of the results suggests that proteins involved in the lipid metabolism and intermediary metabolism and respiration may have a critical role to play in the success of the RD^{Rio} LAM lineage. Furthermore, it is evident from the summary of evolutionary changes shown in Section 4.3, together with what is shown in this section, that the genomic- and proteomic features involved in the cell wall and cell processes of the LAM genotype in general differ from what is seen in the reference strain, *M. tuberculosis* H37Rv, possibly contributing to the success of this strain family world-wide.

CHAPTER 5

SUMMARY & CONCLUSION

The current research used whole genome sequencing to describe the genomes of strains representative of the different sub-lineages of the LAM genotype from Cape Town, South Africa. In addition, proteomic data was analysed to identify different physiological traits associated with two different LAM strain families, one of which (RD^{Rio}) was shown to cause large outbreaks and is thought to be hyper-virulent.

The customised NGS analysis pipeline established during this study successfully identified the mechanism of genetic variation within the previously sequenced *M. tuberculosis* strain of the LAM genotype (F11). This conferred high confidence to the genetic variation detected in other strains that had not been previously sequenced. The pipeline also allowed for the identification of micro-evolutionary events that distinguish strains and strain families from each other, with high confidence. The unmatched resolution that WGS provides was exploited with this multi-software approach to call genetic variation for the purpose of inferring evolutionary relationships between the isolates analysed in the present study.

Previous studies have shown that SNPs significantly contribute to mycobacterial genome variation and have been proved to be suitable to be able to distinguish closely related isolates from each other (Gutacker *et al*., 2002). Based on this principle it was possible to reconstruct the phylogenetic history of *M. tuberculosis* strains representative of the different IS*6110* RFLP families and members of the *M. tuberculosis* complex. The resulting phylogenetic tree predicted a clonal population structure consisting of distinct lineages. IS*6110* RFLP families and spoligotypes were unambiguously associated with defined lineages, confirming the robustness of the SNP-generated phylogeny. This phylogeny of strains from the Western Cape Province of South Africa proved to be congruent with previous evolutionary scenarios for *M. tuberculosis* isolated from other parts of the world (Baker *et al*., 2004; Filliol *et al*., 2006; Gagneux *et al*., 2006). This demonstrates that the TB epidemic in this region is reflective of the global epidemic and probably arose as the result of these strains being imported from both East and West through sea trade.

The *M. tuberculosis* strain families in PGG2 contribute significantly to the TB epidemic in Cape Town, South Africa and represent more than 50% of all the TB cases in this setting (Streicher, 2007). Strains with a PGG2 classification have also been shown to contribute significantly to the global epidemic and are referred to as members of the Euro-American Clade (Gagneux *et al*., 2006). This detailed analysis of the PGG2 LAM genotype is the first to be done on this genotype with the view to understand the genetic mechanisms that contribute to the success of this genotype. The present study used whole genome sequencing data (SNPs and LSPs) together with previously published markers (i.e. SNP, LSP, IS*6110* insertion site mapping and spoligotyping) to determine the evolutionary relationships among IS*6110*-RFLP strain families of the LAM genotype of *M. tuberculosis*. This analysis demonstrated that the LAM genotype evolved from a common ancestor characterised by an IS*6110* insertion in the DR region, the deletion of DVRs 21-24, 33-36, the PGG2 defining *katG*463CTG→CGG polymorphism, and 55 unique LAM specific SNPs. Thereafter six LAM sub-lineages evolved, namely IS*6110* RFLP families F9, F11, F13, F14, F15, and F26. Each of the sub-lineages are characterized by a uniqe set of SNPs and reflect evolution in genes involved in cell wall, cell processes, intermediary metabolism and respiration. The same phenomenon was observed when non-LAM isolates were analysed. This suggests that the evolution of *M. tuberculosis* (in this setting) is unidirectional and the high number of non-synonymous changes observed is evident of the positive selective pressures experienced in the host by these strains.

The on-going whole genomic- and proteomic studies on *M. tuberculosis,* such as the present study, have proved that the genomes of circulating *M. tuberculosis* strains are dynamically evolving. The genomic changes that circulating *M. tuberculosis* strains undergo may cause detectable modifications in the phenotypes of these strains, which can be quantitatively measured with highly sensitive protein mass spectrometry. The role of these genetic changes in the fluctuating levels of virulence and strain fitness remains to be clarified.

Quantitative proteomics studies are superior to conventional gene expression (transcriptional) studies in the sense that they provide insight into the proteins that are truly present at a given point in time. This study confirmed that the proteomes of *M. tuberculosis* isolates of the same lineage (LAM) show vast differences that can be ascribed to the underlying genetic variation. SNPs, small in/dels and the absence or truncation of genes due to the presence of RDs have an immense effect on the expression of the corresponding- and functionally related proteins.

The functional classification of the regulated proteins in the representative of the RD^{Rio}-LAM lineage of *M. tuberculosis* suggests that proteins involved in the MoCo biosynthesis pathway, ESX-1 secretion system, and other proteins generally involved in the lipid metabolism and intermediary metabolism and respiration may be the key to the pathogenic nature of the RD^{Rio} LAM lineage (Lazzarini *et al*., 2008). A combination of the LAM SNP analysis and the LAM RD^{Rio}/non-RD^{Rio} comparison showed that the overall genomic- and proteomic features involved in the cell wall and cell processes of the LAM genotype differ to a large extent from what is seen in the reference strain, *M. tuberculosis* H37Rv. This difference in physiology possibly contributes to the global success of the LAM lineage.

The proteome is a dynamic, constantly interacting collection of components that provide a holistic view of the biology of a living organism. In contrast, the genome of an organism provides a static view of the underlying features of biological characteristics. It is for this reason that combinatorial studies are becoming more popular and shows promise to uncover critically important and unanswered questions regarding the virulence, pathogenesis, persistence, transmissibility, general evolution and more specifically, the evolution of drug resistance in *M. tuberculosis*. The whole genome sequencing of hundreds of *M. tuberculosis* isolates world-wide, combined with a multitude of genetic typing techniques, has not provided an explicit comprehension of the fundamental mechanisms of pathogenicity. This emphasises the need for a systems biology approach to correlate genomic data with clinical and phenotypic data to better understand the fundamental aspects of *M. tuberculosis* pathogenesis.

From the current research, proteome homeostasis appears to be more complex than a mere representation of the underlying genomic features, as the altered abundance of one protein has the ability to set in place a cascade of events, not directly associated with a specific genomic variant. The entire interactome drives protein expression so that a polymorphism, deletion or truncation in/of a gene may influence the expression of an array of interrelated proteins. Our interpretation is further curtailed by a significant knowledge gap in the mechanisms of transcriptional regulation.

Despite certain shortcomings, this is one of the first studies to combine whole genome- and whole proteome data to investigate the biological characteristics of the LAM genotype of *M. tuberculosis* and it provided insightful findings. The current genome wide phylogenetic study is the first of its kind in a South African context. The present study is one of the first investigations that address the inter-LAM genomic variation as well as the genomic characteristics that separate the LAM genotype from other *M. tuberculosis* strains. Evolutionary studies, such as this one, lay the foundation for phenotypic studies, which are instrumental to novel drug- and vaccine development. The aims of this study were successfully achieved.

CHAPTER 6

FUTURE STUDIES

EXPERIMENTAL WORK

1. Investigate the effects of small genomic in/dels on the evolution of the *M. tuberculosis* LAM genotype by comparing in/dels that are common and unique in different clades of the LAM genotype.
2. Inspect reads that did not map to the *M. tuberculosis* H37Rv reference genome for genes or genomic regions that are present in the strains of interest but absent from the reference genome used for the alignment. This will be done by extracting unmapped reads, doing *de novo* assembly on these reads and performing a nucleotide BLAST on the resulting contigs.
3. PCR verify and determine the exact boundaries of newly identified LSPs that were not experimentally verified in the current study.
4. Determine whether the LSPs identified to be unique or common to clades of the LAM genotype are found in a larger mixed *M. tuberculosis* sample set, consisting of LAM and non-LAM isolates.
5. Construct a comprehensive phylogeny of *M. tuberculosis* strains from Cape Town, South Africa with representatives of all IS*6110* RFLP families, including unpublished in-house Spoligotyping data, family specific IS*6110* insertion sites, known LSPs and verified SNPs.
6. Confirm a subset of overly abundant proteins implicated in virulence in the RD^{Rio}-LAM representative strain with Western-blot analysis.
7. Inspect proteome data in order to identify candidate proteins for future screening of possible MoCo biosynthesis regulators in functional studies.

PUBLICATIONS

Title: The evolution of *M. tuberculosis* clinical isolates from Cape Town, South Africa
Authors: Dippenaar, A, van der Merwe, R, Gey van Pittius, N.C, Warren, R.M, Abdallah, A.M, Pain, A.
Status: Manuscript in preparation

Title: The proteome of *M. tuberculosis* H37Rv revisited
Authors: Dippenaar, A, Botha, L, Warren, R.M, Gey van Pittius, N.C, van Helden, P.D, Smit, S.
Status: Manuscript in advanced stage of preparation
Journal: Proteomics

APPENDICES

APPENDIX A

RECIPES AND PROTOCOLS

7H9 LIQUID MEDIA

4.7g of 7H9 powder in 900 ml dd H_2O

Add 2 ml of 100% Glycerol

Add 0.5g of Tween 80 (2.5 ml of 20% Tween)

Autoclave at 121 °C for 10 minutes. Aseptically add 100ml of ADC or DC (for proteomics experiments), Filter sterilise and store at 4 ˚C.

ALBUMIN DEXTROSE CATALASE (ADC)

25 g BSA

10 g Glucose

375 µl Catalase

Make up to 500 ml with dH_2O, dissolve with magnetic stirrer and filter sterilise before use. Store at 4°C.

DEXTROSE CATALASE (DC)

10 g Glucose

375 µl Catalase

Make up to 500 ml with dH_2O, dissolve with magnetic stirrer and filter sterilise before use. Store at 4°C.

PROTEIN EXTRACTION BUFFER

MAKE UP A WORKING STOCK OF PROTEIN INHIBITORS:

50 µl Protease cocktail in 1 ml ddH2O (filter sterilised) = working stock to be used in recipe for protein extraction buffer.

TRIS-HCL PROTEASE INHIBITOR (PROTEIN EXTRACTION) BUFFER:

- 200 ul protease inhibitor cocktail
- 50 µl 1 M Tris

- 3 µl 20% Tween 80 (filter sterilised)
- 9.47 ml ddH2O
- Total: 10 ml

RC-DC ASSAY

Add 5 µl of DC Reagent S to each 250 µl of DC Reagent A that will be needed for the run. This solution is referred to as Reagent A. (Each standard or sample assayed will require 127 µl of Reagent A)

Prepare 3-5 dilutions of a protein standard from 0.2 mg/ml to 1.5 mg/ml protein: (Rehydration buffer= Protein extraction buffer left over from protein (whole cell lysate) extraction)

BSA dilution (mg/ml)	BSA (µl)	Rehydration Buffer (µl)
0	0	25
0.2	2.5	22.5
0.5	6.25	18.75
1.0	12.5	12.5
1.5	18.75	6.25

BSA stock:

20 µl BSA + 80 µl H_2O = BSA with concentration of 2 mg/ml (0.002 g)

Pipette 25 µl of standards into clean 1.5 ml Eppendorf tubes (5 tubes for standard curve).

Pipette 5 µl of clean protein samples and 20 µl of Rehydration buffer into clean tubes, can make dilution for protein samples.

Add 125 µl RC Reagent I into each tube, vortex. Incubate tubes for 1 minute at rt.

Add 125 µl RC Reagent II into each tube, vortex. Centrifuge the tubes at rt at maximum speed for 5 minutes.

Discard the supernatant by inverting the tubes on clean absorbent tissue paper. Allow liquid to drain completely from tubes.

Discard the supernatant by inverting the tubes on clean absorbent tissue paper. Pulse centrifuge tubes. Pipette remaining fluid in tubes, without disturbing the pellet.

Add 127 µl Reagent A (that you prepared from Reagent S and Reagent A) to each tube, vortex. Incubate tubes at room temperature for 5 minutes, or until precipitate is completely dissolved. Vortex before proceeding to next step.

Add 1 ml of DC Reagent B to each tube and vortex immediately. Incubate at room temperature for 15 minutes.

After 15 minutes incubation, absorbencies can be read at 595 nm. The absorbencies will be stable for at least one hour.

Draw standard curve on Excel and determine protein sample concentrations.

Load 60 µg per lane on SDS-page.

PROTEOMICS SAMPLE PREPERATION, TRYPSYN DIGESTION, DESALTING

ABBREVIATIONS USED:

2-DE: Two-dimensional gel electrophoresis

ABC: Ammonium bicarbonate

ACN: Acetonitrile

DTT: DL-Dithiothreitol

FA: Formic acid

IAA: Iodoacetamide

REAGENTS NEEDED:

Ammonium Bicarbonate: Sigma #09830

DTT: Sigma #43817

IAA: Sigma #I1149

Water: Water LCMS Chromasolv: Sigma #39253 or good MilliQ water

Acetonitrile: Acetonitrile LCMS Chromasolv: Sigma #34967

Trypsin: Sequencing grade trypsin: Promega #PRV5111

SOLUTIONS NEEDED:

50 mM Ammonium bicarbonate: Dissolve 0.08 g ammonium bicarbonate in 20 ml MilliQ water (make fresh before use every time)

50% Acetonitrile (ACN): 5 ml ACN + 5 ml MilliQ water

70% Acetonitrile (ACN): 1400 µl ACN + 600 µl MilliQ water

10 mM DTT: 0.031 g in 20 ml 25 mM ammonium bicarbonate (make fresh before use)

55 mM IAA: 0.2 g in 20 ml 25 mM ammonium bicarbonate

Trypsin:

Dissolve a single vial of trypsin (20 µg) in 200 µl of the re-suspension solution provided with the trypsin. Divide into aliquots of 20 µl each and store at -80°C. For experimental use: take an aliquot from -80°C and add 180 µl of the 50 mM ammonium bicarbonate for active trypsin. Remember that you cannot re-use this trypsin, it must be discarded after use and made just before use. Keep one ice.

STAGE TIP PROTOCOL FOR DESALTING PEPTIDES

Prepare desalting columns:

Punch out discs of C18 empore filter with needle

Place in P200 pipette tip

Securely wedge disc into tip

Prepare sample:

Dry sample to ~30ul to remove all ACN

Acidify sample by adding 20ul 5% FA (final volume ~50ul)

Cleanup: (Use 2.5ml combitip)

Activate column with 50ul MeOH, check for back pressure

Wash column with 50ul 5% FA

Run sample through column 2x

Wash column with 50ul 5% FA

Elute sample peptides with 50ul 80% ACN, 5% FA

Dry samples in speedyvac (20-25min) and store at -20˚C until mass spectrometry is done (no longer than one month)

10 X SODIUM TETRA BORIDE (SB) BUFFER

38.14 g Sodium tetra boride in 1 l Milipore water

Pour half of the water in a conical flask and put on the magnetic stirrer, add the weighed off SB, rinse beaker with some of the water and add to conical flask. Add the rest of the water and allow to dissolve whilst stirring.

4 X LAEMMLI SAMPLE BUFFER

For 50 ml:

12 ml 1 M Tris (pH 6.8)

4g SDS powder

20 ml Glycerol

2.5 ml 0.2% Bromophenol blue

5 ml β-mercaptoethanol

14 ml ddWater

Dissolve on magnetic stirrer, can be stored at room temperature.

6 X LOADING BUFFER (PH 8)

30 ml 100% Glycerol

60 mg bromophenol blue

0,6 g SDS

Make up to 100 ml using TE, autoclave at 121 °C for 10 minutes. Store at room temperature

APPENDIX B

PRIMER SEQUENCES

Primers							
					Amplicon sizes		
RD/LSP	Name	Sequence	TM	Length	Region present	Region deleted	Comments
RDmsrA	msrA_Rv0138_Forward	5'ACATGGAGATGGATGTGGTG3'	61.96	20 bp	501 bp	853 bp	Not real deletion
	msrA_Rv0138_Int_reverse	5'TCTTCTTCCAGATCCACGAC3'	61.29	20 bp			
	msrA_Rv0138_Reverse	5'GCATAACTGCTGGTGAAGAC3'	61.34	20 bp			
RDRv0336	Rv0336_Forward	5'AGTCATCCACGTTATCCACAG3'	61.91	21 bp	500 bp	849 bp	Not real deletion
	Rv0336_Int_reverse	5'ATCAGGGTCAACGATCAAGTC'3	61.19	22 bp			
	Rv0336_Reverse	5'TAACTTCCTGGTCAGTTACCG3'	61.5	23 bp			
RD^{Rio}	RDRio_new_fwd	5' ATGACCATTCCTCCAACACC 3'	62.19	20 bp	843 bp	2041 bp	
	RDRio_new_int_rev	5' ATGAAATAAGAAGTCCTCCCAG 3'	59.97	22 bp			
	RDRio_new_rev	5' GATAGACCGTTTCCTGCCAC 3'	62.91	20 bp			
RD149	RD149_Tosl_fwd	5' CATGTCACCCTGGCCCGACGGGTC 3'	71	24 bp	1170 bp	669 bp	
	RD149_intrev	5' ACCGGCGGATTTCTTTGTCGTCGCTGTG 3'	69	28 bp			
	RD149_Tsol_rev	5' CCAGGCGATCTTCGACCACGGCACAC 3'	70	26 bp			
RD152	RD152_Tsol_fw	5' CCGGGTTGAGCAATGCGGATATCAGTGGAC 3'	68	30 bp	640 bp	639 bp	
	RD152_Tsol_int_rev	5' TCAAATCCGGTCAGGTACACGGTATTCC 3'	69.76	28 bp			
	RD152_Tsol_rev	5' TGGGATAGTTCAGGTGGCCATCGTGGGCAT 3'	70	30 bp			
	RD152_new_fw	5' GCAACCCATCTGTTCATCTC 3'	61.06	20 bp	321 bp	537 bp	
	RD152_new_int_rev	5' GTATCGGGTTCCTGATCCAC 3'	62.08	21 bp			
	RD152_new_rev	5' TGACCTATACCGACCGACTG '3	62.94	22 bp			
RD174	RD174_fwd	5’ GCGGTCCTCGTTGTCCTC 3’	60.19	18 bp	264 bp	400 bp	
	RD174_int_rev	5’ ACGGCATCCATTGTGATCTG 3’	60.33	20 bp			
	RD174_rev	5’ TGGCGATGGTCAGAAACTTG 3’	60.62	20 bp			
RD115	RD115_F	5’ CTGGTGGCACGTTCGTATTC 3’	59.96	20 bp	240 bp	352 bp	

	RD115_R1	5’ GGATCTCGGTGCCCGAAC 3’	62.39	18 bp			
	RD115_R2	5’ GCCATGTCCTGCTCAATGAC 3’	60.06	20 bp			
RDwbbL	wbbl2_fw	5' TGTCCCTGATGATTACGGTG 3'	61.65	20 bp	751 bp	672 bp	Not real deletion
	wbbL_rev	5' GTGAGCGGAAATGACTTGTG 3'	61.89	20 bp			
	LSP2_R1	5' TGCAGCCGACGAAAGTCGTT 3'	64.93	20 bp			

TM= melting temperature

APPENDIX C

IDENTIFIED IN/DELS SNPS IN GENES FOR DIFFERENT GROUPS OF ISOLATES

Table 1

	Locus	**Gene name**	**Size (bp)**	**Insertion/ Deletion**
SAWC 4498	Intergenic us. of Rv0441c	-	1	Insertion
	Intergenic us. of Rv0759c		1	Insertion
	Rv0848	*cysK2*	9	Insertion
	Intergenic us. of Rv1161	-	1	Deletion
	Rv2790c	*ltp1*	1	Insertion
	Rv2896c	*dprA*	1	Deletion
	Rv3196		3	Deletion
	Intergenic upstream of Rv3219	-	13	Deletion
SAWC 4498 and SAWC 5165	*Rv0064*		1	Deletion
	Rv0278c	*PE_PGRS3*	1	Insertion
	Rv1243c	*PE_PGRS23*	1	Deletion
	Rv2490c	*PE_PGRS43*	1	Deletion
	Rv3347c	*PPE55*	1	Deletion
	Rv3854c	*ethA*	1	Insertion
SAWC 5165	TBFG_10555	*F11 conserved transmembrane protein*	1	Insertion
	Intergenic	-	1	Insertion
	Intergenic us. of *Rv2101*	-	3	Deletion
	Intergenic us. of *Rv2683*	-	1	Insertion

Table 2

Position	**Gene**	**Gene name**	**Ref**	**Alt**	**Codon number**	**Ref codon**	**Ref AA**	**Alt codon**	**Ref AA**	**Syn/non-syn**	**Gene description**	**Functional category**
33457	*Rv0030*		C	T	78	CAC	H	CAT	H	SYN	hypothetical protein	Conserved hypothetical
74059	*Rv0066c*	*icd2*	C	A	151	AAG	K	AAA	K	SYN	isocitrate dehydrogenase	Intermediary metabolism and respiration
180025	*Rv0152c*	*PE2*	C	A	291	GGG	G	GAG	E	NONSYN	PE family protein	PE/PPE families
203269	*Rv0172*	*mce1D*	C	T	265	GCG	A	GTG	V	NONSYN	MCE-family protein	Virulence, detoxification and adaptation
207226	*Rv0175*		T	C	138	ATG	M	ACG	T	NONSYN	MCE-associated membrane protein	Cell wall and cell processes
212353	*Rv0181c*		C	A	220	CGC	R	CAC	H	NONSYN	conserved hypothetical protein	Conserved hypothetical
472705	Intergenic	-	T	C								
726703	*Rv0631c*	*recC*	C	T	535	AGG	R	ATG	M	NONSYN	exonuclease V gamma chain	Information pathways
764995	*Rv0668*	*rpoC*	C	G	542	GCC	A	GCG	A	SYN	DNA-directed RNA polymerase beta chain	Information pathways
796509	*Rv0696*		G	T	331	GGC	G	TGC	C	NONSYN	membrane sugar transferase	Intermediary metabolism and respiration
942479	Intergenic	-	T	C								
1142266	*Rv1020*	*mfd*	A	C	1100	CTA	L	CTC	L	SYN	transcription-repair coupling factor	Information pathways
1199547	*Rv1075c*		G	T	275	CCA	P	CTA	L	NONSYN	conserved exported protein	Cell wall and cell processes
1389738	*Rv1248c*	*kgd*	G	T	1088	GAC	D	GAT	D	SYN	2-oxoglutarate dehydrogenase sucA	Intermediary metabolism and respiration
1428506	*Rv1278*		G	T	365	GCC	A	TCC	S	NONSYN	hypothetical protein	Conserved hypothetical

