Selection and identification of novel *Mycobacterium tuberculosis* phage-displayed biomarkers by immunoscreening against patients' serum samples

By

Thamsanqa Emmanuel Chiliza

Submitted in partial fulfilment of the academic requirements of Doctor of Philosophy in Microbiology, School of Life Sciences, College of Agriculture, Engineering and Science, University of KwaZulu-Natal

Supervisor: Prof. Balakrishna Pillay

Co-Supervisor: Prof. Manormoney Pillay

August 2018

PREFACE

The research contained in this thesis was completed by the candidate while based in the Discipline of Microbiology, School of Life Sciences of the College of Agriculture, Engineering and Science, Westville Campus, University of KwaZulu-Natal, South Africa. The research was financially supported by the National Research Foundation, the School of Life Science and the University's Research Office.

The contents of this work have not been submitted in any form to another university and, except where the work of others is acknowledged in the text, the results reported are due to investigations by the candidate.

_____ _____

Supervisor: **Prof. Bala Pillay** **Co-supervisor: Prof. M. Pillay**

Date: _____ Date: _____

ETHICAL STATEMENT

This study was approved by the Biomedical Research Ethics Committee of the University of KwaZulu-Natal (BE236/13).

DECLARATION 1: PLAGIARISM

I, Thamsanqa Emmanuel Chiliza, declare that:

(i) the research reported in this thesis, except where otherwise indicated or acknowledged, is my original work;

(ii) this thesis has not been submitted in full or in part for any degree or examination to any other university;

(iii) this thesis does not contain other persons' data, pictures, graphs or other information, unless specifically acknowledged as being sourced from other persons;

(iv) this thesis does not contain other persons' writing, unless specifically acknowledged as being sourced from other researchers. Where other written sources have been quoted, then:

a) their words have been re-written but the general information attributed to them has been referenced;

b) where their exact words have been used, their writing has been placed inside quotation marks, and referenced;

(v) where I have used material for which publications followed, I have indicated in detail my role in the work;

(vi) this thesis is primarily a collection of material, prepared by myself, published as journal articles or presented as a poster and oral presentations at conferences. In some cases, additional material has been included;

(vii) this thesis does not contain text, graphics or tables copied and pasted from the Internet, unless specifically acknowledged, and the source being detailed in the thesis and in the References sections.

_____

Signed: **TE Chiliza**

Date: _____

DECLARATION 2: PUBLICATIONS AND PRESENTATIONS

I, Thamsanqa Emmanuel Chiliza, declare that I performed all experiments, collected and analysed data. Therefore the work reported on publications and presentations listed below is mainly my original work.

PUBLICATIONS

1. **Chiliza, T.E**., Pillay, M. and Pillay, B*. 2017. Identification of unique essential proteins from a *Mycobacterium tuberculosis* F15/LAM4/KZN phage secretome library. *Pathogens and Disease* 75(1), ftx001, https://doi.org/10.1093/femspd/ftx001.

   Authors' contributions

   My supervisors, Prof B. Pillay and Prof M. Pillay and I designed the study. I conducted all the experimental work, processed and analyzed the data and wrote the manuscript. My supervisors provided valuable input through reviewing all manuscript drafts and providing critical comments.

2. **Chiliza, T.E.**, Pillay, M., Naidoo K. and Pillay, B*. 2018. Immunoscreening of the *M. tuberculosis* F15/LAM4/KZN secretome library against TB patients' sera identifies unique active- and latent-TB specific biomarkers. *Tuberculosis*, reviewers assigned by the journal. Submitted on the 19th of January 2018 and the Manuscript number is TUBE_2018_11

   Authors' contributions

   My supervisors, Prof B. Pillay and Prof M. Pillay and I designed the study. Dr. K. Naidoo established study participants recruitment team and assisted with study protocol design and implementation at recruitment site. I conducted all the experimental work, processed and analyzed the data and wrote the manuscript. My supervisors provided valuable input through reviewing all manuscript drafts and providing critical comments.

3. **Chiliza, T.E**., Pillay, M. and Pillay, B*. 2018. B-cell epitope derived T-cell epitopes as vaccine candidates to confer antibody and cellular mediated immunity against *M. tuberculosis* infection. Manuscript in preparation.

   Authors' contributions

   I designed the study and I conducted all the experimental work, processed and analyzed the data and wrote the manuscript. My supervisors, Prof B. Pillay and Prof M. Pillay provided valuable input through reviewing all manuscript drafts and providing critical comments.

CONFERENCE PRESENTATIONS

4. **Chiliza, T.E**., Pillay, M. and Pillay, B. 2016. Selective identification of novel secretory *M. tuberculosis* complex biomarkers using phage-display technique. Paper presentation to the South African Society for Microbiology Biennial Congress 2016, 17th to 20th January, 2016, Durban, South Africa. **Oral presentation**.