1445474	Intergenic	-	G	A								
1446923	*Rv1292*	*argS*	T	G	182	ATC	I	AGC	S	NONSYN	arginyl-tRNA synthetase	Information pathways
1480024	*Rv1318c*		G	A	267	TTC	F	TTA	L	NONSYN	adenylate cyclase	Intermediary metabolism and respiration
1613960	*Rv1436*	*gap*	G	T	218	GCG	A	GCT	A	SYN	glyceraldehyde 3-phosphate dehydrogenase	Intermediary metabolism and respiration
1692795	Intergenic	-	G	C								
1718761	*Rv1524*		C	T	12	GGC	G	GGT	G	SYN	glycosyltransferase	Intermediary metabolism and respiration
1734994	*Rv1534*		C	T	87	GCC	A	GCT	A	SYN	transcriptional regulator	Regulatory
1947903	*Rv1722*		G	T	15	GTG	V	TTG	L	NONSYN	carboxylase	lipid metabolism
1981056	Intergenic	-	C	T								
2207525	Intergenic	-	C	T								
2283030	*Rv2037c*		A	C	231	ATT	I	ACT	T	NONSYN	conserved membrane protein	Cell wall and cell processes
2484255	*Rv2216*		T	G	210	GCT	A	GCG	A	SYN	conserved hypothetical protein	Conserved hypothetical
2502073	*Rv2228c*		C	A	222	CGG	R	CGA	R	SYN	conserved hypothetical protein	Information pathways
2621058	*Rv2343c*	*dnaG*	G	T	465	CCC	P	CCT	P	SYN	DNA primase	Information pathways
2736434	Intergenic	-	C	A								
2745889	*Rv2446c*		C	A	84	GCC	A	ACC	T	NONSYN	conserved membrane protein	Cell wall and cell processes
2871048	*Rv2551c*		C	A	49	CTG	L	CTA	L	SYN	conserved hypothetical protein	Conserved hypothetical
2881455	*Rv2561*		A	G	16	TAC	Y	TGC	C	NONSYN	conserved hypothetical protein	Conserved hypothetical
2968913	Intergenic	-	T	C								
3077039	*Rv2768c*	*PPE43*	C	T	347	GGC	G	GTC	V	NONSYN	PPE family protein	PE/PPE families
3101119	*Rv2791c*		G	A	155	CGC	R	AGC	S	NONSYN	transposase	Insertion seqs and phages
3165074	*Rv2854*		T	C	308	GTG	V	GCG	A	NONSYN	hypothetical protein	Conserved hypothetical
3175702	*Rv2864c*		T	G	522	ATC	I	GTC	V	NONSYN	penicillin-binding lipoprotein	Cell wall and cell

												processes
3191027	*Rv2881c*	*cdsA*	G	T	199	CTG	L	TTG	L	SYN	membrane phosphatidate cytidylyltransferase	lipid metabolism
3243630	Intergenic	-	G	A								
3426795	*Rv3062*	*ligB*	C	G	404	TCC	S	TCG	S	SYN	ATP-dependent DNA ligase	Information pathways
3429202	*Rv3063*	*cstA*	T	G	654	TAC	Y	GAC	D	NONSYN	carbon starvation protein A	Virulence, detoxification and adaptation
3460986	*Rv3092c*		G	T	250	CCG	P	CTG	L	NONSYN	conserved membrane protein	Cell wall and cell processes
3548641	*Rv3179*		T	C	342	TAT	Y	CAT	H	NONSYN	conserved hypothetical protein	Conserved hypothetical
3594124	*Rv3217c*		C	A	38	GCC	A	ACC	T	NONSYN	conserved membrane protein	Cell wall and cell processes
3610441	*Rv3234c*		C	A	250	CGT	R	CAT	H	NONSYN	conserved hypothetical protein	lipid metabolism
3983271	*Rv3544c*	*fadE28*	T	C	292	ATC	I	CTC	L	NONSYN	acyl-CoA dehydrogenase	lipid metabolism
4075957	*Rv3636*		C	A	69	GCC	A	GAC	D	NONSYN	transposase	Insertion seqs and phages
4135112	*Rv3693*		G	A	129	ATG	M	ATA	I	NONSYN	conserved membrane protein	Cell wall and cell processes
4163944	*Rv3720*		A	G	70	CAT	H	CGT	R	NONSYN	fatty-acid synthase	lipid metabolism
4220174	*Rv3775*	*lipE*	G	A	164	GAC	D	AAC	N	NONSYN	lipase	Intermediary metabolism and respiration
4233299	*Rv3786c*		G	T	100	ACC	T	ATC	I	NONSYN	hypothetical protein	Conserved hypothetical
4340330	*Rv3864*		T	G	21	TTG	L	GTG	V	NONSYN	conserved hypothetical protein	Cell wall and cell processes
4373496	*Rv3889c*		C	C	45	GTG	V	GTC	V	SYN	hypothetical protein	Cell wall and cell processes
4391553	*Rv3906c*		C	A	18	CCG	P	CCA	P	SYN	conserved hypothetical protein	Conserved hypothetical

Ref = reference, Alt = alternative, AA = amino acid, syn = synonymous, non-syn/NONSYN = non-synonymous

Table 3

Position	Gene	Gene name	Ref	Alt	Codon number	Ref codon	Ref AA	Alt codon	Ref AA	Syn/non-syn	Gene description	Functional category
35097	*Rv0032*	*bioF2*	T	C	268	ATC	I	ACC	T	NONSYN	8-amino-7-oxononanoate synthase	Intermediary metabolism and respiration
101727	*Rv0092*	*ctpA*	G	A	382	GGA	G	GAA	E	NONSYN	cation transporter P-type ATPase A	Cell wall and cell processes
176303	*Rv0149*		C	T	202	CAC	H	TAC	Y	NONSYN	quinone oxidoreductase	Intermediary metabolism and respiration
234051	*Rv0197*		G	A	607	CCG	P	CCA	P	SYN	oxidoreductase	Intermediary metabolism and respiration
478358	*Rv0399c*	*lpqK*	C	A	67	GAG	E	AAG	K	NONSYN	lipoprotein	Cell wall and cell processes
557133	*Rv0466*		G	A	226	GTT	V	ATT	I	NONSYN	conserved hypothetical protein	Conserved hypothetical
571943	*Rv0483*	*lprQ*	G	A	78	GTG	V	GTA	V	SYN	lipoprotein	Cell wall and cell processes
667659	*Rv0574c*		C	A	246	GAC	D	AAC	N	NONSYN	conserved hypothetical protein	Conserved hypothetical
733798	*Rv0638*	*secE1*	G	C	21	AGC	S	ACC	T	NONSYN	preprotein translocase	Cell wall and cell processes
790180	*Rv0690c*		A	C	298	GAT	D	GAC	D	SYN	conserved hypothetical protein	Conserved hypothetical
1037355	*Rv0930*	*pstA1*	T	C	119	ACT	T	ACC	T	SYN	phosphate-transport membrane ABC transporter	Cell wall and cell processes
1071797	*Rv0959*		C	G	181	GGC	G	GGG	G	SYN	conserved hypothetical protein	Conserved hypothetical
1281443	*Rv1154c*		C	C	14	AAG	K	AAC	N	NONSYN	hypothetical protein	Conserved hypothetical
1297327	*Rv1166*	*lpqW*	G	A	392	GTG	V	GTA	V	SYN	lipoprotein	Cell wall and cell processes
1297999	*Rv1166*	*lpqW*	T	G	616	TCT	S	TCG	S	SYN	lipoprotein	Cell wall and cell processes
1691799	*Rv1500*		C	T	317	ACC	T	ATC	I	NONSYN	glycosyltransferase	Intermediary metabolism and respiration
1709899	*Rv1518*		A	C	86	AAC	N	CAC	H	NONSYN	conserved hypothetical	Conserved

											protein	hypothetical
1719322	*Rv1524*		G	A	199	CTG	L	CTA	L	SYN	glycosyltransferase	Intermediary metabolism and respiration
1736638	*Rv1536*	*ileS*	C	T	40	CGC	R	CGT	R	SYN	isoleucyl-tRNA synthetase	Information pathways
1847811	*Rv1639c*		A	G	216	CCT	P	CCG	P	SYN	conserved membrane protein	Cell wall and cell processes
1861274	*Rv1650*	*pheT*	G	A	506	CGC	R	CAC	H	NONSYN	phenylalanyl-tRNA synthetase beta chain	Information pathways
2007785	Intergenic	-	C	T								
2126366	*Rv1877*		G	C	155	GTC	V	CTC	L	NONSYN	conserved membrane protein	Cell wall and cell processes
2213265	*Rv1969*	*mce3D*	G	A	137	CGG	R	CGA	R	SYN	MCE-family protein	Virulence, detoxification and adaptation
2257780	*Rv2008c*		T	C	55	ATC	I	CTC	L	NONSYN	conserved hypothetical protein	Conserved hypothetical
2527676	*Rv2252*		G	A	230	GGC	G	AGC	S	NONSYN	conserved hypothetical protein	lipid metabolism
2599821	*Rv2326c*		C	T	43	GCG	A	TCG	S	NONSYN	transmembrane ATP-binding protein ABC transorter	Cell wall and cell processes
2664299	*Rv2380c*	*mbtE*	G	G	939	ACC	T	ACG	T	SYN	peptide synthetase	lipid metabolism
2698585	*Rv2402*		C	T	19	TAC	Y	TAT	Y	SYN	conserved hypothetical protein	Conserved hypothetical
2857014	*Rv2531c*		G	T	256	ACC	T	ACT	T	SYN	amino acid decarboxylase	Intermediary metabolism and respiration
2868659	*Rv2547*		C	G	18	GCC	A	GCG	A	SYN	conserved hypothetical protein	Virulence, detoxification and adaptation
2894854	*Rv2570*		C	T	115	CAG	Q	TAG	_	NONSYN	conserved hypothetical protein	Conserved hypothetical
2946157	*Rv2617c*		T	G	44	AAT	N	AGT	S	NONSYN	transmembrane protein	Cell wall and cell processes
3099269	*Rv2790c*	*ltp1*	A	G	301	TTC	F	GTC	V	NONSYN	lipid-transfer protein	Cell wall and cell processes
3207297	*Rv2897c*		G	A	216	CTG	L	ATG	M	NONSYN	conserved hypothetical protein	Conserved hypothetical
3312632	*Rv2959c*		C	A	69	TGG	W	TGA	_	NONSYN	methyltransferase/methylase	Intermediary metabolism and respiration

3377326	Intergenic	-	G	A								
3874191	Intergenic	-	T	C								
3881187	*Rv3463*		G	A	94	GGC	G	GAC	D	NONSYN	conserved hypothetical protein	Conserved hypothetical
3977226	*Rv3538*		G	A	55	TTG	L	TTA	L	SYN	dehydrogenase	Intermediary metabolism and respiration
4042761	*Rv3598c*	*lysS*	G	T	60	GAC	D	GAT	D	SYN	lysyl-tRNA synthetase 1	Information pathways
4109354	*Rv3667*	*acs*	A	C	521	CTA	L	CTC	L	SYN	acetyl-CoA synthetase	lipid metabolism
4146330	*Rv3703c*		A	C	188	TTG	L	CTG	L	SYN	conserved hypothetical protein	Conserved hypothetical
4311871	*Rv3838c*	*pheA*	G	T	267	CAC	H	TAC	Y	NONSYN	prephenate dehydratase	Intermediary metabolism and respiration
4319652	Intergenic	-	G	A								
4372661	*Rv3888c*		T	G	16	ATC	I	GTC	V	NONSYN	conserved membrane protein	Cell wall and cell processes
4383094	*Rv3897c*		A	C	183	TGT	C	CGT	R	NONSYN	conserved hypothetical protein	Conserved hypothetical
4395964	*Rv3909*		C	A	591	ACC	T	ACA	T	SYN	conserved hypothetical protein	Conserved hypothetical

Ref = reference, Alt = alternative, AA = amino acid, syn = synonymous, non-syn/NONSYN = non-synonymous

Table 4

Position	**Gene**	**Gene name**	**Ref**	**Alt**	**Codon number**	**Ref codon**	**Ref AA**	**Alt codon**	**Ref AA**	**Syn/non-syn**	**Gene description**	**Functional category**
170263	Intergenic	-	G	A								
197047	*Rv0167*	*yrbE1A*	A	G	63	ATG	M	GTG	V	NONSYN	hypothetical membrane protein	Virulence, detoxification and adaptation
338012	*Rv0279c*	*PE_PGRS4*	C	A	354	CTG	L	CTA	L	SYN	PE-PGRS family protein	PE/PPE families
361362	*Rv0297*	*PE_PGRS5*	T	C	10	ATG	M	ACG	T	NONSYN	PE-PGRS family protein	PE/PPE families
393780	*Rv0327c*	*cyp135A1*	T	C	89	GAG	E	GCG	A	NONSYN	cytochrome P450 135A1	Intermediary metabolism and respiration
831206	*Rv0739*		C	T	118	CAG	Q	TAG	_	NONSYN	conserved hypothetical protein	Conserved hypothetical
1002044	*Rv0897c*		A	C	124	GAT	D	GAC	D	SYN	oxidoreductase	Intermediary metabolism and respiration
1106649	*Rv0989c*	*grcC2*	T	G	245	TAC	Y	TGC	C	NONSYN	polyprenyl-diphosphate synthase	Intermediary metabolism and respiration
1150143	Intergenic	-	G	T								
1251955	*Rv1128c*		C	A	340	GCG	A	ACG	T	NONSYN	conserved hypothetical protein	Insertion seqs and phages
1274335	*Rv1146*	*mmpL13b*	G	A	327	TTG	L	TTA	L	SYN	transmembrane transport protein	Cell wall and cell processes
1364434	*Rv1221*	*sigE*	C	T	8	CGG	R	TGG	W	NONSYN	alternative RNA polymerase sigma factor	Information pathways
1439711	*Rv1286*	*cysN*	G	A	269	GCG	A	ACG	T	NONSYN	bifunctional sulfate adenyltransferase/adenylylsulfate kinase cysN/cysC	Intermediary metabolism and respiration
1692824	Intergenic	-	A	G								
1953648	*Rv1727*		T	C	127	TGG	W	CGG	R	NONSYN	conserved hypothetical protein	Conserved hypothetical
2133668	*Rv1883c*		C	T	9	GGT	G	TGT	C	NONSYN	conserved hypothetical protein	Conserved hypothetical
2199684	*Rv1949c*		G	T	117	TTC	F	TTT	F	SYN	conserved hypothetical protein	Conserved hypothetical
2210740	*Rv1967*	*mce3B*	A	C	47	AAC	N	ACC	T	NONSYN	MCE-family protein	Virulence, detoxification and adaptation
2316912	*Rv2061c*		G	T	58	AAC	N	AAT	N	SYN	conserved hypothetical protein	Conserved hypothetical
3118976	*Rv2813*		C	G	251	GAC	D	GAG	E	NONSYN	conserved hypothetical protein	Conserved hypothetical
3244674	*Rv2930*	*fadD26*	G	A	326	CGG	R	CGA	R	SYN	fatty-acid-CoA ligase	lipid metabolism
3285945	*Rv2942*	*mmpL7*	C	G	292	GCC	A	GCG	A	SYN	transmembrane transport protein	Cell wall and cell processes

3558749	Intergenic	-	G	T								
3700783	*Rv3312A*		G	T	78	CCC	P	CCT	P	SYN	secreted protein antigen	Cell wall and cell processes
4068928	*Rv3629c*		G	T	43	CTG	L	TTG	L	SYN	conserved membrane protein	Cell wall and cell processes
4121030	Intergenic	-	G	A								
4288212	Intergenic	-	C	T								
4345420	*Rv3869*		G	T	128	GTT	V	TTT	F	NONSYN	conserved membrane protein	Cell wall and cell processes
4398141	*Rv3910*		G	A	515	TCG	S	TCA	S	SYN	conserved membrane protein	Cell wall and cell processes
4404694	*Rv3916c*		G	T	158	GGC	G	GGT	G	SYN	conserved hypothetical protein	Conserved hypothetical

Ref = reference, Alt = alternative, AA = amino acid, syn = synonymous, non-syn/NONSYN = non-synonymous

Table 5

Position	Gene	Gene name	Ref	Alt	Codon number	Ref codon	Ref AA	Alt codon	Ref AA	Syn/non-syn	Gene description	Functional category
40162	*Rv0037c*		C	A	347	ATG	M	ATA	I	NONSYN	conserved membrane protein	Cell wall and cell processes
282188	*Rv0235c*		G	T	143	CGC	R	TGC	C	NONSYN	conserved membrane protein	Cell wall and cell processes
623021	*Rv0532*	*PE_PGRS6*	G	T	77	GTG	V	TTG	L	NONSYN	PE-PGRS family protein	PE/PPE families
632330	*Rv0539*		G	T	196	CGG	R	CGT	R	SYN	dolichyl-phosphate sugar synthase	Intermediary metabolism and respiration
784440	*Rv0684*	*fusA1*	G	T	652	GCG	A	GCT	A	SYN	elongation factor G	Information pathways
990533	*Rv0890c*		T	G	688	ACA	T	ACG	T	SYN	transcriptional regulator, luxR-family	Regulatory
1055049	*Rv0946c*	*pgi*	C	A	546	CGC	R	CAC	H	NONSYN	glucose-6-phosphate isomerase	Intermediary metabolism and respiration
1132368	*Rv1013*	*pks16*	C	T	248	ACC	T	ACT	T	SYN	polyketide synthase	lipid metabolism
1373170	*Rv1230c*		G	G	343	CCG	P	CGG	R	NONSYN	membrane protein	Cell wall and cell processes
1724120	*Rv1527c*	*pks5*	G	T	1430	GAC	D	GAT	D	SYN	polyketide synthase	lipid metabolism
1755599	*Rv1551*	*plsB1*	C	T	52	GCC	A	GTC	V	NONSYN	acyltransferase	lipid metabolism
1769099	*Rv1563c*	*treY*	C	T	112	GAT	D	TAT	Y	NONSYN	maltooligosyltrehalose synthase	Virulence, detoxification and adaptation
1771320	*Rv1564c*	*treX*	G	T	94	GAC	D	GAT	D	SYN	maltooligosyltrehalose synthase	Virulence, detoxification and adaptation
2019942	*Rv1783*		A	G	229	CAG	Q	CGG	R	NONSYN	conserved membrane protein	Cell wall and cell processes
2084526	*Rv1836c*		G	T	37	CCC	P	CCT	P	SYN	conserved hypothetical protein	Conserved hypothetical
2278442	*Rv2030c*		C	C	15	CGG	R	CGC	R	SYN	conserved hypothetical protein	Conserved hypothetical
2327492	*Rv2069*	*sigC*	C	T	183	CTC	L	CTT	L	SYN	RNA polymerase sigma factor	Information pathways
2385695	*Rv2124c*	*metH*	C	A	125	GGG	G	AGG	R	NONSYN	5-methyltetrahydrofolate-	Intermediary metabolism and

											homocystein methyltransferase	respiration
2420535	*Rv2158c*	*murE*	C	T	25	GGC	G	GTC	V	NONSYN	UDP-N-acetylmuramoylalanyl-D-glutamate-2,6-diaminopimelat E ligase	Cell wall and cell processes
2443188	*Rv2180c*		G	T	9	CAC	H	CAT	H	SYN	conserved membrane protein	Cell wall and cell processes
3006767	*Rv2689c*		A	G	99	CTT	L	CGT	R	NONSYN	conserved alanine, valine and glycine rich protein	Conserved hypothetical
3089679	*Rv2782c*	*pepR*	G	T	228	CCA	P	CTA	L	NONSYN	zinc protease	Intermediary metabolism and respiration
3158935	*Rv2850c*		G	G	374	CGC	R	GGC	G	NONSYN	magnesium chelatase	Intermediary metabolism and respiration
3267743	*Rv2935*	*ppsE*	A	G	3	ATC	I	GTC	V	NONSYN	phenolpthiocerol synthesis type-I polyketide synthase	lipid metabolism
3504930	*Rv3138*	*pflA*	C	T	246	CTG	L	TTG	L	SYN	pyruvate formate lyase activating protein	Intermediary metabolism and respiration
3514512	*Rv3148*	*nuoD*	G	C	392	GGT	G	GCT	A	NONSYN	NADH dehydrogenase I chain D	Intermediary metabolism and respiration
3561155	*Rv3193c*		G	T	673	GCC	A	GTC	V	NONSYN	conserved membrane protein	Cell wall and cell processes
3847378	*Rv3429*	*PPE59*	G	C	72	GAC	D	CAC	H	NONSYN	PPE family protein	PE/PPE families
3847380	*Rv3429*	*PPE59*	C	A	72	GAC	D	GAA	E	NONSYN	PPE family protein	PE/PPE families
4043365	*Rv3600c*		G	A	165	GCC	A	GCA	A	SYN	conserved hypothetical protein	Conserved hypothetical
4218350	*Rv3773c*		T	G	159	AAG	K	AGG	R	NONSYN	conserved hypothetical protein	Conserved hypothetical
4288405	*Rv3823c*	*mmpL8*	G	A	1042	GCA	A	GAA	E	NONSYN	membrane transport protein	Cell wall and cell processes
4359195	*Rv3879c*		G	T	196	GGC	G	GGT	G	SYN	hypothetical alanine and proline rich protein	Cell wall and cell processes
4377447	*Rv3894c*		G	T	1002	GAC	D	GAT	D	SYN	conserved membrane protein	Cell wall and cell processes

4403900	*Rv3915*	A	G	237	ATG	M	GTG	V	NONSYN	hydrolase	Intermediary metabolism and respiration

Ref = reference, Alt = alternative, AA = amino acid, syn = synonymous, non-syn/NONSYN = non-synonymous

Table 6

Position	Gene	Gene name	Ref	Alt	Codon number	Ref codon	Ref AA	Alt codon	Ref AA	Syn/non-syn	Gene description	Functional category
8040	*Rv0006*	*gyrA*	G	A	247	GGC	G	AGC	S	NONSYN	DNA gyrase subunit A	Information pathways
387353	*Rv0319*	*pcp*	G	A	69	GGC	G	GAC	D	NONSYN	pyrrolidone-carboxylate peptidase	Intermediary metabolism and respiration
403364	*Rv0338c*		G	T	826	CCC	P	CCT	P	SYN	iron-sulfur-binding reductase	Intermediary metabolism and respiration
403920	*Rv0338c*		C	A	641	CGC	R	CAC	H	NONSYN	iron-sulfur-binding reductase	Intermediary metabolism and respiration
479632	*Rv0400c*	*fadE7*	G	A	41	ACC	T	AAC	N	NONSYN	acyl-CoA dehydrogenase	lipid metabolism
535695	*Rv0447c*	*ufaA1*	C	A	271	GGG	G	GGA	G	SYN	cyclopropane-fatty-acyl-phospholipid synthase	lipid metabolism
558501	*Rv0467*	*icl*	C	T	325	GAC	D	GAT	D	SYN	isocitrate lyase	Intermediary metabolism and respiration
752802	Intergenic	-	A	C								
812808	*Rv0717*	*rpsN1*	G	A	61	TGG	W	TAG	_	NONSYN	30S ribosomal protein S14	Information pathways
919393	*Rv0825c*		G	G	54	ACA	T	AGA	R	NONSYN	conserved hypothetical protein	Conserved hypothetical
1087279	*Rv0974c*	*accD2*	T	G	23	AAG	K	GAG	E	NONSYN	acetyl-/propionyl-CoA carboxylase beta subunit	lipid metabolism
1144664	*Rv1023*	*eno*	G	A	34	CGG	R	CAG	Q	NONSYN	enolase	Intermediary metabolism and respiration
1162274	*Rv1039c*	*PPE15*	C	A	67	GCC	A	ACC	T	NONSYN	PPE family protein	PE/PPE families
1519847	*Rv1353c*		C	C	47	GGC	G	CGC	R	NONSYN	transcriptional regulator	Regulatory
1546530	*Rv1373*		G	A	173	GAG	E	GAA	E	SYN	glycolipid sulfotransferase	Intermediary metabolism and respiration
1651306	*Rv1463*		C	T	197	GCC	A	GCT	A	SYN	ABC transporter ATP-binding protein	Cell wall and cell processes
1797027	*Rv1595*	*nadB*	C	T	408	ACC	T	ATC	I	NONSYN	L-aspartate oxidase	Intermediary metabolism and respiration
1815604	*Rv1615*		G	A	118	GCC	A	ACC	T	NONSYN	hypothetical membrane protein	Cell wall and cell processes
2037716	*Rv1798*		T	C	339	CGT	R	CGC	R	SYN	conserved hypothetical protein	Cell wall and cell processes
2077253	*Rv1832*	*gcvB*	G	A	459	ACG	T	ACA	T	SYN	glycine dehydrogenase	Intermediary metabolism and respiration
2108838	Intergenic	-	C	T								

2144617	*Rv1897c*		G	T	89	CCG	P	CTG	L	NONSYN	conserved hypothetical protein	Conserved hypothetical
2315669	*Rv2059*		G	A	166	GTC	V	ATC	I	NONSYN	conserved hypothetical protein	Conserved hypothetical
2401883	Intergenic	-	C	T								
2518919	*Rv2245*	*kasA*	G	A	269	GGT	G	AGT	S	NONSYN	3-oxoacyl-[acyl-carrier protein] synthase 1	lipid metabolism
2814961	*Rv2501c*	*accA1*	G	T	640	GGC	G	GGT	G	SYN	acetyl-/propionyl-CoA carboxylase alpha subunit	lipid metabolism
3004091	*Rv2687c*		G	T	220	CGG	R	TGG	W	NONSYN	antibiotic-transport membrane leucine and valine rich protein ABC transporter	Cell wall and cell processes
3061615	*Rv2748c*	*ftsK*	T	G	298	ATG	M	GTG	V	NONSYN	cell division transmembrane protein	Cell wall and cell processes
3351926	*Rv2994*		T	C	220	TCG	S	CCG	P	NONSYN	conserved membrane protein	Cell wall and cell processes
3423184	*Rv3060c*		C	A	10	ACG	T	ACA	T	SYN	transcriptional regulator, gntR-family	Regulatory
3894032	Intergenic	-	T	A								
3973954	*Rv3535c*		C	A	183	GGG	G	AGG	R	NONSYN	acetaldehyde dehydrogenase	Intermediary metabolism and respiration
4316046	*Rv3843c*		A	C	184	GTC	V	GCC	A	NONSYN	conserved membrane protein	Cell wall and cell processes
4363069	*Rv3882c*		T	G	118	ATC	I	GTC	V	NONSYN	conserved membrane protein	Cell wall and cell processes
4376098	Intergenic	-	G	A								
4385530	*Rv3900c*		G	G	260	CCA	P	CGA	R	NONSYN	conserved alanine rich protein	Conserved hypothetical
4390753	*Rv3905c*	*esxF*	C	C	93	TCG	S	TCC	S	SYN	Esat-6 like protein	Cell wall and cell processes

Ref = reference, Alt = alternative, AA = amino acid, syn = synonymous, non-syn/NONSYN = non-synonymous

Table 7

Position	Gene	Gene name	Ref	Alt	Codon number	Ref codon	Ref AA	Alt codon	Ref AA	Syn/non-syn	Gene description	Functional category
12061	*Rv0008c*		G	G	84	GCC	A	GGC	G	NONSYN	membrane protein	Cell wall and cell processes
19185	*Rv0016c*	*pbpA*	T	C	350	GCA	A	GCC	A	SYN	penicillin-binding protein	Cell wall and cell processes
52856	*Rv0049*		C	G	9	TCC	S	TGC	C	NONSYN	conserved hypothetical protein	Conserved hypothetical
86811	*Rv0078*		C	T	95	GCA	A	GTA	V	NONSYN	transcriptional regulator	Regulatory
119689	*Rv0102*		T	G	659	ATG	M	AGG	R	NONSYN	conserved membrane protein	Cell wall and cell processes
191766	*Rv0161*		C	T	387	GCG	A	GTG	V	NONSYN	oxidoreductase	Intermediary metabolism and respiration
227923	*Rv0194*		T	C	349	GTG	V	GCG	A	NONSYN	drugs-transport transmembrane ATP-binding protein ABC transporter	Cell wall and cell processes
236412	*Rv0198c*		C	T	32	TGG	W	TGT	C	NONSYN	zinc metalloprotease	Intermediary metabolism and respiration
294183	*Rv0244c*	*fadE5*	C	A	484	GGT	G	GAT	D	NONSYN	acyl-CoA dehydrogenase	lipid metabolism
364984	*Rv0302*		C	T	127	GCG	A	GTG	V	NONSYN	transcriptional regulator, tetR/acrR-family	Regulatory
402161	*Rv0337c*	*aspC*	C	A	334	CAG	Q	CAA	Q	SYN	aspartate aminotransferase	Intermediary metabolism and respiration
429936	*Rv0355c*	*PPE8*	G	T	1582	CAG	Q	TAG	_	NONSYN	PPE family protein	PE/PPE families
501957	*Rv0415*	*thiO*	T	C	270	GAT	D	GAC	D	SYN	thiamine biosynthesis oxidoreductase	Intermediary metabolism and respiration
527770	*Rv0439c*		G	A	182	GCC	A	GAC	D	NONSYN	dehydrogenase/reductase	Intermediary metabolism and respiration
626865	*Rv0535*	*pnp*	C	T	137	CTG	L	TTG	L	SYN	5-methylthioadenosine phosphorylase	Intermediary metabolism and respiration
629571	*Rv0537c*		G	T	54	ACA	T	ATA	I	NONSYN	membrane protein	Cell wall and cell processes
654508	*Rv0563*	*htpX*	C	T	210	GAC	D	GAT	D	SYN	protease transmembrane protein heat shock protein	Virulence, detoxification and adaptation
661901	*Rv0570*	*nrdZ*	C	T	203	CAA	Q	TAA	_	NONSYN	ribonucleoside-	Information pathways

											diphosphate reductase large subunit	
666545	*Rv0573c*		T	T	233	GAC	D	GTC	V	NONSYN	conserved hypothetical protein	Intermediary metabolism and respiration
668666	*Rv0575c*		C	A	360	GCG	A	GCA	A	SYN	oxidoreductase	Intermediary metabolism and respiration
693403	*Rv0594*	*mce2F*	C	T	56	GCC	A	GTC	V	NONSYN	MCE-family protein	Virulence, detoxification and adaptation
696089	*Rv0597c*		A	C	272	GTG	V	GCG	A	NONSYN	conserved hypothetical protein	Conserved hypothetical
696362	*Rv0597c*		C	A	181	CGA	R	CAA	Q	NONSYN	conserved hypothetical protein	Conserved hypothetical
707548	*Rv0613c*		C	T	656	CCG	P	CCT	P	SYN	hypothetical protein	Conserved hypothetical
744571	*Rv0648*		G	A	618	GGC	G	GAC	D	NONSYN	alpha-mannosidase	Intermediary metabolism and respiration
779675	*Rv0679c*		C	A	122	GCG	A	GCA	A	SYN	conserved threonine rich protein	Conserved hypothetical
907124	*Rv0812*		C	T	234	TAC	Y	TAT	Y	SYN	amino acid aminotransferase	Intermediary metabolism and respiration
913355	*Rv0820*	*phoT*	T	C	210	GCT	A	GCC	A	SYN	phosphate-transport ATP-binding protein ABC transporter	Cell wall and cell processes
932523	*Rv0836c*		T	G	137	GAC	D	GGC	G	NONSYN	hypothetical protein	Conserved hypothetical
951237	*Rv0854*		C	T	19	CTG	L	TTG	L	SYN	conserved hypothetical protein	Conserved hypothetical
978545	*Rv0879c*		G	G	71	GCC	A	GGC	G	NONSYN	conserved membrane protein	Cell wall and cell processes
1001680	*Rv0897c*		G	T	246	CTG	L	TTG	L	SYN	oxidoreductase	Intermediary metabolism and respiration
1063397	*Rv0952*	*sucD*	G	A	86	CCG	P	CCA	P	SYN	succinyl-CoA synthetase alpha chain	Intermediary metabolism and respiration
1133442	*Rv1014c*	*pth*	G	A	156	CCG	P	CAG	Q	NONSYN	peptidyl-tRNA hydrolase	Intermediary metabolism and respiration
1154958	*Rv1030*	*kdpB*	T	G	412	ATC	I	AGC	S	NONSYN	potassium-transporting ATPase B chain	Cell wall and cell processes
1177688	*Rv1056*		C	T	21	CGA	R	TGA	_	NONSYN	conserved hypothetical protein	Conserved hypothetical
1246372	*Rv1123c*	*bpoB*	C	A	227	CAG	Q	CAA	Q	SYN	peroxidase	Virulence, detoxification and adaptation
1263722	Intergenic	-	G	T								
1382804	*Rv1239c*	*corA*	T	G	80	GAA	E	GGA	G	NONSYN	magnesium and cobalt	Cell wall and cell

											transport transmembrane protein	processes
1384034	*Rv1240*	*mdh*	C	T	274	TAC	Y	TAT	Y	SYN	malate dehydrogenase	Intermediary metabolism and respiration
1394772	*Rv1250*		C	A	198	CCC	P	CCA	P	SYN	drug-transport membrane protein	Cell wall and cell processes
1399635	*Rv1252c*	*lprE*	T	G	90	AAA	K	AAG	K	SYN	lipoprotein	Cell wall and cell processes
1437929	*Rv1285*	*cysD*	G	A	7	ATG	M	ATA	I	NONSYN	sulfate adenylyltransferase subunit 2	Intermediary metabolism and respiration
1452590	*Rv1296*	*thrB*	G	A	198	GTG	V	GTA	V	SYN	homoserine kinase	Intermediary metabolism and respiration
1599431	*Rv1424c*		G	T	75	CCG	P	CTG	L	NONSYN	membrane protein	Cell wall and cell processes
1628121	*Rv1449c*	*tkt*	G	T	693	GCC	A	GCT	A	SYN	transketolase	Intermediary metabolism and respiration
1648024	*Rv1461*		A	G	346	ACC	T	GCC	A	NONSYN	conserved hypothetical protein	Conserved hypothetical
1655718	*Rv1468c*	*PE_PGRS29*	T	G	335	CAG	Q	CGG	R	NONSYN	PE-PGRS family protein	PE/PPE families
1720865	*Rv1526c*		G	T	399	GCC	A	GTC	V	NONSYN	glycosyltransferase	Intermediary metabolism and respiration
1723708	*Rv1527c*	*pks5*	T	G	1568	ACC	T	GCC	A	NONSYN	polyketide synthase	lipid metabolism
1744607	*Rv1542c*	*glbN*	A	C	77	ATG	M	ACG	T	NONSYN	hemoglobin	Intermediary metabolism and respiration
1795283	*Rv1594*	*nadA*	T	C	176	TGT	C	TGC	C	SYN	quinolinate synthetase	Intermediary metabolism and respiration
1823150	*Rv1621c*	*cydD*	G	G	42	CCA	P	GCA	A	NONSYN	transmembrane ATP-binding protein ABC transporter	Intermediary metabolism and respiration
1865421	Intergenic	-	T	C								
1917190	*Rv1692*		G	A	165	GGC	G	AGC	S	NONSYN	phosphatase	Intermediary metabolism and respiration
1925758	*Rv1700*		G	A	59	ATG	M	ATA	I	NONSYN	conserved hypothetical protein	Information pathways
1948374	*Rv1722*		T	G	172	TCG	S	GCG	A	NONSYN	carboxylase	lipid metabolism
2050007	*Rv1808*	*PPE32*	G	A	29	GCG	A	GCA	A	SYN	PPE family protein	PE/PPE families
2092031	*Rv1842c*		C	A	19	GGC	G	AGC	S	NONSYN	conserved membrane protein	Cell wall and cell processes
2107515	*Rv1859*	*modC*	C	T	314	GTC	V	GTT	V	SYN	molybdenum-transport ATP-binding protein ABC transporter	Cell wall and cell processes

2125239	Intergenic	-	C	T								
2129505	*Rv1879*		C	T	43	GCC	A	GCT	A	SYN	conserved hypothetical protein	Conserved hypothetical
2187314	Intergenic	-	T	C								
2209726	*Rv1966*	*mce3A*	G	A	134	GCG	A	ACG	T	NONSYN	MCE-family protein	Virulence, detoxification and adaptation
2309071	*Rv2051c*	*ppm1*	G	G	562	GCC	A	GGC	G	NONSYN	polyprenol-monophosphomannose synthase	Cell wall and cell processes
2378805	*Rv2119*		G	A	140	TTG	L	TTA	L	SYN	conserved hypothetical protein	Conserved hypothetical
2382238	*Rv2123*	*PPE37*	G	A	390	GCA	A	ACA	T	NONSYN	PPE family protein	PE/PPE families
2388969	*Rv2127*	*ansP1*	C	T	118	ACC	T	ACT	T	SYN	L-asparagine permease	Cell wall and cell processes
2424422	*Rv2162c*	*PE_PGRS38*	A	C	139	GGT	G	GGC	G	SYN	PE-PGRS family protein	PE/PPE families
2424425	*Rv2162c*	*PE_PGRS38*	A	C	138	GGT	G	GGC	G	SYN	PE-PGRS family protein	PE/PPE families
2448663	*Rv2187*	*fadD15*	G	C	168	CCG	P	CCC	P	SYN	long-chain fatty-acid-CoA ligase	lipid metabolism
2469415	*Rv2205c*		C	T	350	GGG	G	GTG	V	NONSYN	conserved hypothetical protein	Conserved hypothetical
2523283	*Rv2249c*	*glpD1*	G	T	503	CGC	R	CGT	R	SYN	glycerol-3-phosphate dehydrogenase	Intermediary metabolism and respiration
2578942	Intergenic	-	C	G								
2604872	*Rv2331A*		A	C	45	AGG	R	CGG	R	SYN	hypothetical protein	Conserved hypothetical
2653956	*Rv2373c*	*dnaJ2*	A	C	11	GTG	V	GCG	A	NONSYN	chaperone protein	Virulence, detoxification and adaptation
2715432	Intergenic	-	G	A								
2794377	*Rv2486*	*echA14*	A	C	10	AGC	S	CGC	R	NONSYN	enoyl-CoA hydratase	lipid metabolism
2836124	*Rv2519*	*PE26*	G	A	114	GAG	E	AAG	K	NONSYN	PE family protein	PE/PPE families
2857169	*Rv2531c*		G	T	205	CGG	R	TGG	W	NONSYN	amino acid decarboxylase	Intermediary metabolism and respiration
2890577	*Rv2567*		G	A	261	GTG	V	GTA	V	SYN	conserved alanine and leucine rich protein	Conserved hypothetical
2904696	Intergenic	-	T	C								
2955638	Intergenic	-	C	T								
2957961	*Rv2631*		C	T	130	GAC	D	GAT	D	SYN	conserved hypothetical protein	Conserved hypothetical
2974895	*Rv2650c*		G	T	114	CTG	L	TTG	L	SYN	phiRv2 phage protein	Insertion seqs and phages

2987533	*Rv2671*	*ribD*	G	C	232	GGG	G	GCG	A	NONSYN	bifunctional riboflavin biosynthesis protein	Intermediary metabolism and respiration
3016313	*Rv2701c*	*suhB*	A	G	141	GCT	A	GCG	A	SYN	extragenic suppressor protein	Intermediary metabolism and respiration
3201547	*Rv2892c*	*PPE45*	C	A	158	GCG	A	GCA	A	SYN	PPE family protein	PE/PPE families
3208017	*Rv2898c*		A	C	104	GAT	D	GAC	D	SYN	conserved hypothetical protein	Conserved hypothetical
3360508	*Rv3001c*	*ilvC*	G	T	27	CAC	H	TAC	Y	NONSYN	ketol-acid reductoisomerase	Intermediary metabolism and respiration
3375293	*Rv3015c*		C	A	124	CTG	L	CTA	L	SYN	conserved hypothetical protein	Conserved hypothetical
3380083	*Rv3021c*	*PPE47*	C	C	124	GAA	E	CAA	Q	NONSYN	PPE family protein	PE/PPE families
3397503	*Rv3037c*		T	C	263	CAG	Q	CCG	P	NONSYN	conserved hypothetical protein	Conserved hypothetical
3411980	Intergenic	-	A	G								
3417213	*Rv3056*	*dinP*	T	G	170	CTC	L	CGC	R	NONSYN	DNA-damage-inducible protein P	Information pathways
3539905	*Rv3171c*	*hpx*	G	T	281	CTG	L	TTG	L	SYN	non-heme haloperoxidase	Virulence, detoxification and adaptation
3588423	*Rv3211*	*rhlE*	T	A	209	TTT	F	TAT	Y	NONSYN	ATP-dependent RNA helicase	Information pathways
3610134	*Rv3233c*		C	T	80	GCC	A	TCC	S	NONSYN	conserved hypothetical protein	lipid metabolism
3627602	*Rv3247c*	*tmk*	G	G	154	GCC	A	GCG	A	SYN	thymidylate kinase	Intermediary metabolism and respiration
3706620	*Rv3318*	*sdhA*	C	G	541	CCC	P	GCC	A	NONSYN	succinate dehydrogenase flavoprotein subunit	Intermediary metabolism and respiration
3746187	*Rv3347c*	*PPE55*	G	T	2333	ACG	T	ATG	M	NONSYN	PPE family protein	PE/PPE families
3769440	*Rv3354*		A	C	110	GCA	A	GCC	A	SYN	conserved hypothetical protein	Conserved hypothetical
3769771	*Rv3355c*		C	C	13	GCC	A	CCC	P	NONSYN	conserved hypothetical protein	Cell wall and cell processes
3823862	*Rv3403c*		A	C	1	ATG	M	ACG	T	NONSYN	hypothetical protein	Conserved hypothetical
3842461	*Rv3425*	*PPE57*	T	A	75	TCT	S	ACT	T	NONSYN	PPE family protein	PE/PPE families
3843433	*Rv3426*	*PPE58*	G	T	133	AGA	R	ATA	I	NONSYN	PPE family protein	PE/PPE families
3858312	*Rv3439c*		C	T	451	GGA	G	TGA	_	NONSYN	conserved alanine and proline rich protein	Conserved hypothetical
3886136	*Rv3469c*	*mhpE*	A	C	316	AGT	S	AGC	S	SYN	4-hydroxy-2-oxovalerate aldolase	Intermediary metabolism and respiration
3923531	*Rv3504*	*fadE26*	A	T	354	AAT	N	ATT	I	NONSYN	acyl-CoA dehydrogenase	lipid metabolism

3969205	Intergenic	-	G	A								
4046289	Intergenic	-	C	A								
4207533	*Rv3762c*		C	A	448	GCG	A	GCA	A	SYN	hydrolase	Intermediary metabolism and respiration
4219577	*Rv3774*	*echA21*	C	T	243	GTC	V	GTT	V	SYN	enoyl-CoA hydratase	lipid metabolism
4230850	*Rv3784*		A	G	199	ATC	I	GTC	V	NONSYN	dTDP-glucose 4,6-dehydratase	Intermediary metabolism and respiration
4233817	*Rv3787c*		C	A	240	TGG	W	TGA	_	NONSYN	conserved hypothetical protein	Conserved hypothetical
4247209	*Rv3795*	*embB*	A	G	232	GCA	A	GCG	A	SYN	membrane indolylacetylinositol arabinosyltransferase	Cell wall and cell processes
4304481	*Rv3829c*		G	T	176	GAC	D	GAT	D	SYN	dehydrogenase	Intermediary metabolism and respiration
4405182	Intergenic	-	C	T								
4409137	*Rv3921c*		T	G	311	CCA	P	CCG	P	SYN	conserved membrane protein	Cell wall and cell processes