5. **Chiliza, T.E.**, Pillay, M., Naidoo K. and Pillay, B. 2016. Selective identification of latent- and active-TB biomarkers using *Mycobacterium tuberculosis* secretome phage library and immunoscreening. Poster presentation to the 26th European Congress of Clinical Microbiology and Infectious Diseases, 9th to 12th April 2016, Amsterdam, Netherlands. **Poster presentation.**

6. **Chiliza, T.E.**, Pillay, M., Naidoo K. and Pillay, B. 2017. Immunoscreening of *M. tuberculosis* F15/LAM4/KZN secretome library against TB patients' sera and identification of latent-TB and active-TB specific biomarkers. Paper presentation to the International Union of Microbiological Societies 2017 (IUMS 2017), 17th to 21st July 2017, Sands Convention and Exhibition Centre, Singapore. **Oral presentation.**

_____

Signed: **TE Chiliza**

Date: _____

ABSTRACT

Mycobacterium tuberculosis (*Mtb*) derived protein biomarkers are urgently needed for the development of point-of-care diagnostics, new drugs and vaccines for improved management of TB. *Mtb* extracellular and secreted proteins facilitate host invasion, interfere with host defence mechanisms during host-pathogen interaction and are essential for growth and survival of *Mtb* during infection. Therefore, the present study aimed to identify unique *Mtb* secretory proteins that can be used to diagnose and differentiate ATB from LTBI, design new drugs for treatment of drug-resistant TB and develop effective TB vaccines. Using the phage display system specifically designed to study the bacterial secretome, a whole genome library was constructed from genomic DNA of the XDR *Mtb* F15/LAM4/KZN strain. The analysis of *Mtb* secretome sub-library (~8×10^3 clones) by DNA sequencing confirmed that the library consisted mainly (>90%) of extracellular proteins including secreted and cell wall associated proteins. The *Mtb* phage secretome sub-library was screened by biopanning against immobilized polyclonal sera from TB negative (*n*=20) individuals, ATB (*n*=20) and LTBI (*n*=15) patients, in order to identify proteins recognized by TB patients' antibodies. DNA sequence analysis of randomly selected ATB and LTBI phage clones revealed 118 and 96 open reading frames, respectively. Of these, 23 proteins overlapped between ATB and LTBI, including the resuscitation-promoting factor (RpfB) which promotes the resuscitation and growth of dormant cells, and the immunogenic proteins, e.g., Ag85B, Mpt63 and Mpt64. Using different *Mtb* databases, proteins essential for growth, virulence and metabolic processes were identified as good targets for development of diagnostic tools, new drugs and vaccines. The ATB-specific biomarkers included TrpG, Alr, TreY, BfrA and EspR, with no human homologs. The LTBI-specific biomarkers included NarG, PonA1, PonA2 and HspR, which are known to be involved in stationary-phase survival under non-replicating conditions. Reverse vaccinology was used to

analyse all ATB-specific and LTBI-specific proteins, 40 proteins were selected and ranked according to predicted subcellular localization, transmembrane domains, adhesive properties and antigenicity. B-cell and T-cell ($CD8^+$/$CD4^+$) epitopes were identified for the design of a new polypeptide TB vaccine. Two novel $CD8^+$ T-cell epitopes, RMPTGMPPK (Rv0361) and QLPPTDPRY (Rv3682) with binding affinity for HLA-E*01:01 and HLA-E*01:03 alleles respectively were identified. Since HLA-E alleles are not down-regulated by HIV co-infection and since HLA-E epitopes demonstrated the highest population coverage (99.88%) among the world's population, these epitopes are attractive TB vaccine candidates. In conclusion, antigenic *Mtb* derived biomarkers essential for *in vivo* growth, intracellular survival and virulence were successfully identified in this study. The B- and T-cells epitopes identified represent potential candidates that may invoke both humoral and cellular immune responses. These findings will potentially accelerate and advance the design of effective and cost-effective vaccines and diagnostic tests against *Mtb* infection. Future studies will evaluate the potential application of these biomarkers in TB diagnosis, monitoring and prognosis of TB disease, and for the design of effective vaccines against this dreaded disease.

ACKNOWLEDGEMENTS

I would like to take this opportunity to extend my sincere gratitude and appreciation to all those who made this PhD thesis possible.

Firstly, I would like to express my sincere gratitude to my supervisor, **Prof. Balakrishna Pillay** for entertaining and nurturing my ideas, for his dedicated help, advice, inspiration, encouragement and continuous support, throughout my PhD. Secondly, **Prof. Manormoney Pillay** my co-supervisor, for the continuous support of my PhD study and related research activities, for her patience, motivation, and immense knowledge. Her guidance helped me throughout my research and writing of this thesis. I could not have imagined having a better advisor and mentor.

My sincere thanks also goes to **Dr. Kogieleum Naidoo** who provided me with the opportunity to work with her team, and who provided access to the CAPRISA eThekwini Clinical Research Site. Without her precious support it would not be possible to conduct this research. I would also like to acknowledge and thank the following CAPRISA staff members who made this research possible: Vinotha Naik, Dhineshree Govender, Phindile Sing, Thandiwe Sithole, Senzo Hlathi, Barnabas Khumalo, Goodness Gumede, Zanele Gwamanda, Dudu Dlamini, Refiwe Sondezi, Gloria Ntanjana, Bhavna Maharaj and Anushka Naidoo.

I would like to thank **Dr. Jasna Rakonjac** from the Institute for Molecular Biosciences, Massey University, New Zealand, for the donation of the pDJ01 phage display system.

To Prof. Samson Mukaratirwa and Prof. Ademola Olaniran, thank you for the support and encouragement. To Dr. Martin Bubb thank you for always reminding me of my career aspirations and encouraging me to pursue my PhD studies.

Thank you to Dr. Charissa Naidoo for her assistance and help during *Mtb* genomic DNA isolation; and Dr. Nontobeko Mvubu, Dr. Chika Nnadozie, Dr. Natasha Naidoo and Dr. Oliver Zishiri for their support and words of encouragement. To Lab 1 family, Charlotte Ramadhin, Sphelele Sibisi, Koobashnee Pillay, Suventha Moodley, Nombali Gumede and Microbiology

Discipline staff and students who indirectly contributed to this research, your kindness means a lot to me. Thank you very much.