Ref = reference, Alt = alternative, AA = amino acid, syn = synonymous, non-syn/NONSYN = non-synonymous

Table 8

Position	Gene	Gene name	Ref	Alt	Codon number	Ref codon	Ref AA	Alt codon	Ref AA	Syn/non-syn	Gene description	Functional category
4057146	Intergenic	-	C	T								
4096985	*Rv3659c*		T	C	337	GAC	D	GCC	A	NONSYN	conserved hypothetical protein	Conserved hypothetical
4101905	*Rv3662c*		G	G	44	GCC	A	GGC	G	NONSYN	conserved hypothetical protein	Conserved hypothetical
4156467	*Rv3711c*	*dnaQ*	A	C	88	GTC	V	GCC	A	NONSYN	DNA polymerase III epsilon subunit	Information pathways
4175390	*Rv3728*		C	T	173	GCG	A	GTG	V	NONSYN	conserved membrane protein	Cell wall and cell processes
4209511	*Rv3763*	*lpqH*	G	C	155	GAG	E	GAC	D	NONSYN	19 kda lipoprotein antigen precursor	Cell wall and cell processes
4325029	Intergenic	-	T	C								
4371779	*Rv3888c*		C	T	310	GTG	V	TTG	L	NONSYN	conserved membrane protein	Cell wall and cell processes
4384930	*Rv3899c*		C	C	150	CTG	L	CTC	L	SYN	conserved hypothetical protein	Conserved hypothetical
4388606	*Rv3903c*		G	T	609	ACC	T	ACT	T	SYN	hypothetical alanine and proline rich protein	Conserved hypothetical
4394450	*Rv3909*		G	C	87	GTC	V	CTC	L	NONSYN	conserved hypothetical protein	Conserved hypothetical
4397368	*Rv3910*		A	C	258	ATG	M	CTG	L	NONSYN	conserved membrane protein	Cell wall and cell processes

Ref = reference, Alt = alternative, AA = amino acid, syn = synonymous, non-syn/NONSYN = non-synonymous

Table 9

Position	**Gene**	**Gene name**	**Ref**	**Alt**	**Codon number**	**Ref codon**	**Ref AA**	**Alt codon**	**Ref AA**	**Syn/non-syn**	**Gene description**	**Functional category**
88486	*Rv0079*		G	C	95	GAT	D	CAT	H	NONSYN	hypothetical protein	Conserved hypothetical
111360	*Rv0101*	*nrp*	T	C	454	TCC	S	CCC	P	NONSYN	peptide synthetase	lipid metabolism
485409	*Rv0404*	*fadD30*	C	A	478	CCC	P	CAC	H	NONSYN	fatty-acid-CoA ligase	lipid metabolism
559052	*Rv0468*	*fadB2*	A	C	53	GAG	E	GCG	A	NONSYN	3-hydroxybutyryl-CoA dehydrogenase	lipid metabolism
637871	*Rv0546c*		T	G	33	GAA	E	GAG	E	SYN	conserved hypothetical protein	Conserved hypothetical
751947	*Rv0655*	*mkl*	T	C	144	GTC	V	GCC	A	NONSYN	ribonucleotide-transport ATP-binding protein ABC transporter	Cell wall and cell processes
754198	*Rv0658c*		C	A	71	CGC	R	CAC	H	NONSYN	conserved membrane protein	Cell wall and cell processes
817856	*Rv0724A*		C	C	4	CGG	R	CCG	P	NONSYN	conserved hypothetical protein	Conserved hypothetical
861283	*Rv0768*	*aldA*	C	T	124	TCC	S	TCT	S	SYN	aldehyde dehydrogenase NAD-dependent	Intermediary metabolism and respiration
1083576	*Rv0972c*	*fadE12*	G	A	59	CCC	P	ACC	T	NONSYN	acyl-CoA dehydrogenase	lipid metabolism
1264381	*Rv1135A*		A	G	23	TAC	Y	TGC	C	NONSYN	acetyl-CoA acetyltransferase	lipid metabolism
1295418	*Rv1165*	*typA*	G	A	417	CTG	L	CTA	L	SYN	GTP-binding translation elongation factor	Information pathways
1325165	*Rv1184c*		C	A	149	CAG	Q	CAA	Q	SYN	hypothetical exported protein	Cell wall and cell processes
1332535	*Rv1189*	*sigI*	C	A	148	GCC	A	GCA	A	SYN	alternative RNA polymerase sigma factor	Information pathways
1476056	Intergenic	-	G	A								
1489708	*Rv1325c*	*PE_PGRS24*	C	A	86	GCG	A	GCA	A	SYN	PE-PGRS family protein	PE/PPE families
1540976	Intergenic	-	C	T								
1791607	*Rv1591*		C	T	13	CCT	P	CTT	L	NONSYN	transmembrane protein	Cell wall and cell processes
1848184	*Rv1639c*		C	A	92	TGG	W	TAG	_	NONSYN	conserved membrane protein	Cell wall and cell processes
1864767	*Rv1651c*	*PE_PGRS30*	C	C	206	GGG	G	CGG	R	NONSYN	PE-PGRS family protein	PE/PPE families

1929252	*Rv1703c*		T	C	157	AAG	K	ACG	T	NONSYN	catechol-O-methyltransferase	Intermediary metabolism and respiration
1968640	*Rv1742*		T	A	156	ATT	I	ATA	I	SYN	hypothetical protein	Hypothetical
1993334	*Rv1760*		T	C	61	TTG	L	TCG	S	NONSYN	conserved hypothetical protein	lipid metabolism
2049666	Intergenic	-	T	G								
2175487	*Rv1923*	*lipD*	A	C	105	AAA	K	AAC	N	NONSYN	lipase	Intermediary metabolism and respiration
2210644	*Rv1967*	*mce3B*	T	C	15	TTC	F	TCC	S	NONSYN	MCE-family protein	Virulence, detoxification and adaptation
2223051	*Rv1979c*		C	C	38	GAG	E	GAC	D	NONSYN	permease	Cell wall and cell processes
2390534	*Rv2129c*		T	G	219	AAC	N	AGC	S	NONSYN	oxidoreductase	Intermediary metabolism and respiration
2418211	*Rv2157c*	*murF*	A	C	265	GTG	V	GCG	A	NONSYN	UDP-N-acetylmuramoylalanyl-D-glutamyl-2,6-diaminopimelate-D-alanyl-D-alanyl ligase	Cell wall and cell processes
2569403	*Rv2298*		C	A	108	CTG	L	ATG	M	NONSYN	conserved hypothetical protein	Conserved hypothetical
2575350	Intergenic	-	T	G								
2581689	Intergenic	-	A	C								
2767371	*Rv2464c*		C	A	99	GCA	A	ACA	T	NONSYN	DNA glycosylase	Information pathways
2812703	*Rv2498c*	*citE*	C	A	158	GTG	V	GTA	V	SYN	citrate (pro-3s)-lyase beta subunit	Intermediary metabolism and respiration
2903463	*Rv2578c*		C	A	23	TTG	L	TTA	L	SYN	conserved hypothetical protein	Conserved hypothetical
2969308	Intergenic	-	A	C								
3065824	*Rv2752c*		G	T	123	CCG	P	CTG	L	NONSYN	conserved hypothetical protein	Conserved hypothetical
3132784	Intergenic	-	G	A								
3169584	*Rv2858c*	*aldC*	C	A	380	GCG	A	GCA	A	SYN	aldehyde dehydrogenase	Intermediary metabolism and respiration
3238703	*Rv2924c*	*fpg*	T	C	256	GAA	E	GAC	D	NONSYN	formamidopyrimidine-DNA glycosylase	Information pathways
3333516	*Rv2977c*	*thiL*	C	C	91	GAG	E	GAC	D	NONSYN	thiamine-monophosphate kinase	Intermediary metabolism and respiration
3420056	*Rv3059*	*cyp136*	G	A	189	GTC	V	ATC	I	NONSYN	cytochrome P450 136	Intermediary metabolism and respiration
3537404	*Rv3169*		A	G	56	AAC	N	AGC	S	NONSYN	conserved hypothetical	Conserved hypothetical

											protein	
3581314	Intergenic	-	C	T								
3588184	*Rv3211*	*rhlE*	G	A	129	GTG	V	GTA	V	SYN	ATP-dependent RNA helicase	Information pathways
3620280	*Rv3240c*	*secA1*	G	T	84	GAC	D	GAT	D	SYN	preprotein translocase	Cell wall and cell processes
3633957	*Rv3254*		A	C	95	ACG	T	CCG	P	NONSYN	conserved hypothetical protein	Conserved hypothetical
3750163	*Rv3347c*	*PPE55*	A	C	1008	TTG	L	CTG	L	SYN	PPE family protein	PE/PPE families
3938660	*Rv3510c*		A	C	200	TAC	Y	CAC	H	NONSYN	conserved hypothetical protein	Conserved hypothetical
3959536	*Rv3523*	*ltp3*	G	A	3	GGA	G	GAA	E	NONSYN	lipid carrier protein or keto acyl-CoA thiolase	lipid metabolism
3998613	*Rv3558*	*PPE64*	C	T	212	CTC	L	TTC	F	NONSYN	PPE family protein	PE/PPE families
4072240	*Rv3633*		G	A	150	GCG	A	GCA	A	SYN	conserved hypothetical protein	Conserved hypothetical
4099344	Intergenic	-	T	C								
4156279	*Rv3711c*	*dnaQ*	C	C	151	GGT	G	CGT	R	NONSYN	DNA polymerase III epsilon subunit	Information pathways
4163246	*Rv3719*		G	C	314	AGG	R	ACG	T	NONSYN	conserved hypothetical protein	Conserved hypothetical
4177264	*Rv3728*		C	T	798	CGT	R	TGT	C	NONSYN	conserved membrane protein	Cell wall and cell processes
4399385	*Rv3910*		C	T	930	GCG	A	GTG	V	NONSYN	conserved membrane protein	Cell wall and cell processes

Ref = reference, Alt = alternative, AA = amino acid, syn = synonymous, non-syn/NONSYN = non-synonymous

Table 10

Position	Gene	Gene name	Ref	Alt	Codon number	Ref codon	Ref AA	Alt codon	Ref AA	Syn/non-syn	Gene description	Functional category
41257	Intergenic	-	A	G								
47699	*Rv0043c*		G	T	134	CGC	R	CGT	R	SYN	transcriptional regulator, gntR-family	Regulatory
69025	*Rv0064*		G	C	136	GTG	V	CTG	L	NONSYN	conserved membrane protein	Cell wall and cell processes
113838	*Rv0101*	*nrp*	C	T	1280	CTT	L	TTT	F	NONSYN	peptide synthetase	lipid metabolism
147655	*Rv0120c*	*fusA2*	G	G	39	CTC	L	CTG	L	SYN	elongation factor G	Information pathways
165665	*Rv0138*		A	G	115	ATC	I	GTC	V	NONSYN	conserved hypothetical protein	Conserved hypothetical
243118	*Rv0204c*		G	T	33	CGC	R	TGC	C	NONSYN	conserved membrane protein	Cell wall and cell processes
243535	*Rv0205*		C	T	51	GCG	A	GTG	V	NONSYN	conserved membrane protein	Cell wall and cell processes
244719	*Rv0206c*	*mmpL3*	G	T	867	CCG	P	CTG	L	NONSYN	transmembrane transport protein	Cell wall and cell processes
244911	*Rv0206c*	*mmpL3*	G	G	803	CCA	P	CGA	R	NONSYN	transmembrane transport protein	Cell wall and cell processes
265154	*Rv0221*		G	C	363	TGG	W	TCG	S	NONSYN	conserved hypothetical protein	lipid metabolism
285876	Intergenic	-	C	T								
287904	*Rv0237*	*lpqI*	T	C	240	GTC	V	GCC	A	NONSYN	lipoprotein	Cell wall and cell processes
295917	Intergenic	-	C	T								
302660	Intergenic	-	T	C								
326039	*Rv0270*	*fadD2*	C	T	491	GCC	A	GCT	A	SYN	fatty-acid-CoA ligase	lipid metabolism
338533	*Rv0279c*	*PE_PGRS4*	C	A	181	GGC	G	AGC	S	NONSYN	PE-PGRS family protein	PE/PPE families

361596	*Rv0297*	*PE_PGRS5*	C	T	88	GCC	A	GTC	V	NONSYN	PE-PGRS family protein	PE/PPE families
370905	*Rv0304c*	*PPE5*	G	T	620	AAC	N	AAT	N	SYN	PPE family protein	PE/PPE families
370911	*Rv0304c*	*PPE5*	C	C	618	TTG	L	TTC	F	NONSYN	PPE family protein	PE/PPE families
371265	*Rv0304c*	*PPE5*	C	A	500	TCG	S	TCA	S	SYN	PPE family protein	PE/PPE families
392914	*Rv0327c*	*cyp135A1*	C	A	378	GGC	G	AGC	S	NONSYN	cytochrome P450 135A1	Intermediary metabolism and respiration
429247	*Rv0355c*	*PPE8*	C	A	1811	GGG	G	GGA	G	SYN	PPE family protein	PE/PPE families
453780	*Rv0376c*		A	C	198	ATG	M	ACG	T	NONSYN	conserved hypothetical protein	Conserved hypothetical
457032	*Rv0381c*		T	G	264	CTA	L	CTG	L	SYN	hypothetical protein	Conserved hypothetical
466096	*Rv0386*		C	T	896	CGC	R	TGC	C	NONSYN	transcriptional regulator, luxR/uhpA-family	Regulatory
515125	*Rv0425c*	*ctpH*	T	C	66	AAC	N	ACC	T	NONSYN	metal cation transporting P-type ATPase	Cell wall and cell processes
533781	*Rv0444c*		T	G	3	GAA	E	GAG	E	SYN	conserved hypothetical protein	Information pathways
544683	*Rv0453*	*PPE11*	A	G	504	ATG	M	GTG	V	NONSYN	PPE family protein	PE/PPE families
646243	*Rv0554*	*bpoC*	T	C	259	AGT	S	AGC	S	SYN	non-haem peroxidase	Virulence, detoxification and adaptation
701336	Intergenic	-	C	T								
705595	*Rv0610c*		C	A	105	GCG	A	GCA	A	SYN	hypothetical protein	Conserved hypothetical
712773	*Rv0620*	*galK*	T	C	20	TAC	Y	CAC	H	NONSYN	galactokinase	Intermediary metabolism and respiration
784581	*Rv0684*	*fusA1*	G	C	699	ACG	T	ACC	T	SYN	elongation factor G	Information pathways
814641	*Rv0721*	*rpsE*	T	C	105	GTA	V	GCA	A	NONSYN	30S ribosomal protein S5	Information pathways

848652	Intergenic	-	A	G								
853469	*Rv0758*	*phoR*	C	A	358	GGC	G	GGA	G	SYN	two component system sensor kinase	Regulatory
928525	*Rv0834c*	*PE_PGRS14*	C	A	654	AGC	S	AAC	N	NONSYN	PE-PGRS family protein	PE/PPE families
980100	*Rv0881*		C	T	247	CGG	R	TGG	W	NONSYN	rRNA methyltransferase	Intermediary metabolism and respiration
980360	*Rv0882*		G	T	46	GGT	G	TGT	C	NONSYN	transmembrane protein	Cell wall and cell processes
986563	Intergenic	-	C	T								
1047683	*Rv0938*		G	T	516	TTG	L	TTT	F	NONSYN	ATP dependent DNA ligase	Information pathways
1055672	*Rv0946c*	*pgi*	C	T	338	CTG	L	CTT	L	SYN	glucose-6-phosphate isomerase	Intermediary metabolism and respiration
1070476	*Rv0958*		C	T	198	CTC	L	CTT	L	SYN	magnesium chelatase	Intermediary metabolism and respiration
1090172	*Rv0976c*		G	T	2	CGT	R	TGT	C	NONSYN	conserved hypothetical protein	Conserved hypothetical
1095687	*Rv0980c*	*PE_PGRS18*	G	G	255	ACC	T	ACG	T	SYN	PE-PGRS family protein	PE/PPE families
1118026	*Rv1001*	*arcA*	T	C	281	ATG	M	ACG	T	NONSYN	arginine deiminase	Intermediary metabolism and respiration
1157768	*Rv1032c*	*trcS*	A	G	63	CTG	L	CGG	R	NONSYN	two component system sensor histidine kinase	Regulatory
1163404	Intergenic	-	C	G								
1205269	Intergenic	-	C	T								
1231349	*Rv1104*		G	A	17	GGC	G	AGC	S	NONSYN	para-nitrobenzyl esterase	Intermediary metabolism and respiration
1232496	*Rv1105*		T	C	62	TAT	Y	TAC	Y	SYN	para-nitrobenzyl esterase	Intermediary metabolism

												and respiration
1277731	*Rv1148c*		G	T	6	TGC	C	TGT	C	SYN	conserved hypothetical protein	Insertion seqs and phages
1290326	*Rv1161*	*narG*	G	A	1000	CGT	R	CAT	H	NONSYN	respiratory nitrate reductase alpha chain	Intermediary metabolism and respiration
1334260	*Rv1191*		C	A	110	GGC	G	GGA	G	SYN	conserved hypothetical protein	Conserved hypothetical
1354765	*Rv1212c*		C	C	299	ATG	M	ATC	I	NONSYN	glycosyltransferase	Intermediary metabolism and respiration
1440150	*Rv1286*	*cysN*	T	C	415	TTA	L	TCA	S	NONSYN	bifunctional sulfate adenyltransferase/adenylylsulfate kinase cysN/cysC	Intermediary metabolism and respiration
1451542	*Rv1295*	*thrC*	C	T	282	GCC	A	GCT	A	SYN	threonine synthase	Intermediary metabolism and respiration
1472337	Intergenic	-	C	T								
1592015	*Rv1416*	*ribH*	C	T	109	GGC	G	GGT	G	SYN	riboflavin synthase beta chain	Intermediary metabolism and respiration
1636851	*Rv1452c*	*PE_PGRS28*	G	G	460	ACC	T	AGC	S	NONSYN	PE-PGRS family protein	PE/PPE families
1692205	*Rv1501*		G	A	106	GGC	G	AGC	S	NONSYN	conserved hypothetical protein	Conserved hypothetical
1710070	*Rv1518*		C	G	143	CAA	Q	GAA	E	NONSYN	conserved hypothetical protein	Conserved hypothetical
1778557	*Rv1571*		C	T	7	CTG	L	TTG	L	SYN	conserved hypothetical protein	Conserved hypothetical
1799806	*Rv1599*	*hisD*	C	T	75	GCG	A	GTG	V	NONSYN	histidinol dehydrogenase	Intermediary metabolism and respiration
1801706	*Rv1600*	*hisC1*	G	T	271	GCC	A	TCC	S	NONSYN	histidinol-phosphate aminotransferase	Intermediary metabolism

												and respiration
1810757	*Rv1611*	*trpC*	A	G	173	CAG	Q	CGG	R	NONSYN	indole-3-glycerol phosphate synthase	Intermediary metabolism and respiration
1813255	*Rv1614*	*lgt*	T	A	29	TGC	C	AGC	S	NONSYN	prolipoprotein diacylglyceryl transferases	Cell wall and cell processes
1838153	*Rv1633*	*uvrB*	G	A	360	CGC	R	CAC	H	NONSYN	excinuclease ABC subunit B	Information pathways
1842280	Intergenic	-	C	T								
1862066	*Rv1650*	*pheT*	T	A	770	CTG	L	CAG	Q	NONSYN	phenylalanyl-tRNA synthetase beta chain	Information pathways
1862335	Intergenic	-	A	G								
1885073	*Rv1662*	*pks8*	A	G	1124	ATG	M	GTG	V	NONSYN	polyketide synthase	lipid metabolism
1887410	*Rv1663*	*pks17*	C	G	300	CCG	P	CGG	R	NONSYN	polyketide synthase	lipid metabolism
2030355	*Rv1792*	*esxM*	A	G	3	TCA	S	TCG	S	SYN	esat-6 like protein	Cell wall and cell processes
2034662	*Rv1796*	*mycP5*	A	G	312	ATG	M	GTG	V	NONSYN	proline rich membrane-anchored mycosin	Intermediary metabolism and respiration
2038913	Intergenic	-	T	G								
2121063	*Rv1870c*		C	A	123	GGT	G	GAT	D	NONSYN	conserved hypothetical protein	Conserved hypothetical
2184371	*Rv1933c*	*fadE18*	G	T	196	GCC	A	GTC	V	NONSYN	acyl-CoA dehydrogenase	lipid metabolism
2191251	*Rv1938*	*ephB*	G	A	75	AAG	K	AAA	K	SYN	epoxide hydrolase	Virulence, detoxification and adaptation
2192350	*Rv1939*		C	T	86	GCC	A	GTC	V	NONSYN	oxidoreductase	Intermediary metabolism and respiration
2227654	*Rv1983*	*PE_PGRS35*	C	T	471	CCC	P	TCC	S	NONSYN	PE-PGRS family protein	PE/PPE families