Thanks to my wife, Gugu Chiliza for taking on my family responsibilities while I was immersed in my studies. My two daughters, Owethu for always asking me when am I finishing my studies and Sbongakonke for being there at night during the thesis write-up.

Many thanks to my family especially my mother, Mrs. Misiwe Chiliza for believing in me and encouraging me to pursue my passion, studying. My brother Musa for being there and supporting me throughout my student life. My sisters Khanyisile, Zanele, and my late brother Themba a.k.a. Professor for their constant support and encouragement.

I would like to thank the National Research Foundation (NRF) of South Africa and the University of KwaZulu-Natal Research Office and College of Agriculture, Engineering and Science, for financial support.

To My God, UShembe, UNyazi LweZulu for the strength and His blessing in completing this thesis. “Dumisani uJehova Amen!”

TABLE OF CONTENTS

Page

LIST OF TABLES

Table **Page**

LIST OF FIGURES

Figure **Page**

LIST OF ABBREVIATIONS

Abbreviation	Abbreviated term
AIDS	acquired immunodeficiency syndrome
ATB	active tuberculosis
BCG	Bacillus Calmette–Guerin (attenuated strain of *Mycobacterium bovis*)
BLAST	Basic Local Alignment Search Tool
bp	base pairs
CDS	coding sequence
CFP-10	culture filtrate protein of 10 kDa
cfu	colony forming units
Cm	chloramphenicol
CTAB	Cetyltrimethylammoniumbromide
DAVID	database for annotation, visualization and integrated discovery
DNA	deoxyribonucleic acid
dNTP	deoxydribonucleotide triphosphate
EDTA	ethylenediaminetetraacetic acid
eg.	example
ELISA	Enzyme Linked Immunosorbent Assay
ESAT-6	early secreted antigenic target of 6 kDa
ESX	ESAT-6 like secretion system
GO	gene ontology
HIV	Human Immunodeficiency Virus
HLA	human leukocyte antigen
hr	hour
i.e.	that is
IEDB	Immune Epitope Database and Analysis Resource
KAN	kanamycin
kb	kilobase
kbp	kilobase pair

kDa	kilodalton
KEGG	Kyoto Encyclopedia of Genes and Genomes
KZN	KwaZulu-Natal
KZN605	F15/LAM4/KZN strain
L	litre
LTBI	latent-TB infection
M	molar
Mbp	megabase pair
MDR	multidrug resistant
MDR-TB	multidrug resistant TB
mg	milligram
$MgCl_2$	magnesium chloride
MHC	major histocompatibility complex
min	minute
mL	millilitre
mm	millimetre
mM	millimolar
mmol	millimoles
MOI	multiplicity of infection
MPPs	metabolic pathway gene encoded proteins
MPTR	major polymorphic tandem repeats
Mtb	*Mycobacterium tuberculosis*
***MTB*C**	*Mycobacterium tuberculosis* complex
N- terminal	amino-terminus of a peptide or protein
NaCl	sodium chloride
ng	nanogram
OD	optical density
OD_{600}	optical density at 600nm
ORF	open reading frame
PATRIC	Pathosystems Resource Integration Center

PBS	phosphate buffered saline
PBST	phosphate buffered saline + Tween 20
PBSTM	phosphate buffered saline + Tween 20 + non-fat dry milk
PCR	polymerase chain reaction
PE	Pro-Glu
PEG	polyethylene glycol
pfu	plaque-forming unit
PGRS	Polymorphic GC-rich repetitive sequences
POC	point-of-care
PPD	purified protein derivative
PPE	Pro-Pro-Glu
RIF	rifampicin
RNA	ribonucleic acid
rpm	revolutions per minute
SDS	sodium dodecyl sulphate
SDS-PAGE	Sodium dodecyl sulphate – polyacrylamide gel electrophoresis
Sec	Secretion
SRP	signal recognition particle
ssDNA	single stranded DNA
Tat	Twin-arginine translocation
TB	tuberculosis
TBE	Tris-borate-EDTA
TDR	totally drug resistant
TE	Tris-EDTA
TepiTool	T-cell Epitope Prediction Tools
TST	TB skin test
U	units
UniProt	Universal Protein Resource database
viz.	namely
w/v	weight/volume

WEGO	Web Gene Ontology Annotation Plotting
WHO	World Health Organization
XDR	extensively drugresistant
XDR-TB	extensively- drug resistant TB
xg	times gravity
µg	microgram
µL	microlitre
µM	micromolar
%	percent
°C	degrees celcius

CHAPTER 1: Introduction and literature review

1.1 INTRODUCTION

Mycobacterium tuberculosis is the causative agent of tuberculosis (TB), one of the leading causes of death from infectious disease worldwide. In 2016, the World Health Organization (WHO) recorded 10.4 million new TB incidents and 1.7 million deaths (including 0.4 million HIV/TB coinfections), globally. TB is the leading killer of people living with human immunodeficiency virus (HIV), responsible for 35% of deaths. The largest number of new TB cases were reported in the South-East Asia (45%) and African Regions (25%) accounting for 70% of global new cases in 2016 (Fig. 1.1). India, Indonesia, China, Philippines, and Pakistan are the leading countries contributing 56% of new TB cases globally. South Africa and Nigeria together account for 4% of the total new TB cases and are ranked 6th and 7th highest burdened countries respectively, in the world (WHO, 2017).

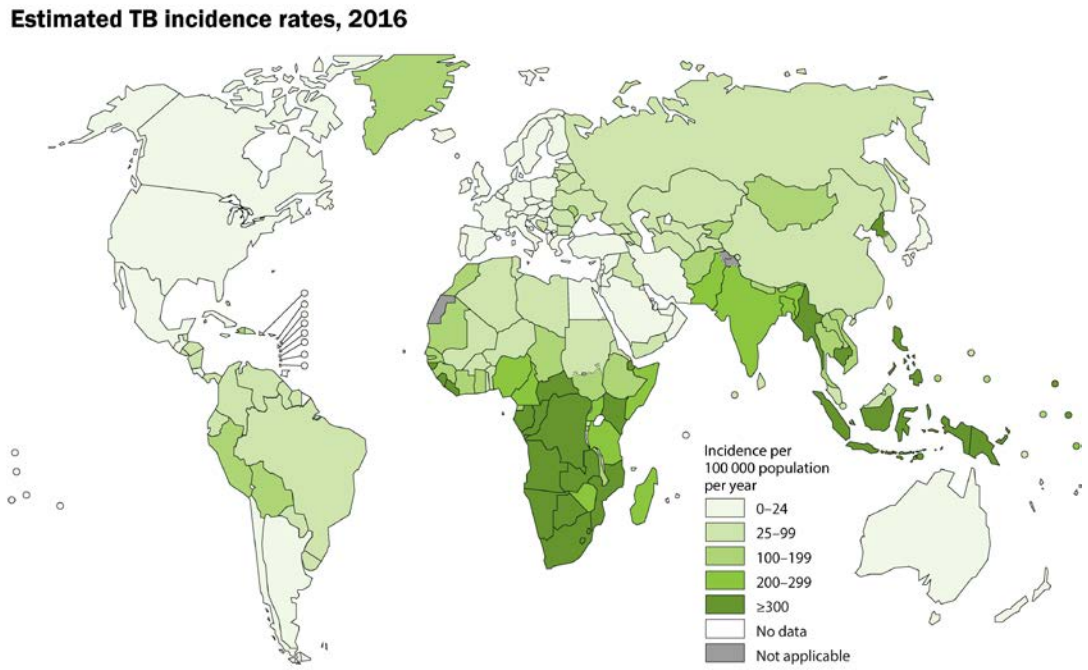

Fig. 1. 1: Global view of the estimated number of TB incidence rates in 2016 (WHO, 2017).

In 2016, South Africa recorded a total of 244053 new and relapse TB cases with 59% of these patients co-infected with HIV. TB is the leading underlying cause of mortality in South Africa, with 33137 deaths in 2015. Most TB deaths were recorded in KwaZulu-Natal, Eastern Cape and Gauteng provinces, respectively (Fig. 1.2). South Africa achieved a treatment success rate of 78% for new and relapse TB cases registered in 2014 and recorded an estimate of at least 3.5% of TB cases with MDR/TB (WHO, 2017).

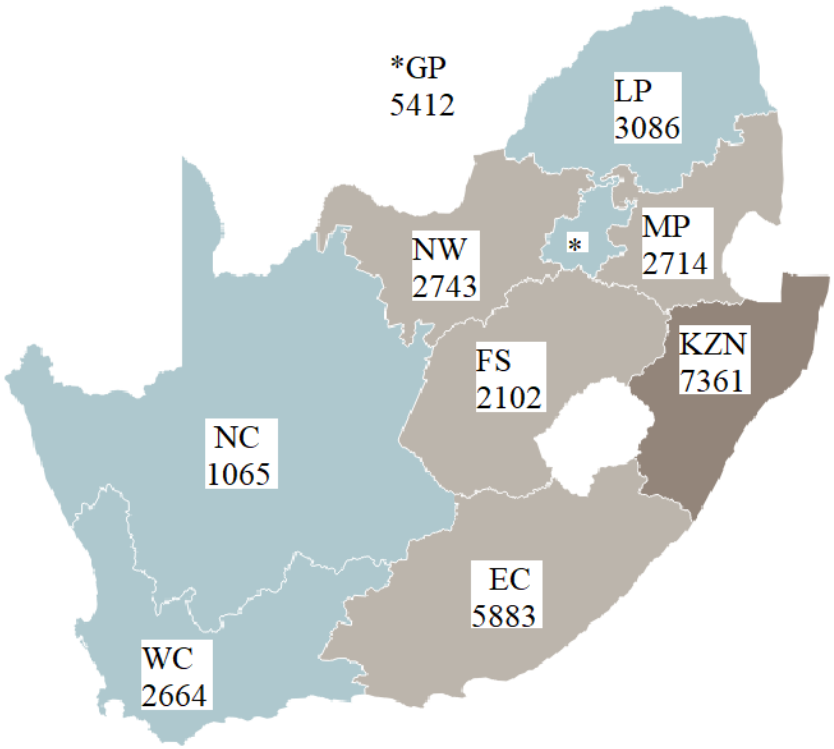

Fig. 1. 2: Estimated number of TB deaths recorded in 2015 for 9 provinces of South Africa which are EC- Eastern Cape, FS- Free State, GP- Gauteng, KZN- KwaZulu-Natal, LP- Limpopo, MP- Mpumalanga, NC- Northern Cape, NW- North West, and WC- Western Cape. The data was retrieved from Statistics South Africa cause of deaths report for 2015.