2249403	*Rv2003c*		C	A	6	CGG	R	CGA	R	SYN	conserved hypothetical protein	Conserved hypothetical
2260339	Intergenic	-	G	T								
2380682	*Rv2122c*	*hisE*	T	C	88	GAC	D	GCC	A	NONSYN	phosphoribosyl-AMP pyrophosphatase	Intermediary metabolism and respiration
2412160	*Rv2153c*	*murG*	G	G	398	CTG	L	GTG	V	NONSYN	UDP-N-acetylglucosamine-N-acetylmuramyl-(pentapeptide)pyrophosphoryl-undecaprenol-N-acetylglucosamine transferase	Cell wall and cell processes
2438090	*Rv2176*	*pknL*	G	A	50	ATG	M	ATA	I	NONSYN	transmembrane serine/threonine-protein kinase L	Regulatory
2453383	Intergenic	-	A	G								
2461815	*Rv2197c*		A	C	112	TCC	S	CCC	P	NONSYN	conserved membrane protein	Cell wall and cell processes
2537602	*Rv2264c*		A	G	250	CTT	L	CGT	R	NONSYN	conserved proline rich protein	Conserved hypothetical
2546359	*Rv2273*		C	T	86	CGC	R	CGT	R	SYN	conserved membrane protein	Cell wall and cell processes
2562004	*Rv2289*	*cdh*	C	A	110	TAC	Y	TAA	_	NONSYN	cdp-diacylglycerol pyrophosphatase	lipid metabolism
2649855	*Rv2368c*	*phoH1*	G	T	40	GAC	D	GAT	D	SYN	phosphate starvation-inducible protein	Intermediary metabolism and respiration
2654643	*Rv2374c*	*hrcA*	C	A	151	GTG	V	ATG	M	NONSYN	heat shock protein transcriptional repressor	Virulence, detoxification and adaptation
2659711	*Rv2379c*	*mbtF*	G	T	792	CCG	P	CTG	L	NONSYN	peptide synthetase	lipid metabolism
2754600	*Rv2454c*		A	C	49	AGT	S	AGC	S	SYN	oxidoreductase beta subunit	Intermediary metabolism and respiration
2793360	*Rv2485c*	*lipQ*	C	A	210	AGC	S	AAC	N	NONSYN	carboxylesterase	Intermediary metabolism and

												respiration
2835576	Intergenic	-	G	A								
2844689	*Rv2524c*	*fas*	G	G	1548	CTC	L	CTG	L	SYN	fatty-acid synthase	lipid metabolism
2871911	*Rv2552c*	*aroE*	G	T	35	GAC	D	GAT	D	SYN	shikimate 5-dehydrogenase	Intermediary metabolism and respiration
2907629	*Rv2582*	*ppiB*	G	A	272	CAG	Q	CAA	Q	SYN	peptidyl-prolyl-cis-trans-isomerase B	Information pathways
2932318	*Rv2605c*	*tesB2*	A	G	275	GGT	G	GGG	G	SYN	acyl-CoA thioesterase II	lipid metabolism
2945838	Intergenic	-	G	A								
2964024	*Rv2637*	*dedA*	G	A	147	GTC	V	ATC	I	NONSYN	transmembrane protein	Cell wall and cell processes
2976541	*Rv2652c*		G	A	5	GCA	A	GAA	E	NONSYN	phiRv2 phage protein	Insertion seqs and phages
3007840	*Rv2690c*		A	C	457	GTG	V	GCG	A	NONSYN	conserved alanine, valine and leucine rich membrane protein	Cell wall and cell processes
3011834	Intergenic	-	C	T								
3051068	Intergenic	-	C	T								
3095606	*Rv2787*		G	A	166	GGG	G	AGG	R	NONSYN	conserved alanine rich protein	Conserved hypothetical
3107859	*Rv2799*		C	T	31	GCG	A	GTG	V	NONSYN	membrane protein	Cell wall and cell processes
3130682	*Rv2823c*		G	T	364	GGC	G	GGT	G	SYN	conserved hypothetical protein	Conserved hypothetical
3195557	*Rv2886c*		G	T	292	CAC	H	CAT	H	SYN	resolvase	Insertion seqs and phages
3195975	*Rv2886c*		G	T	153	GCG	A	GTG	V	NONSYN	resolvase	Insertion seqs and phages
3196434	*Rv2887*		G	A	2	GGT	G	AGT	S	NONSYN	transcriptional regulator	Regulatory
3284726	*Rv2941*	*fadD28*	C	G	464	CTC	L	CTG	L	SYN	fatty-acid-CoA ligase	lipid metabolism

3340615	*Rv2984*	*ppk*	C	G	254	TTC	F	TTG	L	NONSYN	polyphosphate kinase	Intermediary metabolism and respiration
3378404	*Rv3018A*	*PE27A*	G	A	4	AGC	S	AGA	R	NONSYN	PE family protein	PE/PPE families
3378419	Intergenic	-	G	C								
3382748	Intergenic	-	C	G								
3388166	*Rv3029c*	*fixA*	C	C	235	ACG	T	ACC	T	SYN	electron transfer flavoprotein beta subunit	Intermediary metabolism and respiration
3392462	*Rv3032*		T	A	310	CTG	L	CAG	Q	NONSYN	transferase	Intermediary metabolism and respiration
3401951	*Rv3042c*	*serB2*	G	G	404	GTC	V	GTG	V	SYN	phosphoserine phosphatase	Intermediary metabolism and respiration
3414791	*Rv3053c*	*nrdH*	G	G	56	GCC	A	GCG	A	SYN	glutaredoxin-like electron transport protein	Information pathways
3436296	*Rv3072c*		A	C	9	GAT	D	GAC	D	SYN	conserved hypothetical protein	Conserved hypothetical
3451670	*Rv3085*		G	C	251	GTG	V	CTG	L	NONSYN	short-chain type dehydrogenase/reductase	Intermediary metabolism and respiration
3458491	*Rv3090*		G	A	94	GGT	G	GAT	D	NONSYN	hypothetical alanine and valine rich protein	Conserved hypothetical
3495658	*Rv3130c*		G	G	237	CCT	P	GCT	A	NONSYN	conserved hypothetical protein	lipid metabolism
3524903	*Rv3157*	*nuoM*	G	A	258	GCG	A	ACG	T	NONSYN	NADH dehydrogenase I chain M	Intermediary metabolism and respiration
3564790	*Rv3195*		G	A	143	GAC	D	AAC	N	NONSYN	conserved hypothetical protein	Conserved hypothetical
3600717	*Rv3224A*		T	G	28	ATC	I	AGC	S	NONSYN	conserved hypothetical protein	Conserved hypothetical
3688944	*Rv3302c*	*glpD2*	C	A	167	GGC	G	AGC	S	NONSYN	glycerol-3-phosphate dehydrogenase	Intermediary metabolism

												and respiration
3717205	*Rv3331*	*sugI*	C	T	39	CCC	P	CTC	L	NONSYN	sugar-transport membrane protein	Cell wall and cell processes
3722592	*Rv3335c*		C	C	3	GAG	E	GAC	D	NONSYN	conserved membrane protein	Cell wall and cell processes
3770768	Intergenic	-	G	T								
3771563	*Rv3359*		G	T	74	GGC	G	TGC	C	NONSYN	oxidoreductase	Intermediary metabolism and respiration
3788840	*Rv3375*	*amiD*	G	T	74	GCC	A	TCC	S	NONSYN	amidase	Intermediary metabolism and respiration
3802878	*Rv3388*	*PE_PGRS52*	G	A	409	GGG	G	GAG	E	NONSYN	PE-PGRS family protein	PE/PPE families
3831558	Intergenic	-	C	G								
3849867	*Rv3431c*		C	A	91	GTG	V	GTA	V	SYN	transposase	Insertion seqs and phages
3883226	*Rv3465*	*rmlC*	G	A	131	ATG	M	ATA	I	NONSYN	dTDP-4-dehydrorhamnose 3,5-epimerase	Intermediary metabolism and respiration
3894176	*Rv3477*	*PE31*	T	C	28	AAT	N	AAC	N	SYN	PE family protein	PE/PPE families
3968114	*Rv3531c*		A	C	277	AAT	N	AAC	N	SYN	hypothetical protein	Conserved hypothetical
4005862	*Rv3564*	*fadE33*	T	C	206	TAC	Y	CAC	H	NONSYN	acyl-CoA dehydrogenase	lipid metabolism
4033954	*Rv3591c*		G	T	30	TCC	S	TTC	F	NONSYN	hydrolase	Intermediary metabolism and respiration
4045645	*Rv3603c*		C	A	158	GCG	A	GCA	A	SYN	conserved alanine and leucine rich protein	Conserved hypothetical
4050877	*Rv3610c*	*ftsH*	G	G	669	GCC	A	GCG	A	SYN	membrane-bound ell division protein	Cell wall and cell processes

4076546	*Rv3637*		G	A	21	GTG	V	GTA	V	SYN	transposase	Insertion seqs and phages
4084599	*Rv3646c*	*topA*	G	A	887	CTG	L	ATG	M	NONSYN	DNA topoisomerase I	Information pathways
4133356	*Rv3691*		G	C	280	GGC	G	GCC	A	NONSYN	conserved hypothetical protein	Conserved hypothetical
4137657	*Rv3695*		G	A	151	CGT	R	CAT	H	NONSYN	conserved membrane protein	Cell wall and cell processes
4141036	*Rv3698*		G	T	182	GAT	D	TAT	Y	NONSYN	conserved hypothetical protein	Conserved hypothetical
4147496	*Rv3704c*	*gshA*	C	A	231	GCG	A	ACG	T	NONSYN	glutamate-cysteine ligase	Intermediary metabolism and respiration
4148529	Intergenic	-	T	G								
4225647	*Rv3779*		G	A	221	GCG	A	GCA	A	SYN	conserved alanine and leucine rich membrane protein	Cell wall and cell processes
4251087	*Rv3797*	*fadE35*	G	A	1	ATG	M	ATA	I	NONSYN	acyl-CoA dehydrogenase	lipid metabolism
4266668	Intergenic	-	C	T								
4295631	*Rv3825c*	*pks2*	C	T	1325	GTG	V	GTT	V	SYN	polyketide synthase	lipid metabolism
4297793	*Rv3825c*	*pks2*	T	G	605	ACC	T	GCC	A	NONSYN	polyketide synthase	lipid metabolism
4316114	*Rv3843c*		G	T	161	GCC	A	GCT	A	SYN	conserved membrane protein	Cell wall and cell processes
4349688	*Rv3871*		C	T	288	CCC	P	TCC	S	NONSYN	conserved hypothetical protein	Cell wall and cell processes
4350446	*Rv3871*		G	C	540	TCG	S	TCC	S	SYN	conserved hypothetical protein	Cell wall and cell processes
4361162	*Rv3881c*		T	C	255	CAG	Q	CCG	P	NONSYN	conserved alanine and glycine rich protein	Cell wall and cell processes
4388153	*Rv3903c*		G	T	760	GGC	G	GGT	G	SYN	hypothetical alanine and proline rich protein	Conserved hypothetical

4389202	*Rv3903c*	C	A	411	GAC	D	AAC	N	NONSYN	hypothetical alanine and proline rich protein	Conserved hypothetical

Ref = reference, Alt = alternative, AA = amino acid, syn = synonymous, non-syn/NONSYN = non-synonymous

Table 11

Position	**Gene**	**Gene name**	**Ref**	**Alt**	**Codon number**	**Ref codon**	**Ref AA**	**Alt codon**	**Ref AA**	**Syn/non-syn**	**Gene description**	**Functional category**
196522	*Rv0166*	*fadD5*	C	T	510	GAC	D	GAT	D	SYN	fatty-acid-CoA ligase	lipid metabolism
327897	*Rv0271c*	*fadE6*	C	A	189	GTT	V	ATT	I	NONSYN	acyl-CoA dehydrogenase	lipid metabolism
1040050	*Rv0932c*	*pstS2*	C	A	333	GCG	A	GCA	A	SYN	periplasmic phosphate-binding lipoprotein	Cell wall and cell processes
2071576	*Rv1825*		C	T	181	CCT	P	TCT	S	NONSYN	conserved hypothetical protein	Conserved hypothetical
2437971	*Rv2176*	*pknL*	G	A	11	GAG	E	AAG	K	NONSYN	transmembrane serine/threonine-protein kinase L	Regulatory
3314629	*Rv2962c*		C	A	165	TGG	W	TAG	_	NONSYN	glycosyltransferase	Intermediary metabolism and respiration
3454986	*Rv3088*		C	A	216	GCG	A	GAG	E	NONSYN	conserved hypothetical protein	lipid metabolism
3643630	*Rv3263*		G	A	152	GGC	G	AGC	S	NONSYN	DNA methylase	Information pathways
3847684	*Rv3429*	*PPE59*	G	A	174	GGG	G	AGG	R	NONSYN	PPE family protein	PE/PPE families
3895691	Intergenic	-	C	G	-	-	-	-	-	-	-	-
3936761	Intergenic	-	A	G	-	-	-	-	-	-	-	-
4136581	*Rv3694c*		C	A	178	GTG	V	GTA	V	SYN	conserved membrane protein	Cell wall and cell processes
4150648	*Rv3707c*		G	T	131	GGC	G	GGT	G	SYN	conserved hypothetical protein	Conserved hypothetical

Ref = reference, Alt = alternative, AA = amino acid, syn = synonymous, non-syn/NONSYN = non-synonymous

Table 12

Position	**Gene**	**Gene name**	**Ref**	**Alt**	**Pos. in gene**	**Stop position**	**New stop position**	**Change**	**Functional group**
125830	Rv0107c	*ctpI*	G	GA	4712	1633	1571	INSERTION	Cell wall and cell processes
131174	Intergenic, upstream of *Rv0109,* upstream of *Rv0108c*	-	T	TG					
234496	*Rv0197*		C	CGT	2266	763	762	INSERTION	Intermediary metabolism and respiration
293704	Intergenic	-	CT	C					
373282	*Rv0305c*	*PPE6*	TA	T	2430	964	0	DELETION	PE/PPE families
424320	*Rv0354c*	*PPE7*	T	TC	375	142	0	INSERTION	PE/PPE families
467508	*Rv0388c*	*PPE9*	C	CG	494	181	180	INSERTION	PE/PPE families
874835	*Rv0781*	*ptrBa*	C	CCG	603	237	0	INSERTION	Intermediary metabolism and respiration
929995	*Rv0834c*	*PE_PGRS14*	GCGGCACCCC	G	491	883	880	DELETION	PE/PPE families
968426	*Rv0872c*	*PE_PGRS15*	A	AGCCGGGTTG	1819	607	0	INSERTION	PE/PPE families
1010204	*Rv0907*		C	CG	69	533	45	INSERTION	Cell wall and cell processes
1061676	Intergenic, upstream of *Rv0951*, upstream of *Rv0950c*	-	GTGC	G					
1101832	*Rv0986*		TA	T	30	249	12	DELETION	Cell wall and cell processes
1165521	Intergenic, upstream of *Rv1042c*	-	T	TA					
1168715	*Rv1046c*		C	CT	514	175	0	INSERTION	Hypothetical
1264980	Intergenic	-	TCGC	T					
1365837	Intergenic, upstream of *Rv1223*	-	C	CG					
1527449	*Rv1358*		G	GT	838	1160	300	INSERTION	Regulatory
1753519	*Rv1549*	*fadD11.1*	G	GC	10	176	20	INSERTION	lipid metabolism
1864078	*Rv1651c*	*PE_PGRS30*	GC	G	1305	1012	550	DELETION	PE/PPE families
1894300	*Rv1668c*		G	GGTCTTGCCGC	1043	373	0	INSERTION	Cell wall and cell

									processes
1944642	*Rv1717*		CT	C	237	117	106	DELETION	Conserved hypothetical
2030340	Intergenic, upstream of *Rv1792*	-	AG	A					
2094911	*Rv1844c*	*gnd1*	ACAGCGT	A	278	486	484	DELETION	Intermediary metabolism and respiration
2109523	Intergenic, upstream of *Rv1862*	-	C	CG					
2133468	Rv1883c		T	TTCGCATGCCGTCACC	225	154	159	INSERTION	Conserved hypothetical
2137521	Intergenic	-	A	ACTCCGATCAC					
2207591	Intergenic, upstream of *Rv1964*, upstream of *Rv1963c*	-	T	TC					
2208071	*Rv1964*	*yrbE3A*	TGC	T	372	266	129	DELETION	Virulence, detoxification and adaptation
2357268	*Rv2098c*	*PE_PGRS36*	TGCC	T	766	435	434	DELETION	PE/PPE families
2368564	Intergenic	-	TA	T					
2382085	*Rv2123*	*PPE37*	AGT	A	1015	474	0	DELETION	PE/PPE families
2523205	Intergenic	-	G	GCGC					
2525722	*Rv2250A*		CG	C	321	140	0	DELETION	Intermediary metabolism and respiration
2534562	*Rv2262c*		GGA	G	991	361	0	DELETION	lipid metabolism
2796131	*Rv2487c*	*PE_PGRS42*	G	GC	1255	695	432	INSERTION	PE/PPE families
2881597	*Rv2561*		AG	A	189	98	0	DELETION	Conserved hypothetical
3131469	*Rv2823c*		T	TTGTCGGCGA	305	810	813	INSERTION	Conserved hypothetical
3190145	*Rv2879c*		TC	T	8	190	33	DELETION	Conserved hypothetical
3415180	Intergenic, upstream of *Rv3053c*	-	ACACCTAGGGGGTGG	A					
3473996	Intergenic, upstream of *Rv3106,* upstream of *Rv3105c*	-	G	GA					
3580636	Intergenic, upstream of	-	CT	C					

	Rv3203, upstream of *Rv3202c*								
3590686	Intergenic	-	G	GC					
3844756	*Rv3428c*		GC	G	1215	411	0	DELETION	Insertion seqs and phages
3853708	*Rv3434c*		C	CT	221	238	143	INSERTION	Cell wall and cell processes
3862472	Intergenic, upstream of *Rv3443c*	-	GA	G					
3936661	*Rv3508*	*PE_PGRS54*	CCGACGGCGA	C	5657	1902	1899	DELETION	PE/PPE families
3998059	*Rv3558*	*PPE64*	AGGC	A	80	553	552	DELETION	PE/PPE families
4093879	*Rv3652*	*PE_PGRS60*	TG	T	248	105	0	DELETION	PE/PPE families
4095001	*Rv3655c*		CG	C	300	126	0	DELETION	Conserved hypothetical
4197138	Intergenic, upstream of *Rv3748*	-	C	CT					
4198611	Intergenic, upstream of *Rv3751*, upstream of *Rv3750c*	-	CG	C					
4336090	Intergenic, upstream of *Rv3860,* upstream of *Rv3859c*	-	A	AT					
4338595	Intergenic, upstream of *Rv3863*, upstream of *Rv3862c*	-	GC	G					
4383144	*Rv3897c*		C	CCGGGG	497	211	0	INSERTION	Conserved hypothetical
4400660	*Rv3911*	*sigM*	AC	A	475	223	197	DELETION	Information pathways

Ref = reference, Alt = alternative, AA = amino acid, syn = synonymous, non-syn/NONSYN = non-synonymous, pos = position

APPENDIX D

REGULATED PROTEINS IN SAWC 3517 AND SAWC 3651

Up-regulated in SAWC 3517, down-regulated in SAWC 3651		Up-regulated in SAWC 3651, down-regulated in SAWC 3517	
Rv0107c	Cell wall and cell processes	Rv0313	Conserved hypothetical
Rv0120c	Information pathways	Rv0351	Virulence, detoxification and adaptation
Rv0166	lipid metabolism	Rv0394c	Cell wall and cell processes
Rv0169	Virulence, detoxification and adaptation	Rv0467	Intermediary metabolism and respiration
Rv0386	Regulatory	Rv0516c	Information pathways
Rv0491	Regulatory	Rv0724	Cell wall and cell processes
Rv0655	Cell wall and cell processes	Rv0991c	Conserved hypothetical
Rv0690c	Conserved hypothetical	Rv1013	lipid metabolism
Rv0757	Regulatory	Rv1094	lipid metabolism
Rv0758	Regulatory	Rv1178	Intermediary metabolism and respiration
Rv0778	Intermediary metabolism and respiration	Rv1185c	lipid metabolism
Rv0932c	Cell wall and cell processes	Rv1275	Cell wall and cell processes
Rv0958	Intermediary metabolism and respiration	Rv1345	lipid metabolism
Rv0966c	Conserved hypothetical	Rv1348	Cell wall and cell processes
Rv1318c	Intermediary metabolism and respiration	Rv1685c	Conserved hypothetical
Rv1397c	Virulence, detoxification and adaptation	Rv1687c	Cell wall and cell processes
Rv1399c	Intermediary metabolism and respiration	Rv1722	lipid metabolism
Rv1421	Conserved hypothetical	Rv1932	Virulence, detoxification and adaptation
Rv1427c	lipid metabolism	Rv2018	Conserved hypothetical
Rv1523	Intermediary metabolism and respiration	Rv2285	lipid metabolism
Rv1532c	Conserved hypothetical	Rv2380c	lipid metabolism
Rv1538c	Intermediary metabolism and respiration	Rv2383c	lipid metabolism
Rv1597	Conserved hypothetical	Rv2386c	lipid metabolism
Rv1654	Intermediary metabolism and respiration	Rv2428	Virulence, detoxification and adaptation
Rv1661	lipid metabolism	Rv2429	Virulence, detoxification and adaptation
Rv1700	Information pathways	Rv2465c	Intermediary metabolism and respiration
Rv1742	Hypothetical	Rv2590	lipid metabolism
Rv1796	Intermediary metabolism and respiration	Rv2627c	Conserved hypothetical
Rv1812c	Intermediary metabolism and respiration	Rv2711	Regulatory
Rv1844c	Intermediary metabolism and respiration	Rv2930	lipid metabolism
Rv1883c	Conserved hypothetical	Rv2931	lipid metabolism
Rv2001	Conserved hypothetical	Rv2947c	lipid metabolism
Rv2155c	Cell wall and cell processes	Rv3093c	Intermediary metabolism and respiration
Rv2161c	Intermediary metabolism and respiration	Rv3094c	Conserved hypothetical
Rv2557	Conserved hypothetical	Rv3161c	Intermediary metabolism and respiration
Rv2568c	Conserved hypothetical	Rv3547	Intermediary metabolism and respiration