The reasons for the high global TB burdens are multifactorial: High prevalence of HIV co-infections especially in under resourced countries, the evolution of drug resistant strains, lack

of rapid point-of-care diagnostics tests, and timeous initiation of anti-TB therapy and ineffective drugs and vaccines.

The rapid and accurate diagnosis of TB is critical for treatment and management of patients and for reducing disease transmission. In 2011, WHO suspended the use of serological tests based on antibody detection for TB diagnosis due to the diverse antibody response of patients, and the lack of a single robust antigen resulting in significant inconsistencies in their performance (Bekmurzayeva *et al.,* 2013). The diagnostic tests for TB disease are sputum smear microscopy, culture-based methods and molecular methods. However, the smear microscopy requires a large number of bacilli (~$1x10^4$ bacterial cells) to be present in order for the result to be detected as positive. The culture-based method remains the reference standard, but its high sensitivity and specificity is offset by delays in the laboratory turnaround time of up to 12 weeks. The only molecular method for detection of *Mtb* and diagnosis of TB is GeneXpert *MTB*/RIF assay that provide results within 2 hours (WHO, 2012).

Drugs for TB treatment have been available since the 1940s. The first-line anti-TB drugs; isoniazid, rifampicin, ethambutol and pyrazinamide are used for treatment of drug-susceptible TB with 85% success rates (WHO, 2017). The poor patient compliance to treatment guidelines has led to the emergence of *Mtb* strains that are increasingly resistant to the available anti-TB drugs leading to the development of mono-drug resistant to multidrug resistant (MDR), extensively drug resistant (XDR), and eventually totally drug resistant (TDR) strains (Pillay *et al.*, 2007, Nguyen, 2016). In 2016, 82% of 600000 new global cases with resistance to rifampicin were MDR-TB. About 30000 MDR-TB cases were XDR-TB. South Africa recorded 19073 MDR-TB including 967 XDR-TB (WHO, 2017). Three provinces; KwaZulu-Natal, Western Cape, and Eastern Cape reported the most cases of MDR-TB and XDR-TB in South Africa (NHLS, 2009).

MDR strains are resistant to the two most powerful anti-TB drugs, isoniazid and rifampicin. MDR-TB is treatable with the second-line anti-TB drugs, however, these drugs are limited, expensive and toxic with a treatment period of up to 2 years. XDR strains are resistant to isoniazid and rifampicin, any fluoroquinolone, and at least one of capreomycin, kanamycin, and amikacin, the second-line anti-TB drugs (Caminero *et al.*, 2010). TDR strains are MDR strains resistant to all second-line anti-TB drug classes (Velayati *et al.*, 2009). Infections with XDR- and TDR-TB are incurable and are generally associated with high mortality rates (Caminero *et al.*, 2010), posing a serious threat to human kind, globally.

In 1890, Robert Koch discovered tuberculin which was first introduced as a tuberculin skin (TST) for TB screening by Von Pirquet in 1909 (Von Pirquet, 1909). Tuberculin was developed into purified protein derivative (PPD) in the 1930s by Florence Seibert (Daniel, 2006) and is now widely used as a diagnostic tool for detection of latent-TB and/or previous TB disease (WHO, 2017). In 1908, Albert Calmette and Camille Guerin developed the only available TB vaccine, Bacillus-Calmette Guerin (BCG) vaccine (Sakula, 1983). BCG vaccine comprises an attenuated strain of *Mycobacterium bovis*, the causative agent of TB in cattle (Sakula, 1983). It is administered to over 90% of new-borns and provide immunity against *M. tuberculosis* infection in new-borns and infants, globally.

There is an urgent need for new *Mtb* derived biomarkers for the design of new diagnostic tools, vaccines and drugs to tackle the current epidemic of drug-resistant TB. This study focused on the identification of unique *Mtb* extracellular proteins involved in host-pathogen interaction with potential as TB biomarkers.

1.2 LITERATURE REVIEW

1.2.1 *Mycobacterium tuberculosis*

M. tuberculosis (*Mtb*), an obligate pathogenic bacterium was confirmed as the causative agent of TB in 1882 by Robert Koch (Koch, 1882). *Mtb* is a member of closely related bacterial strains, referred to as the *Mtb* complex (*MTB*C). *MTB*C are acid-fast bacteria transmitted by inhalation of infectious aerosols generated by patients with pulmonary TB. In the lungs, it is phagocytosed by macrophages, where it spends the majority of it life cycle (Galagan, 2014). *MTB*C consist of *Mtb*, *M. africanum*, *M. bovis*, *M. canettii*, *M. microti*, *M. caprae*, *M. pinnipedii* and *M. orygis* which differ significantly in their genetic repertoire (Brosch *et al.*, 2000). *MTB*C is classified into seven phylogenetic lineages associated with distinct geographical regions, globally (Fig. 1.3) (Galagan, 2014).

Due to genetic variation, different *Mtb* strains present different clinical phenotypes and epidemiological outcomes (Bifani *et al.*, 2002, Gagneux *et al.*, 2007). Since H37Rv strain was derived in a laboratory from H37 (Kubica *et al.*, 1972), it may not represent the actual virulence behavior of wild-type clinical strains.

In 2002, a comparison of the genomes of the clinical strain CDC1551 and laboratory strain H37Rv demonstrated the presence of 17 genes in the former strain that were absent in the latter. The identified differences between H37Rv and CDC1551 may play a part in the respective virulence of these strains (Fleischmann *et al.*, 2002). Okumura and colleagues reported genomic variation among H37Rv and XDR- F15/LAM4/KZN (KZN605), MDR-KZN1435 and KZN4207 (drug-susceptible) clinical strains, respectively (Okumura *et al.*, 2015). Large inversion regions ranging from 0.93 to 3.46Mbp were observed in the genomes of KZN605, KZN1435 and KZN4207. Of interest in the present study is the KZN605 that has inversion points between 932051 and 932052, and between 3479594 and 3459595. In 2005,

the KZN605 strain was responsible for the deadly MDR- and XDR-TB outbreak in Tugela Ferry, KwaZulu-Natal (Gandhi *et al.*, 2006, Pillay *et al.*, 2007).

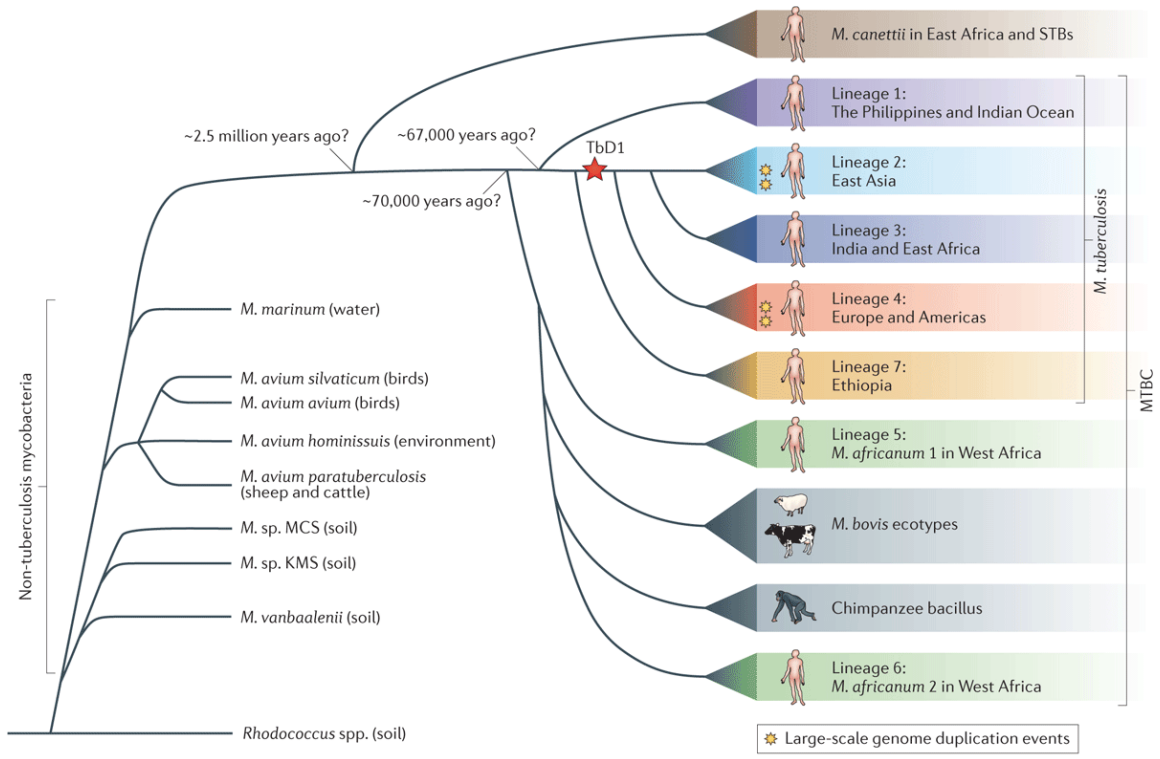

Fig. 1. 3: Evolutionary relationship between selected mycobacteria and members of the *MTB*C. TbD1 indicates the deletion event specific for *Mtb* lineages 2, 3 and 4. Colored branches indicate the seven human-adapted lineages, animal-adapted strains including *M. bovis* and the distantly related *M. canettii* (Galagan, 2014).

Pathosystems Resource Integration Centre (PATRIC) databases were used to sketch the proteome similarity profile between the laboratory H37Rv (ID: 83332.12) and KZN605 (ID: 1417005.3) genomes (Fig. 1.4). The genome lengths of H37Rv and KZN605 were shown to be 4411532 base pairs encoding 4367 genes and 4399718 base pairs that encode 4344 genes, respectively. The comparison of corresponding genes between the two genomes revealed similarity of 23% to 100%. Notably, 215 H37Rv genes were missing in KZN605 genome. Most

of the unaccounted genes encoded PE_PGRS and hypothetical proteins, respectively. It is known that the similarity between the many PE-PPE genes makes it difficult to place small internal sequences into the correct PE-PPE genes. Therefore, previous studies have shown that the PE-PPE genes, the most polymorphic in the chromosome, were often excluded from WGS analyses (Cole, 1998, Cole *et al.*, 1998, McEvoy *et al.*, 2012), which partly explains some of the missing 215 genes in KZN605. Nonetheless, differences in pathogenicity between laboratory and clinical strains and among clinical strains themselves could be due to variation in gene expression profile (Devasundaram *et al.*, 2016, Peters *et al.*, 2016).