Rv2682c	Intermediary metabolism and respiration	Rv3720	lipid metabolism
Rv2702	Intermediary metabolism and respiration	Rv3765c	Regulatory
Rv2821c	Conserved hypothetical	Rv3841	Intermediary metabolism and respiration
Rv2824c	Conserved hypothetical	Rv3849	Regulatory
Rv2906c	Information pathways		
Rv2956	Conserved hypothetical		
Rv2986c	Information pathways		
Rv3077	Intermediary metabolism and respiration		
Rv3106	Intermediary metabolism and respiration		
Rv3203	Intermediary metabolism and respiration		
Rv3213c	Cell wall and cell processes		
Rv3232c	Intermediary metabolism and respiration		
Rv3237c	Conserved hypothetical		
Rv3241c	Information pathways		
Rv3279c	Intermediary metabolism and respiration		
Rv3303c	Intermediary metabolism and respiration		
Rv3320c	Virulence, detoxification and adaptation		
Rv3322c	Intermediary metabolism and respiration		
Rv3323c	Intermediary metabolism and respiration		
Rv3329	Intermediary metabolism and respiration		
Rv3480c	lipid metabolism		
Rv3505	lipid metabolism		
Rv3573c	lipid metabolism		
Rv3600c	Conserved hypothetical		
Rv3603c	Conserved hypothetical		
Rv3684	Intermediary metabolism and respiration		
Rv3691	Conserved hypothetical		
Rv3726	Intermediary metabolism and respiration		
Rv3866	Cell wall and cell processes		
Rv3868	Cell wall and cell processes		
Rv3879c	Cell wall and cell processes		
Rv3918c	Cell wall and cell processes		

APPENDIX E

SCRIPT USED TO ANNOTATE CONFIDANCE VARIANTS

```
#!/usr/bin/perl

# annotating SNPs for one strain
# Note that the H37RvAnno.txt and H37RvGeneSeq.fasta files used in this script were downloaded from
the Tuberculosis database (TBDB)

use strict;

my ($vcf)=@ARGV;

my @headers;



open(MUT, "$vcf") or die "Cannot open $vcf:$!\n";

while (<MUT>) {

    chomp;

       next if (/^##/);

       if (/^#/) {

       @headers = split(/\t/,$_);

       print join ("\t", ("CHROM", "POS", "LOCUS", "SYMBOL", "REFBASE", "ALTBASE",
@headers[5,6,7,8,9], "CODONnr", "REFCODON", "REFAA", "MUTCODON", "MUTAA", "CHANGE")), "\n";

       next;

    }

       my ($CHROM, $POS, $ID, $REFBASE, $ALTBASE, $QUAL, $FILTER, $INFO, $FORMAT,
$STRAIN_1)=split(/\t/,$_);
       my $annofile="/home/adippenaar/Documents/Bioinformatics/Out_groups_output/H37RvAnno.txt";
       my $line=0;
       my $prevGene;
       my $prevStrand;
       my
$geneseqfile="/home/adippenaar/Documents/Bioinformatics/Out_groups_output/H37RvGeneSeq.fasta";
       my $codonsize = 3;
       open(ANNO, "$annofile") or die "Cannot open $annofile:$!\n";
       while (<ANNO>) {
              if ($line==0) {
                 $line++;
                 next;
             }
```

```
        chomp;
        my ($LOCUS, $SYMBOL, $SYNOYM, $LENGHT, $START, $STOP, $STRAND,
$NAME)=split(/\t/,$_);

        #print "'$STRAND'\n";

        if ($POS>$START && $POS<$STOP) {

            open(GENESEQ, "$geneseqfile") or die "Cannot open $geneseqfile:$!\n";
            my $seq;
            while (<GENESEQ>) {

                next unless /^>$LOCUS/;

                while (<GENESEQ>) {

                    last if /^>/;

                    chomp;

                    $seq.=$_;
                }

                last;
            }

            close GENESEQ;

            my $posingene;

            if ($STRAND eq "+") {

                $posingene = ($POS - $START) + 1;
            }

            else {$posingene = ($STOP - $POS) + 1;

            }



            my $codonnr = int(($posingene - 1)/$codonsize + 1);

            #print "'$codonnr'\n";

            my $firstbase = (($codonnr - 1) * $codonsize) + 1 - 1;

            #print "'$firstbase'\n";

            my $lastbase = (($codonnr - 1) * $codonsize) + 3 - 1;

            #print "'$lastbase'\n";

# -1: subst will start to count at 0
```

```
my $codon = substr($seq, $firstbase, $codonsize);
#print "'$seq'\n";
#print "'$codon'\n";
my $aa = &codon2aa($codon);
#print "aa is $aa\n";
my $offset;
if (($posingene % $codonsize) == 1) {
    $offset = 1 - 1;
}
elsif (($posingene % $codonsize) == 2) {
    $offset = 2 - 1;
}
elsif (($posingene % $codonsize) == 0) {
    $offset = 3 - 1;
}
#print "offset value is $offset\n";
my $mutcodon = $codon;
my $mutaa;
my $change;

if ($STRAND eq "+") {
        substr($mutcodon, $offset, 1) = $ALTBASE;
        #print "altbase is $ALTBASE\n";
        #print "my mutcodon is $mutcodon\n";
        $mutaa = &codon2aa($mutcodon);
        if ($mutaa eq $aa) {$change = "SYN";
        }
        elsif ($mutaa ne $aa) {$change = "NONSYN";
        }
}
```

```
else   {

        if ($ALTBASE eq "T") {$ALTBASE = "B";
        }
        if ($ALTBASE eq "A") {$ALTBASE = "D";
        }
        if ($ALTBASE eq "C") {$ALTBASE = "E";
        }
        if ($ALTBASE eq "G") {$ALTBASE = "H";
        }
        if ($ALTBASE eq "B") {$ALTBASE = "A";
        }
        if ($ALTBASE eq "D") {$ALTBASE = "T";
        }
        if ($ALTBASE eq "E") {$ALTBASE = "G";
        }
        if ($ALTBASE eq "H") {$ALTBASE = "C";
        }

        substr($mutcodon, $offset, 1) = $ALTBASE;
        #print "altbase is $ALTBASE\n";
        #print "my mutcodon is $mutcodon\n";
        $mutaa = &codon2aa($mutcodon);
        if ($mutaa eq $aa) {$change = "SYN";
        }
        elsif ($mutaa ne $aa) {$change = "NONSYN";
        }
}
```

```
#                               print join ("\t", ($CHROM, $POS, $LOCUS, $SYMBOL, $REFBASE, $ALTBASE,
$QUAL, $FILTER, $INFO, $FORMAT, $STRAIN_1, $codonnr, $codon, $aa, $mutcodon, $NAME)), "\n";
                                print join ("\t", ($CHROM, $POS, $LOCUS, $SYMBOL, $REFBASE, $ALTBASE,
$QUAL, $FILTER, $INFO, $FORMAT, $STRAIN_1, $codonnr, $codon, $aa, $mutcodon, $mutaa,
$change, $NAME)), "\n";

                                last;

                        }

                        if ($POS<$START){

                              $SYMBOL = "-";

                                my $message="Intergenic";

                                #if($STRAND eq "+") {$message.=".Upstream of $LOCUS";

                                #}

                                #if($prevStrand eq "-") {$message.=".Upstream of $prevGene";

                                #}

                               print join ("\t", ($CHROM, $POS, $message, $SYMBOL, $REFBASE, $ALTBASE,
$QUAL, $FILTER, $INFO, $FORMAT, $STRAIN_1)), "\n";
                                last;
                        }
                        $prevGene=$LOCUS;
                        $prevStrand=$STRAND;



            }
            close ANNO;
}
close MUT;



############################################################
sub codon2aa {

      my $codon = uc shift;

      if       ( $codon =~ m/GC./ )              { return "A" } # Alanine
      elsif ( $codon =~ m/TG[TC]/ )        { return "C" } # Cysteine
      elsif ( $codon =~ m/GA[TC]/ )        { return "D" } # Aspartic Acid
      elsif ( $codon =~ m/GA[AG]/ )        { return "E" } # Glutamic Acid
      elsif ( $codon =~ m/TT[TC]/ )        { return "F" } # Phenylalanine
      elsif ( $codon =~ m/GG./ )           { return "G" } # Glycine
      elsif ( $codon =~ m/CA[TC]/ )        { return "H" } # Histidine
      elsif ( $codon =~ m/AT[TCA]/ )       { return "I" } # Isoleucine
      elsif ( $codon =~ m/AA[AG]/ )        { return "K" } # Lysine
      elsif ( $codon =~ m/TT[AG]|CT./ ) { return "L" } # Leucine
```

```
        elsif ( $codon =~ m/ATG/ )              { return "M" } # Methionine
        elsif ( $codon =~ m/AA[TC]/ )         { return "N" } # Asparagine
        elsif ( $codon =~ m/CC./ )              { return "P" } # Proline
        elsif ( $codon =~ m/CA[AG]/ )         { return "Q" } # Glutamine
        elsif ( $codon =~ m/CG.|AG[AG]/ ) { return "R" } # Arginine
        elsif ( $codon =~ m/TC.|AG[TC]/ ) { return "S" } # Serine
        elsif ( $codon =~ m/AC./ )              { return "T" } # Threonine
        elsif ( $codon =~ m/GT./ )              { return "V" } # Valine
        elsif ( $codon =~ m/TGG/ )              { return "W" } # Tryptophan
        elsif ( $codon =~ m/TA[TC]/ )         { return "Y" } # Tyrosine
        elsif ( $codon =~ m/TA[AG]|TGA/ ) { return "_" } # Stop
        else { die "Bad codon \"$codon\"!\n" }

}
```

SCRIPT USED TO ASSIGN FUNCTIONAL GROUPS TO ANNOTATED HIGH CONFIDANCE VARIANTS

```
#! /usr/bin/perl

#Note that the functional groups file specified in the usage information of this script was compiled from data from the Tuberculist knowledge base. The protein file refers to the protein or SNP file with a list of RV numbers that is used as input.

use strict;

if ($#ARGV < 1) {

        die "usage: $0 [functional_groups file] [protein file]\n";

}

my $fgrps =$ARGV[0];

my $proteinfile=$ARGV[1];

my %grps = ();

open(F, $fgrps);

while (<F>) {

        chomp;

        next if ($_=~/^Rv_number/);

        my (@line) = split(/\t/);

        $grps{$line[0]} = $line[5];

}

close(F);


open(F, $proteinfile);

while (<F>) {

        chomp;

        next if ($_=~/^CHROM/);

        my @line = split(/\t/);

        my $rv_id = $line[2];
        #$rv_id =~s/\"//g;

        print "$_\t".$grps{$rv_id}."\n";

}

close(F);
```

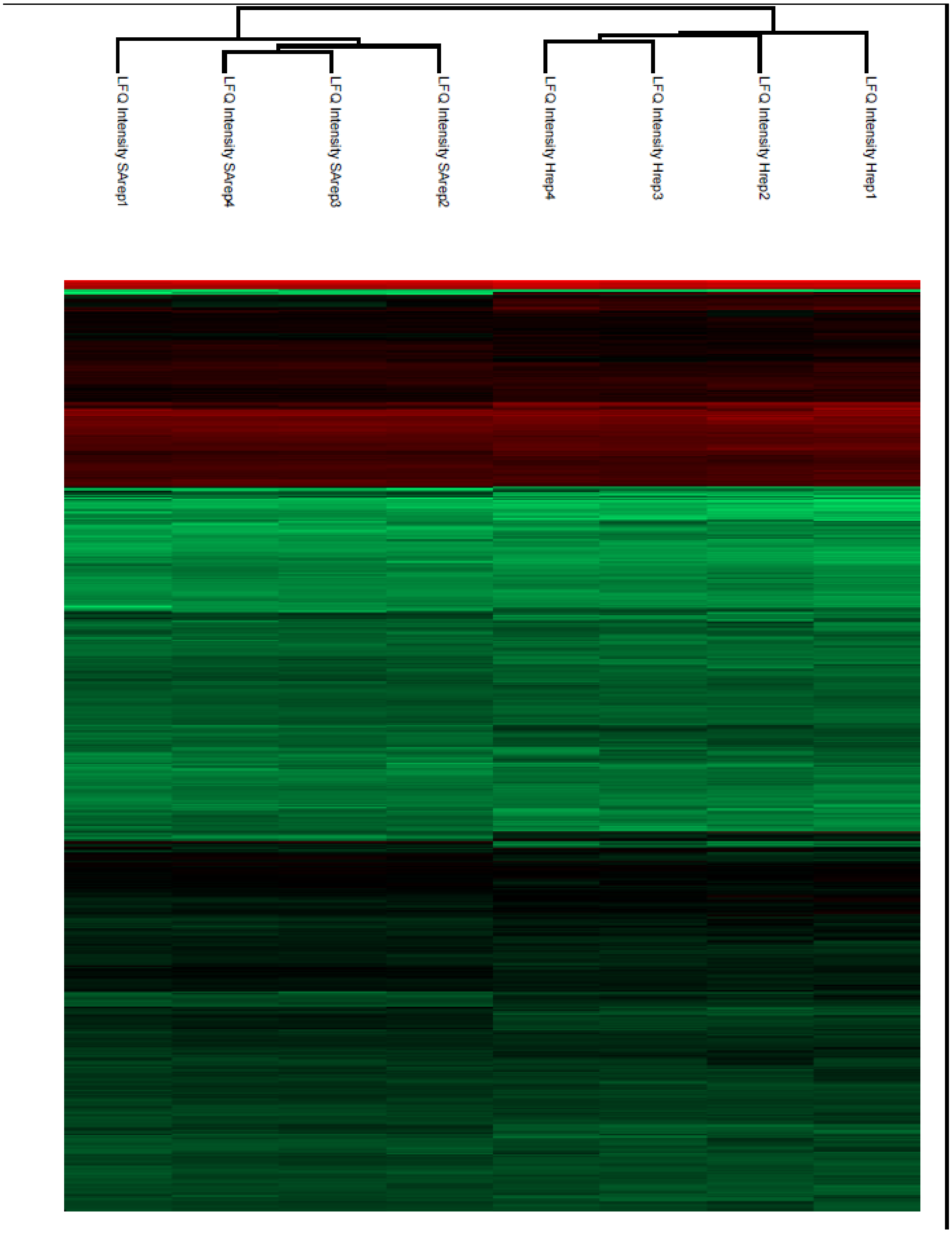
SCRIPT USED TO CREATE SNP STRING FOR PHYLOGENOMIC INFERENCE

```
#This python script assumes:
#1. A single chromosome
#2. Input files with the naming convention
#     pos_alt_<sample_nr>_<list_identifier>.txt
#     Entries in these files contain 2 columns: the position of a variant and its
#     value
#3. A fasta file (reference sequence) with 1 header line and a column length of 60

#The output of this script is <sample_nr>.txt files, one file for each of the
#input files. The content of each file is a string of nucleotides, in order
#of position, for the set of all positions read from the input files. If a
#value for a certain position is not available for a sample, it contains the
#value of the the reference allele, as read from the FASTA file.

nr_fasta_header_lines = 1
col_len = 60
input_file_prefix = 'pos_alt_'

import sys, os

if len(sys.argv) != 4:
        print("Usage: python create_phylo_files.py <fasta_file> <input_dir> " + \
                        "<output_dir>")
        sys.exit(-1)
else:
        fasta_file_name = sys.argv[1]
        input_dir = sys.argv[2]
        output_dir = sys.argv[3]


#Map containing a map of positions and variants, keyed on sample nr
sample_map = {}
#Map containing reference variants, keyed on position
ref_map = {}
#Set the column length to 60

input_dir_list = os.listdir(input_dir)
for file_name in input_dir_list:
        if file_name[0:8] == input_file_prefix:
                print('Processing ' + file_name + ' ...')

                #Determine the sample number
                sample_nr_length = file_name[9:].find('_')
                sample_nr = file_name[8:9+sample_nr_length]

                #Initialize the variant map for this sample, keyed on position
                var_map = {}

                #Read the file content and populate the maps
                in_file = open(os.path.join(input_dir, file_name))
                for line in in_file:
                        data = line.strip().split()
```

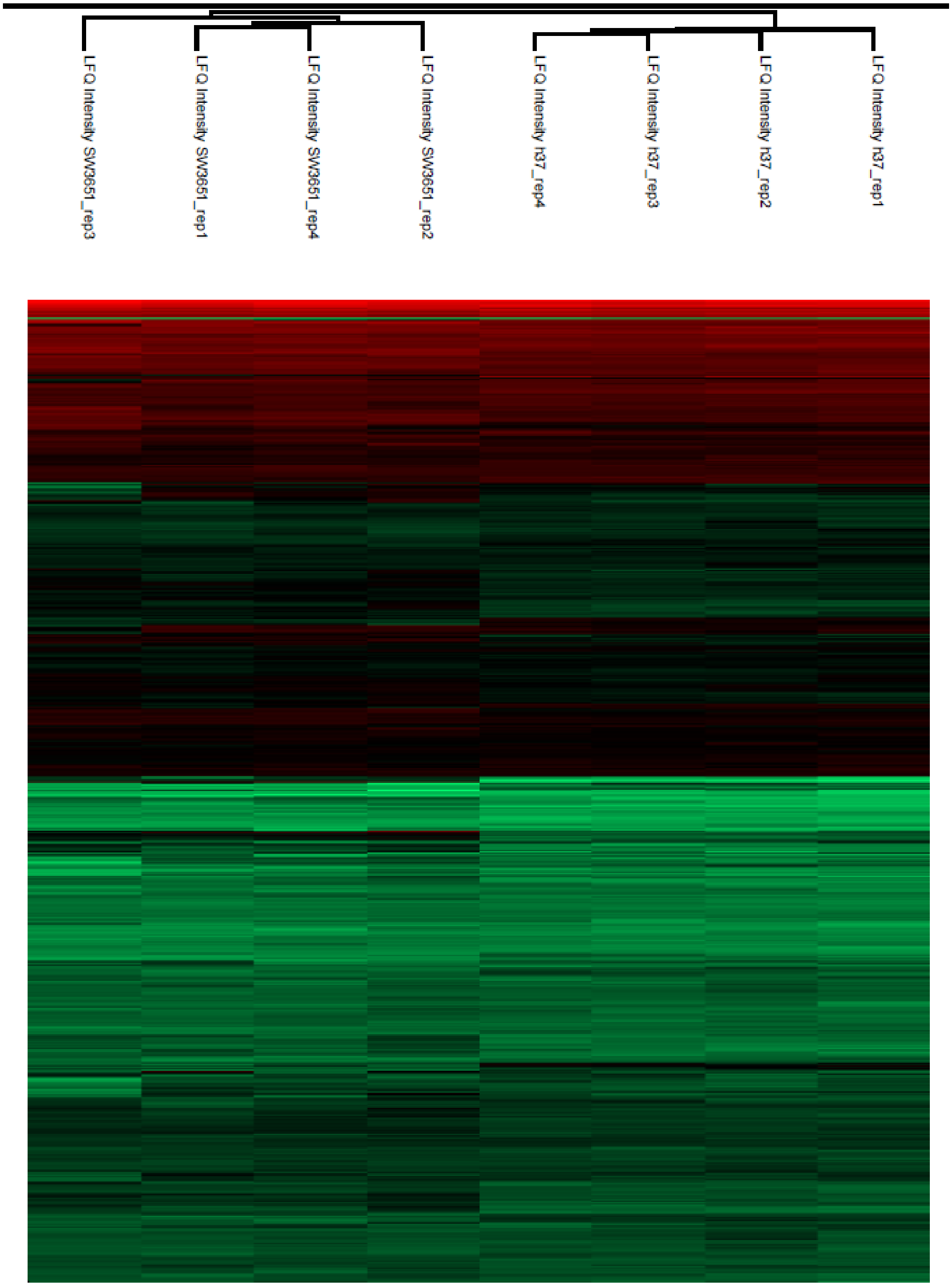
```
            pos, var = int(data[0]), data[1]
            var_map[pos] = var
            if (pos in ref_map) == False:
                #Get the reference allele from the fasta file
                col_nr = str(pos % col_len)
                row_nr = str((pos / col_len) + nr_fasta_header_lines + 1)
                if col_nr == '0':
                    col_nr = str(col_len)
                os.system('head -' + row_nr + ' ' + fasta_file_name + ' | ' + \
                    'tail -1 > tmp_fasta_line.txt')
                ref = os.popen('cut -c' + col_nr + \
                    ' tmp_fasta_line.txt').read().strip().upper()
                os.system('rm tmp_fasta_line.txt')
                ref_map[pos] = ref
        in_file.close()
        sample_map[sample_nr] = var_map

#Sort the positions
positions = sorted(ref_map.keys())

for sample_nr in sample_map.keys():
    #Open the output file for this sample
    print('Writing ' + sample_nr + '.txt ....')
    out_file = open(os.path.join(output_dir, sample_nr) + '.txt', 'w')
    debug_file = open(os.path.join(output_dir, sample_nr) + '_debug.txt', 'w')
    debug_file.write('pos\tref\talt\n')

    #Get the variant map for this sample
    var_map = sample_map[sample_nr]

    #Write the output file
    for pos in positions:
        debug_file.write(str(pos) + '\t')
        if pos in var_map:
            var = var_map[pos]
            debug_file.write('*\t' + var + '\n')
        else:
            var = ref_map[pos]
            debug_file.write(var + '\t*\n')
        out_file.write(var)
    out_file.close()
    debug_file.close()
```

APPENDIX F

HEAT MAP SHOWING PROTEIN CLUSTERING FOR SAWC 3517 AND *M. TUBERCULOSIS* H37RV



HEAT MAP SHOWING PROTEIN CLUSTERING FOR SAWC 3651 AND *M. TUBERCULOSIS* H37RV



REFERENCES

Abadia, E., Zhang, J., dos Vultos, T., Ritacco, V., Kremer, K., Aktas, E., Matsumoto, T., Refregier, G., van Soolingen, D., Gicquel, B., Sola, C., 2010. Resolving lineage assignation on *Mycobacterium tuberculosis* clinical isolates classified by spoligotyping with a new high-throughput 3R SNPs based method. Infect. Genet. Evol. 10, 1066–1074.

Abdallah, A.M., Gey van Pittius, N.C., Champion, P.A.D., Cox, J., Luirink, J., Vandenbroucke-Grauls, C.M.J.E., Appelmelk, B.J., Bitter, W., 2007. Type VII secretion--mycobacteria show the way. Nat. Rev. Microbiol. 5, 883–891.