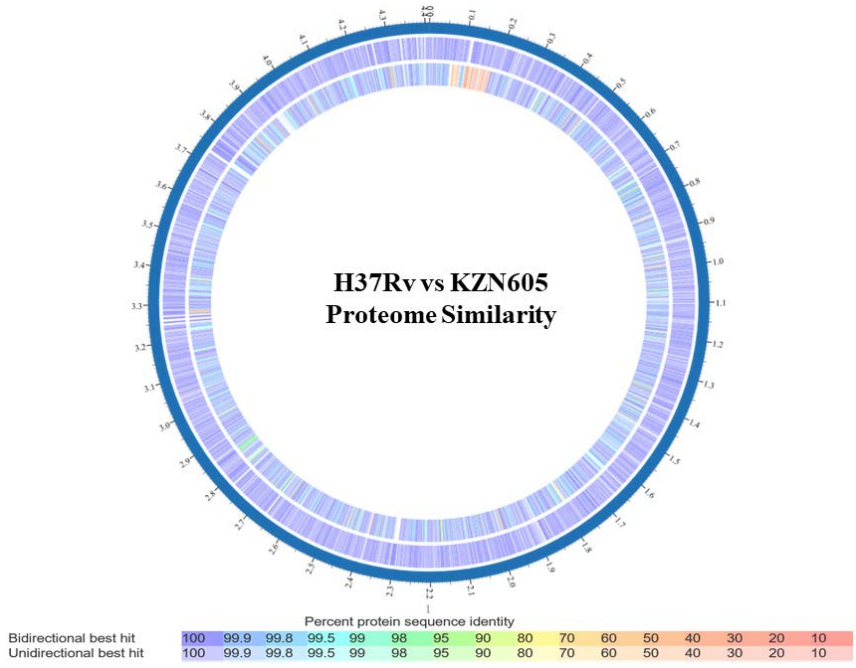

Fig. 1. 4: Circular map for comparison of *Mtb* H37Rv (83332.12) and *Mtb* KZN605 (1417005.3) proteome similarity. The H37Rv is depicted on the outside whilst KZN605 is shown on the outside. The figure was generated using the PATRIC.

1.2.2 *M. tuberculosis* H37Rv Genome Sequence

Mtb H37Rv was first isolated in 1905 and is widely used as the reference strain for phylogenetic and epidemiological studies in comparison to clinical isolates (Kubica *et al.*, 1972). In 1998, Cole and colleagues published the whole genome sequence of *Mtb* H37Rv strain (Cole *et al.*, 1998). The genome was re-annotated in 2002 (Camus *et al.*, 2002) and comprised a 4411529 base pair circular chromosome with high guanine and cytosine (GC) content of 65.6% (Fig. 1.5).

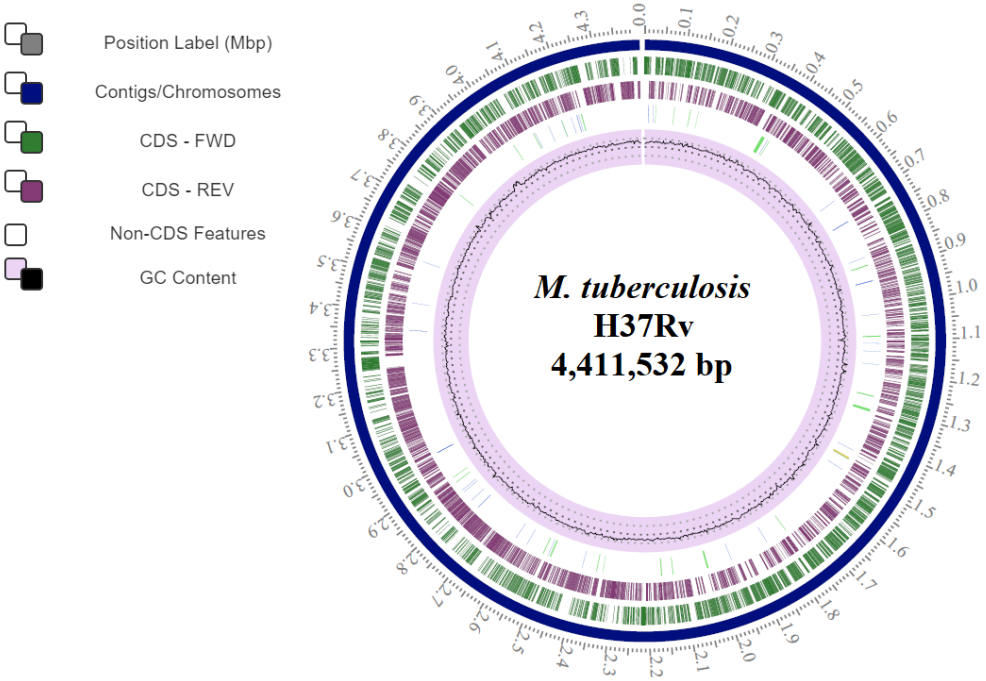

Fig. 1. 5: Circular map of the chromosome of *Mtb* H37Rv first published in 1998. The outer circles (grey and blue) shows the scale in Mbp, with 0.0 representing the origin of replication. The third ring from the exterior denotes the coding sequence by forward strand (clockwise, dark green) and the forth ring inwards shows the coding sequence by reverse strand (anticlockwise, purple); the fifty ring depicts non-coding sequence features (light green). The histogram (centre, pink with black line) represents GC content. The figure was generated using the PATRIC.

H37Rv contains a repertoire of 3924 protein-encoding genes. Since the *Mtb* H37Rv genome sequence became available, many studies have been conducted *in silico* to identify and characterize the encoded open reading frames (proteins). As a result, this genome is well annotated and curated and used as a reference by TB researchers.