Alix, E., Godreuil, S., Blanc-Potard, A.-B., 2006. Identification of a Haarlem genotype-specific single nucleotide polymorphism in the mgtC virulence gene of *Mycobacterium tuberculosis*. J. Clin. Microbiol. 44, 2093–2098.

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

Aranaz, A., Liébana, E., Mateos, A., Dominguez, L., Vidal, D., Domingo, M., Gonzolez, O., Rodriguez-Ferri, E.F., Bunschoten, A.E., Van Embden, J.D., Cousins, D., 1996. Spacer oligonucleotide typing of *Mycobacterium bovis* strains from cattle and other animals: a tool for studying epidemiology of tuberculosis. J. Clin. Microbiol. 34, 2734–2740.

Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data, URL http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (accessed 6.21.13).

Baker, L., Brown, T., Maiden, M.C., Drobniewski, F., 2004. Silent nucleotide polymorphisms and a phylogeny for *Mycobacterium tuberculosis*. Emerging Infect. Dis. 10, 1568–1577.

Behe, M., J., 1998. Darwin's Big Black Box: The Biochemical Challenge to Evolution. Free Press. New York. 307 pp.

Bentley, S.D., Comas, I., Bryant, J.M., Walker, D., Smith, N.H., Harris, S.R., Thurston, S., Gagneux, S., Wood, J., Antonio, M., Quail, M.A., Gehre, F., Adegbola, R.A., Parkhill, J., de Jong, B.C., 2012. The Genome of *Mycobacterium africanum* West African 2 Reveals a Lineage-Specific Locus and Genome Erosion Common to the *M. tuberculosis* Complex. PLoS Negl Trop Dis 6, e1552.

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

Bitter, W., Houben, E.N.G., Bottai, D., Brodin, P., Brown, E.J., Cox, J.S., Derbyshire, K., Fortune, S.M., Gao, L.-Y., Liu, J., Gey van Pittius, N.C., Pym, A.S., Rubin, E.J., Sherman, D.R., Cole, S.T., Brosch, R., 2009. Systematic Genetic Nomenclature for Type VII Secretion Systems. PLoS Pathog 5, e1000507.

Blouin, Y., Hauck, Y., Soler, C., Fabre, M., Vong, R., Dehan, C., Cazajous, G., Massoure, P.-L., Kraemer, P., Jenkins, A., Garnotel, E., Pourcel, C., Vergnaud, G., 2012. Significance of the identification in the Horn of Africa of an exceptionally deep branching *Mycobacterium tuberculosis* clade. PLoS ONE 7, e52841.

Brisse, S., Supply, P., Brosch, R., Vincent, V., Gutierrez, M.C., 2006. "A re-evaluation of M. prototuberculosis": continuing the debate. PLoS Pathog. 2, e95.

Brosch, R., Gordon, S.V., Marmiesse, M., Brodin, P., Buchrieser, C., Eiglmeier, K., Garnier, T., Gutierrez, C., Hewinson, G., Kremer, K., Parsons, L.M., Pym, A.S., Samper, S., van Soolingen, D., Cole, S.T., 2002. A new evolutionary scenario for the *Mycobacterium tuberculosis* complex. Proc. Natl. Acad. Sci. U.S.A. 99, 3684–3689.

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

Brown, T., Nikolayevskyy, V., Velji, P., Drobniewski, F., 2010. Associations between *Mycobacterium* *tuberculosis* strains and phenotypes. Emerging Infect. Dis 16, 272–280.

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

Bryant, J.M., Schürch, A.C., van Deutekom, H., Harris, S.R., de Beer, J.L., de Jager, V., Kremer, K., van Hijum, S.A., Siezen, R.J., Borgdorff, M., Bentley, S.D., Parkhill, J., van Soolingen, D., 2013. Inferring patient to patient transmission of *Mycobacterium tuberculosis* from whole genome sequencing data. BMC Infect. Dis. 13, 110.

Camus, J.-C., Pryor, M.J., Médigue, C., Cole, S.T., 2002. Re-annotation of the genome sequence of *Mycobacterium tuberculosis* H37Rv. Microbiology 148, 2967 –2973.

Carver, T., Harris, S.R., Berriman, M., Parkhill, J., McQuillan, J.A., 2011. Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data. Bioinformatics 28, 464–469.

Caws, M., Thwaites, G., Dunstan, S., Hawn, T.R., Lan, N.T.N., Thuong, N.T.T., Stepniewska, K., Huyen, M.N.T., Bang, N.D., Loc, T.H., Gagneux, S., van Soolingen, D., Kremer, K., van der Sande, M., Small, P., Anh, P.T.H., Chinh, N.T., Quy, H.T., Duyen, N.T.H., Tho, D.Q., Hieu, N.T., Torok, E., Hien, T.T., Dung, N.H., Nhu, N.T.Q., Duy, P.M., van Vinh Chau, N., Farrar, J., 2008. The influence of host and bacterial genotype on the development of disseminated disease with *Mycobacterium tuberculosis*. PLoS Pathog. 4, e1000034.

Chan, J., Halachev, M., Yates, E., Smith, G., Pallen, M., 2012. Whole-genome sequence of the emerging pathogen *Mycobacterium* abscessus strain 47J26. J. Bacteriol. 194, 549.

Chiang, C.-Y., Centis, R., Migliori, G.B., 2010. Drug-resistant tuberculosis: past, present, future. Respirology 15, 413–432.

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

Chihota, V.N., Müller, B., Mlambo, C.K., Pillay, M., Tait, M., Streicher, E.M., Marais, E., van der Spuy, G.D., Hanekom, M., Coetzee, G., Trollip, A., Hayes, C., Bosman, M.E., Gey van Pittius, N., Victor, T.C., van Helden, P.D., Warren, R.M., 2011. The population structure of multi- and extensively drug-resistant tuberculosis in South Africa. Journal of Clinical Microbiology. 3 995-1002

Chohota, V, Müller, B, Mlambo, C.K, Pillay, M, Tait, M, Streicher, E.M, Marais, E, van der Spuy, G.D, Hanekom, M, Coetzee, G, Trollip, A, Hayes, C, Bosman, M.E, Gey van Pittius, N.C, Victor, T, van Helden, P.D, Warren, R.M. 2012. Population structure of multi-and extensively drug-resistent *Mycobacterium tuberculosis* strains in South Africa. Clin. Microbiol. 50(9):995-1002

Chuang, P.-C., Chen, Y.-M.A., Chen, H.-Y., Jou, R., 2010. Single nucleotide polymorphisms in cell wall biosynthesis-associated genes and phylogeny of *Mycobacterium tuberculosis* lineages. Infect. Genet. Evol. 10, 459–466.

Chuang, P.-C., Liu, H., Sola, C., Chen, Y.-M.A., Jou, R., 2008. Spoligotypes of *Mycobacterium tuberculosis* isolates of a high tuberculosis burden aboriginal township in Taiwan. Infect. Genet. Evol. 8, 553–557.

Claesson, M.J., Wang, Q., O'Sullivan, O., Greene-Diniz, R., Cole, J.R., Ross, R.P., O'Toole, P.W., 2010. Comparison of two next-generation sequencing technologies for resolving highly complex microbiota composition using tandem variable 16S rRNA gene regions. Nucleic Acids Res 38, e200.

Click, E.S., Moonan, P.K., Winston, C.A., Cowan, L.S., Oeltmann, J.E., 2012. Relationship Between *Mycobacterium tuberculosis* Phylogenetic Lineage and Clinical Site of Tuberculosis. Clin. Infect. Dis. 54, 211–219.

Cock, P.J.A., Fields, C.J., Goto, N., Heuer, M.L., Rice, P.M., 2010. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Res 38, 1767–1771.

Cole, S.T., 2002. Comparative and functional genomics of the *Mycobacterium tuberculosis* complex. Microbiology (Reading, Engl.) 148, 2919–2928.

Cole, S.T., Brosch, R., Parkhill, J., Garnier, T., Churcher, C., Harris, D., Gordon, S.V., Eiglmeier, K., Gas, S., Barry, C.E., 3rd, Tekaia, F., Badcock, K., Basham, D., Brown, D., Chillingworth, T., Connor, R., Davies, R., Devlin, K., Feltwell, T., Gentles, S., Hamlin, N., Holroyd, S., Hornsby, T., Jagels, K., Krogh, A., McLean, J., Moule, S., Murphy, L., Oliver, K., Osborne, J., Quail, M.A., Rajandream, M.A., Rogers, J., Rutter, S., Seeger, K., Skelton, J., Squares, R., Squares, S., Sulston, J.E., Taylor, K., Whitehead, S., Barrell, B.G., 1998. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. Nature 393, 537–544.

Coll, F., Mallard, K., Preston, M.D., Bentley, S., Parkhill, J., McNerney, R., Martin, N., Clark, T.G., 2012. SpolPred: rapid and accurate prediction of *Mycobacterium tuberculosis* spoligotypes from short genomic sequences. Bioinformatics 28, 2991–2993.

Collins, C.H., Yates, M.D., Grange, J.M., 1982. Subdivision of *Mycobacterium tuberculosis* into five variants for epidemiological purposes: methods and nomenclature. J Hyg (Lond) 89, 235–242.

Comas, I., Coscolla, M., Luo, T., Borrell, S., Holt, K.E., Kato-Maeda, M., Parkhill, J., Malla, B., Berg, S., Thwaites, G., Yeboah-Manu, D., Bothamley, G., Mei, J., Wei, L., Bentley, S., Harris, S.R., Niemann, S., Diel, R., Aseffa, A., Gao, Q., Young, D., Gagneux, S., 2013. Out-of-Africa migration and Neolithic coexpansion of *Mycobacterium tuberculosis* with modern humans. Nat Genet advance online publication.

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

Comas, I, Chakravartti, J, Small, P.M, Galagan, J, Niemann, S, Kremer, K, Ernst, J.D, Gagneux, S. 2010. Human T cell epitopes of *Mycobacterium tuberculosis* are evolutionarily hyperconserved. Nature Genetics. 42: 498-503

Cox, J, Mann, M. 2008. Maxquant enables high peptide identification rates, individualised P.P.B.-range mass accuracies and proteome-wide protein quantification. Nat. Biotechnol. 26(12):1367-72.

Dahl, J.L., Kraus, C.N., Boshoff, H.I.M., Doan, B., Foley, K., Avarbock, D., Kaplan, G., Mizrahi, V., Rubin, H., Barry, C.E., 3rd, 2003. The role of RelMtb-mediated adaptation to stationary phase in long-term persistence of *Mycobacterium tuberculosis* in mice. Proc. Natl. Acad. Sci. U.S.A. 100, 10026–10031.

Daugelat, S., Kowall, J., Mattow, J., Bumann, D., Winter, R., Hurwitz, R., Kaufmann, S.H.E., 2003. The RD1 proteins of *Mycobacterium tuberculosis*: expression in *Mycobacterium* smegmatis and biochemical characterization. Microbes Infect. 5, 1082–1095.

Davies, J., Davies, D., 2010. Origins and Evolution of Antibiotic Resistance. Microbiol. Mol. Biol. Rev. 74, 417–433.

De Souza, G.A., Arntzen, M.Ø., Fortuin, S., Schürch, A.C., Målen, H., McEvoy, C.R.E., van Soolingen, D., Thiede, B., Warren, R.M., Wiker, H.G., 2011. Proteogenomic analysis of polymorphisms and gene annotation divergences in prokaryotes using a clustered mass spectrometry-friendly database. Mol. Cell Proteomics 10, M110.002527.

De Souza, G.A., Fortuin, S., Aguilar, D., Pando, R.H., McEvoy, C.R.E., van Helden, P.D., Koehler, C.J., Thiede, B., Warren, R.M., Wiker, H.G., 2010. Using a label-free proteomics method to identify differentially abundant proteins in closely related hypo- and hypervirulent clinical *Mycobacterium tuberculosis* Beijing isolates. Mol. Cell Proteomics 9, 2414–2423.

DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.A., Hanna, M., McKenna, A., Fennell, T.J., Kernytsky, A.M., Sivachenko, A.Y., Cibulskis, K., Gabriel, S.B., Altshuler, D., Daly, M.J., 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. 43, 491–498.

Devaux, I., Kremer, K., Heersma, H., Van Soolingen, D., 2009. Clusters of multidrug-resistant *Mycobacterium tuberculosis* cases, Europe. Emerging Infect. Dis. 15, 1052–1060.

Domenech, P., Kolly, G.S., Leon-Solis, L., Fallow, A., Reed, M.B., 2010. Massive Gene Duplication Event among Clinical Isolates of the *Mycobacterium tuberculosis* W/Beijing Family. J Bacteriol 192, 4562–4570.

Dos Vultos, T., Mestre, O., Rauzier, J., Golec, M., Rastogi, N., Rasolofo, V., Tonjum, T., Sola, C., Matic, I., Gicquel, B., 2008. Evolution and diversity of clonal bacteria: the paradigm of *Mycobacterium tuberculosis*. PLoS ONE 3, e1538.

Drobniewski, F, Balabanova, Y, Ruddy, M, Weldon, L, Jeltkova, K, Brown, T, Malomanova, N, Elizarova, E, Melenteye, A, Mutovkin, E, Zhakharova, S, Fedorin, I. 2005. Rifampin- and multidrug-resistant Tuberculosis in Russian civilians and prison inmates: dominance of the Beijing strain family. Emerg. Infect. Dis. 8(11): 1320-1326

Dubiley, S., Kirillov, E., Ignatova, A., Stepanshina, V., Shemyakin, I., 2007. Molecular characteristics of the *Mycobacterium tuberculosis* LAM-RUS family prevalent in Central Russia. J. Clin. Microbiol 45, 4036–4038.

Fanning, E., 1994. *Mycobacterium bovis* infection in animals and humans, in: Clinical Tuberculosis. Chapman & Hall, London, pp. 361–365.

Felsenstein, J., 1989. PHYLIP - Phylogeny Inference Package (Version 3.2). Cladistics 5, 164–166.

Feuerriegel, S., Köser, C., Trübe, L., Archer, J., Rüsch Gerdes, S., Richter, E., Niemann, S., 2010. Thr202Ala in thyA is a marker for the Latin American Mediterranean lineage of the *Mycobacterium tuberculosis* complex rather than para-aminosalicylic acid resistance. Antimicrob. Agents Chemother. 54, 4794–4798.

Filliol, I., Driscoll, J.R., Van Soolingen, D., Kreiswirth, B.N., Kremer, K., Valétudie, G., Anh, D.D., Barlow, R., Banerjee, D., Bifani, P.J., Brudey, K., Cataldi, A., Cooksey, R.C., Cousins, D.V., Dale, J.W., Dellagostin, O.A., Drobniewski, F., Engelmann, G., Ferdinand, S., Gascoyne-Binzi, D., Gordon, M., Gutierrez, M.C., Haas, W.H., Heersma, H., Källenius, G., Kassa-Kelembho, E., Koivula, T., Ly, H.M., Makristathis, A., Mammina, C., Martin, G., Moström, P., Mokrousov, I., Narbonne, V., Narvskaya, O., Nastasi, A., Niobe-Eyangoh, S.N., Pape, J.W., Rasolofo-Razanamparany, V., Ridell, M., Rossetti, M.L., Stauffer, F., Suffys, P.N., Takiff, H., Texier-Maugein, J., Vincent, V., De Waard, J.H., Sola, C., Rastogi, N., 2002. Global distribution of *Mycobacterium tuberculosis* spoligotypes. Emerging Infect. Dis. 8, 1347–1349.

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

Filliol, I., Motiwala, A.S., Cavatore, M., Qi, W., Hazbón, M.H., Bobadilla del Valle, M., Fyfe, J., García-García, L., Rastogi, N., Sola, C., Zozio, T., Guerrero, M.I., León, C.I., Crabtree, J., Angiuoli, S., Eisenach, K.D., Durmaz, R., Joloba, M.L., Rendón, A., Sifuentes-Osornio, J., Ponce de León, A., Cave, M.D., Fleischmann, R., Whittam, T.S., Alland, D., 2006. Global phylogeny of *Mycobacterium tuberculosis* based on single nucleotide polymorphism (SNP) analysis: insights into tuberculosis evolution, phylogenetic accuracy of other DNA fingerprinting systems, and recommendations for a minimal standard SNP set. J. Bacteriol. 188, 759–772.

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

Flores, L., Van, T., Narayanan, S., DeRiemer, K., Kato-Maeda, M., Gagneux, S., 2007. Large sequence polymorphisms classify *Mycobacterium tuberculosis* strains with ancestral spoligotyping patterns. J. Clin. Microbiol. 45, 3393–3395.

Ford, C., Yusim, K., Ioerger, T., Feng, S., Chase, M., Greene, M., Korber, B., Fortune, S., 2012. *Mycobacterium tuberculosis* - Heterogeneity revealed through whole genome sequencing. Tuberculosis (Edinburgh, Scotland). 3 194-201

Fortune, S.M., Jaeger, A., Sarracino, D.A., Chase, M.R., Sassetti, C.M., Sherman, D.R., Bloom, B.R., Rubin, E.J., 2005. Mutually dependent secretion of proteins required for mycobacterial virulence. PNAS 102, 10676–10681.

Freeman, Z.N., Dorus, S., Waterfield, N.R., 2013. The KdpD/KdpE Two-Component System: Integrating K+ Homeostasis and Virulence. PLoS Pathog 9, e1003201.

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

Gagneux, S., 2012. Host-pathogen coevolution in human tuberculosis. Philos. Trans. R. Soc. Lond., B, Biol. Sci. 367, 850–859.

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

Gagneux, S., Small, P.M., 2007. Global phylogeography of *Mycobacterium tuberculosis* and implications for tuberculosis product development. The Lancet Infectious Diseases 7, 328–337.

García-Alcalde, F., Okonechnikov, K., Carbonell, J., Cruz, L.M., Götz, S., Tarazona, S., Dopazo, J., Meyer, T.F., Conesa, A., 2012. Qualimap: evaluating next-generation sequencing alignment data. Bioinformatics 28, 2678–2679.

Gardy, J.L., Johnston, J.C., Ho Sui, S.J., Cook, V.J., Shah, L., Brodkin, E., Rempel, S., Moore, R., Zhao, Y., Holt, R., Varhol, R., Birol, I., Lem, M., Sharma, M.K., Elwood, K., Jones, S.J.M., Brinkman, F.S.L., Brunham, R.C., Tang, P., 2011. Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. N. Engl. J. Med. 364, 730–739.

Gibson, A.L., Huard, R.C., Gey van Pittius, N.C., Lazzarini, L.C.O., Driscoll, J., Kurepina, N., Zozio, T., Sola, C., Spindola, S.M., Kritski, A.L., Fitzgerald, D., Kremer, K., Mardassi, H., Chitale, P., Brinkworth, J., Garcia de Viedma, D., Gicquel, B., Pape, J.W., van Soolingen, D., Kreiswirth, B.N., Warren, R.M., van Helden, P.D., Rastogi, N., Suffys, P.N., Lapa e Silva, J., Ho, J.L., 2008. Application of sensitive and specific molecular methods to uncover global dissemination of the major RDRio Sublineage of the Latin American-Mediterranean *Mycobacterium tuberculosis* spoligotype family. J. Clin. Microbiol 46, 1259–1267.

Groenen, P.M., Bunschoten, A.E., van Soolingen, D., van Embden, J.D., 1993. Nature of DNA polymorphism in the direct repeat cluster of *Mycobacterium tuberculosis*; application for strain differentiation by a novel typing method. Mol. Microbiol. 10, 1057–1065.

Guindon, S., Gascuel, O., 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. 52, 696–704.

Gutacker, M.M., Mathema, B., Soini, H., Shashkina, E., Kreiswirth, B.N., Graviss, E.A., Musser, J.M., 2006. Single-nucleotide polymorphism-based population genetic analysis of *Mycobacterium tuberculosis* strains from 4 geographic sites. J. Infect. Dis. 193, 121–128.

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

Hanekom, M., Gey van Pittius, N.C., McEvoy, C., Victor, T.C., Van Helden, P.D., Warren, R.M., 2011. *Mycobacterium tuberculosis* Beijing genotype: a template for success. Tuberculosis (Edinb) 91, 510–523.

Hanekom, M., van der Spuy, G.D., Gey van Pittius, N.C., McEvoy, C.R.E., Hoek, K.G.P., Ndabambi, S.L., Jordaan, A.M., Victor, T.C., van Helden, P.D., Warren, R.M., 2008. Discordance between mycobacterial interspersed repetitive-unit-variable-number tandem-repeat typing and IS6110 restriction fragment length polymorphism genotyping for analysis of *Mycobacterium tuberculosis* Beijing strains in a setting of high incidence of tuberculosis. J. Clin. Microbiol. 46, 3338–3345.

Hermans, P.W., van Soolingen, D., Bik, E.M., de Haas, P.E., Dale, J.W., van Embden, J.D., 1991. Insertion element IS987 from *Mycobacterium bovis* BCG is located in a hot-spot integration region for insertion elements in *Mycobacterium tuberculosis* complex strains. Infect. Immun. 59, 2695–2705.

Hershberg, R., Lipatov, M., Small, P.M., Sheffer, H., Niemann, S., Homolka, S., Roach, J.C., Kremer, K., Petrov, D.A., Feldman, M.W., Gagneux, S., 2008. High Functional Diversity in *Mycobacterium tuberculosis* Driven by Genetic Drift and Human Demography. PLoS Biol 6, e311.

Hirsh, A.E., Tsolaki, A.G., DeRiemer, K., Feldman, M.W., Small, P.M., 2004. Stable association between strains of *Mycobacterium tuberculosis* and their human host populations. Proc. Natl. Acad. Sci. U.S.A 101, 4871–4876.

Homer, N., Merriman, B., Nelson, S.F., 2009. BFAST: an alignment tool for large scale genome resequencing. PLoS ONE 4, e7767.

Huard, R.C., Fabre, M., de Haas, P., Lazzarini, L.C.O., van Soolingen, D., Cousins, D., Ho, J.L., 2006. Novel genetic polymorphisms that further delineate the phylogeny of the *Mycobacterium tuberculosis* complex. J. Bacteriol. 188, 4271–4287.

Huard, R.C., Lazzarini, L.C. de O., Butler, W.R., van Soolingen, D., Ho, J.L., 2003. PCR-based method to differentiate the subspecies of the *Mycobacterium tuberculosis* complex on the basis of genomic deletions. J. Clin. Microbiol. 41, 1637–1650.

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

Ioerger, T.R., Koo, S., No, E.-G., Chen, X., Larsen, M.H., Jacobs, W.R., Jr, Pillay, M., Sturm, A.W., Sacchettini, J.C., 2009. Genome analysis of multi- and extensively-drug-resistant tuberculosis from KwaZulu-Natal, South Africa. PLoS ONE 4, e7778.

Joshi, S.A., Ball, D.A., Sun, M.G., Carlsson, F., Watkins, B.Y., Aggarwal, N., McCracken, J.M., Huynh, K.K., Brown, E.J., 2012. EccA1, a component of the *Mycobacterium* marinum ESX-1 protein virulence factor secretion pathway, regulates mycolic acid lipid synthesis. Chem. Biol. 19, 372–380.

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

Kato-Maeda, M., Rhee, J.T., Gingeras, T.R., Salamon, H., Drenkow, J., Smittipat, N., Small, P.M., 2001. Comparing genomes within the species *Mycobacterium tuberculosis*. Genome Res. 11, 547–554.

Keane, T.M., Creevey, C.J., Pentony, M.M., Naughton, T.J., McInerney, J.O., 2006. Assessment of methods for amino acid matrix selection and their use on empirical data shows that ad hoc assumptions for choice of matrix are not justified. BMC Evolutionary Biology 6, 29.

Lamichhane, G., Zignol, M., Blades, N.J., Geiman, D.E., Dougherty, A., Grosset, J., Broman, K.W., Bishai, W.R., 2003. A postgenomic method for predicting essential genes at subsaturation levels of mutagenesis: application to *Mycobacterium tuberculosis*. Proc. Natl. Acad. Sci. U.S.A. 100, 7213–7218.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., Thompson, J.D., Gibson, T.J., Higgins, D.G., 2007. Clustal W and Clustal X version 2.0. Bioinformatics 23, 2947–2948.

Lazzarini, L.C.O., Huard, R.C., Boechat, N.L., Gomes, H.M., Oelemann, M.C., Kurepina, N., Shashkina, E., Mello, F.C.Q., Gibson, A.L., Virginio, M.J., Marsico, A.G., Butler, W.R., Kreiswirth, B.N., Suffys, P.N., Lapa E Silva, J.R., Ho, J.L., 2007. Discovery of a novel *Mycobacterium tuberculosis* lineage that is a major cause of tuberculosis in Rio de Janeiro, Brazil. J. Clin. Microbiol. 45, 3891–3902.

Lazzarini, L.C.O., Spindola, S.M., Bang, H., Gibson, A.L., Weisenberg, S., da Silva Carvalho, W., Augusto, C.J., Huard, R.C., Kritski, A.L., Ho, J.L., 2008. RDRio *Mycobacterium tuberculosis* infection is associated with a higher frequency of cavitary pulmonary disease. J. Clin. Microbiol 46, 2175–2183.

Letunic, I., Yamada, T., Kanehisa, M., Bork, P., 2008. iPath: interactive exploration of biochemical pathways and networks. Trends Biochem. Sci. 33, 101–103.

Lew, J.M., Kapopoulou, A., Jones, L.M., Cole, S.T., 2011. TubercuList--10 years after. Tuberculosis (Edinb) 91, 1–7.

Lewis, K.N., Liao, R., Guinn, K.M., Hickey, M.J., Smith, S., Behr, M.A., Sherman, D.R., 2003. Deletion of RD1 from *Mycobacterium tuberculosis* mimics bacille Calmette-Guérin attenuation. J. Infect. Dis. 187, 117–123.

Li, H., Durbin, R., 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754–1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079.

Lipsitch, M., Moxon, E.R., 1997. Virulence and transmissibility of pathogens: what is the relationship? Trends Microbiol. 5, 31–37.

Manabe, Y.C., Dannenberg, A.M., Tyagi, S.K., Hatem, C.L., Yoder, M., Woolwine, S.C., Zook, B.C., Pitt, M.L.M., Bishai, W.R., 2003. Different Strains of *Mycobacterium tuberculosis* Cause Various Spectrums of Disease in the Rabbit Model of Tuberculosis. Infect. Immun. 71, 6004–6011.

Marmiesse, M., Brodin, P., Buchrieser, C., Gutierrez, C., Simoes, N., Vincent, V., Glaser, P., Cole, S.T., Brosch, R., 2004. Macro-array and bioinformatic analyses reveal mycobacterial “core” genes, variation in the ESAT-6 gene family and new phylogenetic markers for the *Mycobacterium tuberculosis* complex. Microbiology 150, 483–496.

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

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., DePristo, M.A., 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20, 1297–1303.

Mestre, O., Luo, T., Dos Vultos, T., Kremer, K., Murray, A., Namouchi, A., Jackson, C., Rauzier, J., Bifani, P., Warren, R., Rasolofo, V., Mei, J., Gao, Q., Gicquel, B., 2011. Phylogeny of *Mycobacterium tuberculosis* Beijing strains constructed from polymorphisms in genes involved in DNA replication, recombination and repair. PLoS ONE 6, e16020.

Migliori, G.B., D’Arcy Richardson, M., Sotgiu, G., Lange, C., 2009. Multidrug-resistant and extensively drug-resistant tuberculosis in the West. Europe and United States: epidemiology, surveillance, and control. Clin. Chest Med. 30, 637–665

Miyoshi-Akiyama, T., Matsumura, K., Iwai, H., Funatogawa, K., Kirikae, T., 2012. Complete Annotated Genome Sequence of *Mycobacterium tuberculosis* Erdman. J. Bacteriol. 194, 2770–2770.

Mokrousov, I, Ly, H.M., Otten, T, Lan N.N., Vyshnevskyi, B, Narvskaya, O. 2005. Origin and primary dispersal of the *Mycobacterium tuberculosis* Beijing genotype: clues from human phylogeography. Genome Res.15(10): 1357–1364

Mostowy, S., Behr, M.A., 2005. The origin and evolution of *Mycobacterium tuberculosis*. Clin. Chest Med. 26, 207–216.

Moström, P., Gordon, M., Sola, C., Ridell, M., Rastogi, N., 2002. Methods used in the molecular epidemiology of tuberculosis. Clin. Microbiol. Infect. 8, 694–704.

Mulenga, C., Shamputa, I.C., Mwakazanga, D., Kapata, N., Portaels, F., Rigouts, L., 2010. Diversity of *Mycobacterium tuberculosis* genotypes circulating in Ndola, Zambia. BMC Infect. Dis. 10, 177.

Musser, J.M., Amin, A., Ramaswamy, S., 2000. Negligible genetic diversity of *Mycobacterium tuberculosis* host immune system protein targets: evidence of limited selective pressure. Genetics 155, 7–16.

Nachega, J.B., Chaisson, R.E., 2003. Tuberculosis Drug Resistance: A Global Threat. Clinical Infectious Diseases 36, S24–S30.

Nambu, S., Matsui, T., Goulding, C.W., Takahashi, S., Ikeda-Saito, M., 2013. A new way to degrade heme: the *Mycobacterium tuberculosis* enzyme MhuD catalyzes heme degradation without generating CO. J. Biol. Chem. 288, 10101–10109.

Namouchi, A., Didelot, X., Schöck, U., Gicquel, B., Rocha, E.P.C., 2012. After the bottleneck: Genome-wide diversification of the *Mycobacterium tuberculosis* complex by mutation, recombination, and natural selection. Genome Research. 4, 721-734

Nei, M, Kumar, S. 2000. Molecular Evolution and Phylogenetics. Oxford University Press, New York. 333pp.

Pan, Y., Yang, X., Duan, J., Lu, N., Leung, A.S., Tran, V., Hu, Y., Wu, N., Liu, D., Wang, Z., Yu, X., Chen, C., Zhang, Y., Wan, K., Liu, J., Zhu, B., 2011. Whole-genome sequences of four *Mycobacterium bovis* BCG vaccine strains. J. Bacteriol. 193, 3152–3153.

Parish, T., Smith, D.A., Roberts, G., Betts, J., Stoker, N.G., 2003. The senX3-regX3 two-component regulatory system of *Mycobacterium tuberculosis* is required for virulence. Microbiology (Reading, Engl.) 149, 1423–1435.

Parsons, L.M., Brosch, R., Cole, S.T., Somoskövi, A., Loder, A., Bretzel, G., Van Soolingen, D., Hale, Y.M., Salfinger, M., 2002. Rapid and simple approach for identification of *Mycobacterium tuberculosis* complex isolates by PCR-based genomic deletion analysis. J. Clin. Microbiol. 40, 2339–2345.

Parsons, S.D.C., Drewe, J., Warren, R., Gey Van Pittius, N.C., Van Helden, P.D., 2013. A novel pathogen, *Mycobacterium suricattae*, is the cause of tuberculosis in meerkats (Suricata suricatta,) in South Africa. Emerging Infect. Dis. Ahead of print.

Pepperell, C.S., Casto, A.M., Kitchen, A., Granka, J.M., Cornejo, O.E., Holmes, E.C., Birren, B., Galagan, J., Feldman, M.W., 2013. The Role of Selection in Shaping Diversity of Natural *M. tuberculosis* Populations. PLoS Pathog 9, e1003543.

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

Pym, A.S, Brosch, R. 2000. Tools for the population genomics of the tubercle bacilli. Genome Res. 10(12): 1837-1839

Quinlan, A.R., Hall, I.M., 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841–842.

R Development Core Team. 2005. R: A Language and Environment for Statistical Computing. URL http://www.R-project.org

Reed, M.B., Pichler, V.K., McIntosh, F., Mattia, A., Fallow, A., Masala, S., Domenech, P., Zwerling, A., Thibert, L., Menzies, D., Schwartzman, K., Behr, M.A., 2009. Major *Mycobacterium tuberculosis* Lineages Associate with Patient Country of Origin. J Clin Microbiol 47, 1119–1128.

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

Ritacco, V., Iglesias, M.-J., Ferrazoli, L., Monteserin, J., Dalla Costa, E.R., Cebollada, A., Morcillo, N., Robledo, J., de Waard, J.H., Araya, P., Aristimuño, L., Díaz, R., Gavin, P., Imperiale, B., Simonsen, V., Zapata, E.M., Jiménez, M.S., Rossetti, M.L., Martin, C., Barrera, L., Samper, S., 2011. Conspicuous multidrug-resistant *Mycobacterium tuberculosis* cluster strains do not trespass country borders in Latin America and Spain. Infection, Genetics and Evolution: Journal of Molecular Epidemiology and Evolutionary Genetics in Infectious Diseases. 4, 711-717

Roetzer, A., Diel, R., Kohl, T.A., Ruckert, C., Nubel, U., Blom, J., Wirth, T., Jaenicke, S., Schuback, S., Rüsch-Gerdes, S., Supply, P., Kalinowski, J., Niemann, S., 2013. Whole Genome Sequencing versus Traditional Genotyping for Investigation of a *Mycobacterium tuberculosis* Outbreak: A Longitudinal Molecular Epidemiological Study. PLoS Med 10(2): e1001387.

Rosas-Magallanes, V., Stadthagen-Gomez, G., Rauzier, J., Barreiro, L.B., Tailleux, L., Boudou, F., Griffin, R., Nigou, J., Jackson, M., Gicquel, B., Neyrolles, O., 2007. Signature-tagged transposon mutagenesis identifies novel *Mycobacterium tuberculosis* genes involved in the parasitism of human macrophages. Infect. Immun. 75, 504–507.

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

Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M.-A., Barrell, B., 2000. Artemis: sequence visualization and annotation. Bioinformatics 16, 944–945.

Salie, M, van der Merwe, L, Möller, M, Daya, M., van der Spuy, G.D, van Helden, P.D, Martin, M.P, Gao, X.J, Warren, R.M, Carrington, M, Hoal, E.G. 2013. Associations between human leukocyte antigen class I variants and the *Mycobacterium tuberculosis* subtypes causing disease. J. Infect. Dis. Epub ahead of print.

Sandgren, A., Strong, M., Muthukrishnan, P., Weiner, B.K., Church, G.M., Murray, M.B., 2009. Tuberculosis Drug Resistance Mutation Database. PLoS Med 6, e1000002.

Sarkar, R., Lenders, L., Wilkinson, K.A., Wilkinson, R.J., Nicol, M.P., 2012. Modern Lineages of *Mycobacterium tuberculosis* Exhibit Lineage-Specific Patterns of Growth and Cytokine Induction in Human Monocyte-Derived Macrophages. PLoS ONE 7, e43170.

Sassetti, C.M., Boyd, D.H., Rubin, E.J., 2003. Genes required for mycobacterial growth defined by high density mutagenesis. Mol. Microbiol. 48, 77–84.

Schaaf, H.S., Victor, T.C., Venter, A., Brittle, W., Jordaan, A.M., Hesseling, A.C., Marais, B.J., van Helden, P.D., Donald, P.R., 2009. Ethionamide cross- and co-resistance in children with isoniazid-resistant tuberculosis. Int. J. Tuberc. Lung Dis. 13, 1355–1359.

Schmieder, R., Edwards, R., 2011. Quality control and preprocessing of metagenomic datasets. Bioinformatics 27, 863–864.

Schürch, A.C., van Soolingen, D., 2011. DNA fingerprinting of *Mycobacterium tuberculosis*: From phage typing to whole-genome sequencing. Infect. Genet. Evol. 4, 602-9

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

Shemyakin, I.G., Stepanshina, V.N., Ivanov, I.Y., Lipin, M.Y., Anisimova, V.A., Onasenko, A.G., Korobova, O.V., Shinnick, T.M., 2004. Characterization of drug-resistant isolates of *Mycobacterium tuberculosis* derived from Russian inmates. Int. J. Tuberc. Lung Dis. 8, 1194–1203.

Smith, T. 1898. A comparative study of bovine tubercule bacilli and of human bacilli from sputum. J. Exp. Med. 3: 451-511

Sola, C., Filliol, I., Legrand, E., Mokrousov, I., Rastogi, N., 2001. *Mycobacterium tuberculosis* phylogeny reconstruction based on combined numerical analysis with IS1081, IS6110, VNTR, and DR-based spoligotyping suggests the existence of two new phylogeographical clades. J. Mol. Evol. 53, 680–689.

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

Stamatakis, A., 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22, 2688–2690.

Streicher, E.M., 2007. Application of spoligotyping in the understanding of the dynamics of *Mycobacterium tuberculosis* strains in high incidence communities. PhD Thesis. Stellenbosch University. Tygerberg, South Africa

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

Tamura, K, Peterson, D, Peterson, N, Stecher, G, Nei, M, Kumar, S. 2011. MEGA5: Molecular Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. Mol. Biol. Evol. 28(10):2731–2739. 2011

Thwaites, G., Caws, M., Chau, T.T.H., D'Sa, A., Lan, N.T.N., Huyen, M.N.T., Gagneux, S., Anh, P.T.H., Tho, D.Q., Torok, E., Nhu, N.T.Q., Duyen, N.T.H., Duy, P.M., Richenberg, J., Simmons, C., Hien, T.T., Farrar, J., 2008. Relationship between *Mycobacterium tuberculosis* genotype and the clinical phenotype of pulmonary and meningeal tuberculosis. J. Clin. Microbiol. 46, 1363–1368.

Tsolaki, A.G., Hirsh, A.E., DeRiemer, K., Enciso, J.A., Wong, M.Z., Hannan, M., Goguet de la Salmoniere, Y.-O.L., Aman, K., Kato-Maeda, M., Small, P.M., 2004. Functional and evolutionary genomics of *Mycobacterium tuberculosis*: insights from genomic deletions in 100 strains. Proc. Natl. Acad. Sci. U.S.A. 101, 4865–4870.

Tucker, P.A., Nowak, E., Morth, J.P., 2007. Two-Component Systems of *Mycobacterium tuberculosis*—Structure-Based Approaches, in: Melvin I. Simon, B.R.C. and A.C. (Ed.), Methods in Enzymology, Two-Component Signaling Systems, Part B. Academic Press, pp. 477–501.

Van der Spuy, G.D, Kremer, K, Ndabambi, S.L, Beyers, N, Dunbar, R, Marais, B.J, van Helden, P.D, Warren, R.M. 2009. Changing *Mycobacterium tuberculosis* population highlights clade specific pathogenic characteristics. Tuberculosis. 89(2):120-5.

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

Van Ingen, J., Rahim, Z., Mulder, A., Boeree, M.J., Simeone, R., Brosch, R., van Soolingen, D., 2012. Characterization of *Mycobacterium orygis* as *M. tuberculosis* complex subspecies. Emerging Infect. Dis. 18, 653–655.

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

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

Victor, T.C., de Haas, P.E.W., Jordaan, A.M., van der Spuy, G.D., Richardson, M., van Soolingen, D., van Helden, P.D., Warren, R., 2004. Molecular characteristics and global spread of *Mycobacterium tuberculosis* with a western cape F11 genotype. J. Clin. Microbiol. 42, 769–772.

Victor, T. 2005. A first for Tuberculosis research in South Africa: whole-genome sequence of the South African *Mycobacterium tuberculosis* strain F11 released. S. Afr. J. Sci.

Viegas, S.O., Machado, A., Groenheit, R., Ghebremichael, S., Pennhag, A., Gudo, P.S., Cuna, Z., Miotto, P., Hill, V., Marrufo, T., Cirillo, D.M., Rastogi, N., Källenius, G., Koivula, T., 2010. Molecular diversity of *Mycobacterium tuberculosis* isolates from patients with pulmonary tuberculosis in Mozambique. BMC Microbiol. 10, 195.

Von Groll, A., Martin, A., Felix, C., Prata, P.F.S., Honscha, G., Portaels, F., Vandame, P., da Silva, P.E.A., Palomino, J.C., 2010. Fitness study of the RDRio lineage and Latin American-Mediterranean family of *Mycobacterium tuberculosis* in the city of Rio Grande, Brazil. FEMS Immunol. Med. Microbiol. 58, 119–127.

Walker, T.M., Ip, C.L.C., Harrell, R.H., Evans, J.T., Kapatai, G., Dedicoat, M.J., Eyre, D.W., Wilson, D.J., Hawkey, P.M., Crook, D.W., Parkhill, J., Harris, D., Walker, A.S., Bowden, R., Monk, P., Smith, E.G., Peto, T.E.A., 2013. Whole-genome sequencing to delineate *Mycobacterium tuberculosis* outbreaks: a retrospective observational study. Lancet Infect Dis 13, 137–146.

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

Warren, R.M., Gey van Pittius, N.C., Barnard, M., Hesseling, A., Engelke, E., de Kock, M., Gutierrez, M.C., Chege, G.K., Victor, T.C., Hoal, E.G., van Helden, P.D., 2006. Differentiation of *Mycobacterium tuberculosis* complex by PCR amplification of genomic regions of difference. Int. J. Tuberc. Lung Dis. 10, 818–822.

Warren, R.M., Streicher, E.M., Charalambous, S., Churchyard, G., van der Spuy, G.D., Grant, A.D., van Helden, P.D., Victor, T.C., 2002. Use of spoligotyping for accurate classification of recurrent tuberculosis. J. Clin. Microbiol. 40, 3851–3853.

Weiner, B., Gomez, J., Victor, T.C., Warren, R.M., Sloutsky, A., Plikaytis, B.B., Posey, J.E., van Helden, P.D., Gey van Pittius, N.C., Koehrsen, M., Sisk, P., Stolte, C., White, J., Gagneux, S., Birren, B., Hung, D., Murray, M., Galagan, J., 2012. Independent Large Scale Duplications in Multiple *M. tuberculosis* Lineages Overlapping the Same Genomic Region. PLoS ONE 7, e26038.

Weisenberg, S.A., Gibson, A.L., Huard, R.C., Kurepina, N., Bang, H., Lazzarini, L.C.O., Chiu, Y., Li, J., Ahuja, S., Driscoll, J., Kreiswirth, B.N., Ho, J.L., 2011. Distinct clinical and epidemiological features of tuberculosis in New York City caused by the RD(Rio)*Mycobacterium tuberculosis* sublineage. Infect. Genet. Evol.

WHO, 2011. Global Tuberculosis control 2011. URL http://www.who.int/tb/publications/global_report/2011/en/ (accessed 5.26.12).

WHO, 2012. Global Tuberculosis Control 2012. URL www.who.int/iris/bitstream/10665/75938/1/9789241564502_eng.pdf

Williams, M.J., Kana, B.D., Mizrahi, V., 2011. Functional analysis of molybdopterin biosynthesis in mycobacteria identifies a fused molybdopterin synthase in *Mycobacterium tuberculosis*. J. Bacteriol. 193, 98–106.

Wirth, T., Hildebrand, F., Allix-Béguec, C., Wölbeling, F., Kubica, T., Kremer, K., van Soolingen, D., Rüsch-Gerdes, S., Locht, C., Brisse, S., Meyer, A., Supply, P., Niemann, S., 2008. Origin, spread and demography of the *Mycobacterium tuberculosis* complex. PLoS Pathog. 4, e1000160.

Yamada, T., Letunic, I., Okuda, S., Kanehisa, M., Bork, P., 2011. iPath2.0: interactive pathway explorer. Nucleic Acids Res. 39, W412–415.

Zhang, H., Wang, J., Lei, J., Zhang, M., Yang, Y., Chen, Y., Wang, H., 2007. PPE protein (Rv3425) from DNA segment RD11 of *Mycobacterium tuberculosis*: a potential B-cell antigen used for serological diagnosis to distinguish vaccinated controls from tuberculosis patients. Clin. Microbiol. Infect. 13, 139–145.

Zhang, Y., Gladyshev, V.N., 2008. Molybdoproteomes and evolution of molybdenum utilization. J. Mol. Biol. 379, 881–899.