1.2.3 *M. tuberculosis* H37Rv Annotation

There are several TB databases based on *Mtb* H37Rv genome sequence that are publicly available for use by TB researchers. The *Mtb* genome was annotated using different bioinformatics tools to generate information reported by TB databases. Of interest to the present study are two widely used and well curated databases, the TubercuList and PATRIC databases.

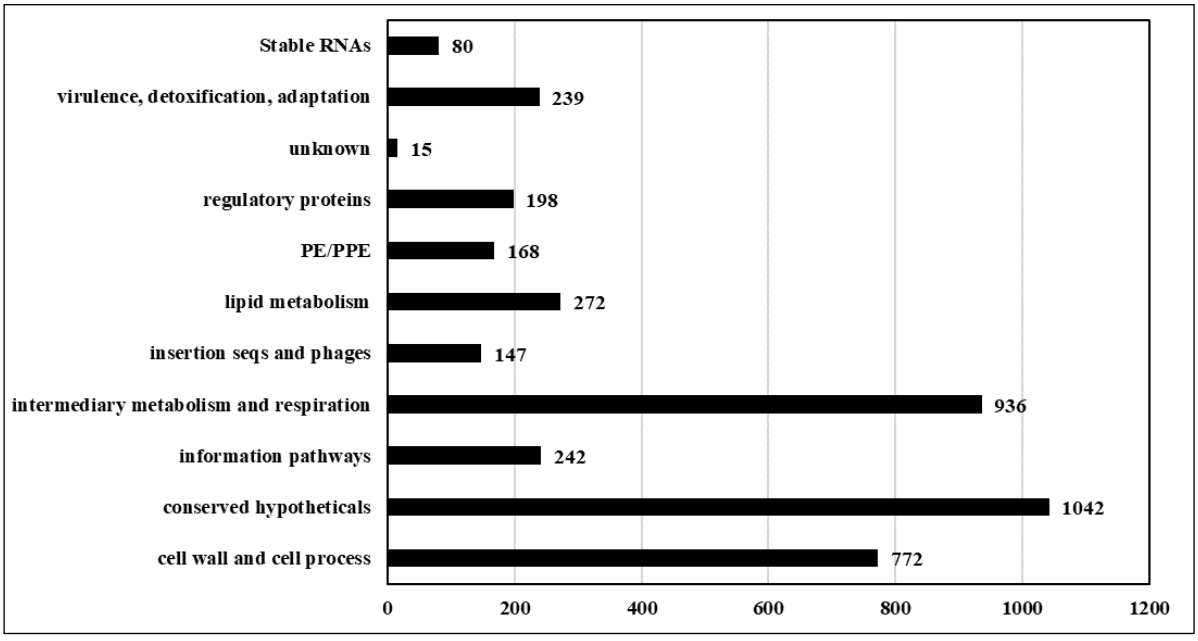

Fig. 1. 6: Distribution of *Mtb* H37Rv genes across eleven functional categories. The X and Y axes represent the number of genes and gene categories, respectively.

The TubercuList database website was created, currently annotated and maintained by Cole's group since 1998 (Lew *et al.*, 2011). TubercuList (www.tuberculist.epfl.ch) is based on *Mtb* H37Rv genome: its integrated genome details, protein information, operon annotation and bibliography feature among information available. To date, 4111 *Mtb* H37Rv genes are classified into eleven functional categories (Fig. 1.6), following continuous annotation and update of TubercuList database (Lew *et al.*, 2011, Lew *et al.*, 2013).

The majority of annotated H37Rv proteins belong to the *conserved hypotheticals* (1042) functional category which are unknown until they are fully characterized and functions are allocated to them. Recently, more proteins from this category were assigned to the *lipid metabolism* (272) functional category. The *intermediary metabolism and respiration*, and, *cell wall and cell process* functional categories were allocated 936 and 772 proteins respectively. The number of proteins associated with *virulence, detoxification, and adaptation* functional category were 239, *regulatory proteins*, 198, and 168 belonged to *PE*/*PPE* category. The *unknown* functional category was least represented with 15 proteins.

PATRIC (www.patricbrc.org) combines genome-scale data, metadata, and analysis tools for bacterial pathogens including *Mtb*. To date, over 9300 *Mtb* genome sequences (as of January 2018) including clinical isolates and reference strain (H37Rv) have been reported. Analysis of *Mtb* H37Rv PATRIC metabolic pathways revealed 14 encoded metabolic pathway classes (Fig. 1.7). Each pathway class comprised different metabolism and/or biosynthesis pathways (Appendix A, Table S1.1). For instance, the amino acid metabolism class have 13 different amino acid pathways such as lysine degradation, arginine and proline metabolism, histidine metabolism, tryptophan metabolism, and valine, leucine and isoleucine biosynthesis, etc. Worth noting, is the immune system class with one pathway, the T-cell receptor-signaling pathway involving only the phosphoserine/threonine phosphatase, PstP (Rv0018c). PstP is the only *Mtb* protein phosphatase reported to dephosphorylate at least five protein kinases (PknA,

PknB, PknD, PknE and PknF) and the penicillin-binding protein PBPA (Chopra *et al.*, 2003, Durán *et al.*, 2005, Dasgupta *et al.*, 2006). In total, there are 138 metabolic pathways encoded by the *Mtb* genome, with most of these required for growth and survival during *in vivo* growth.

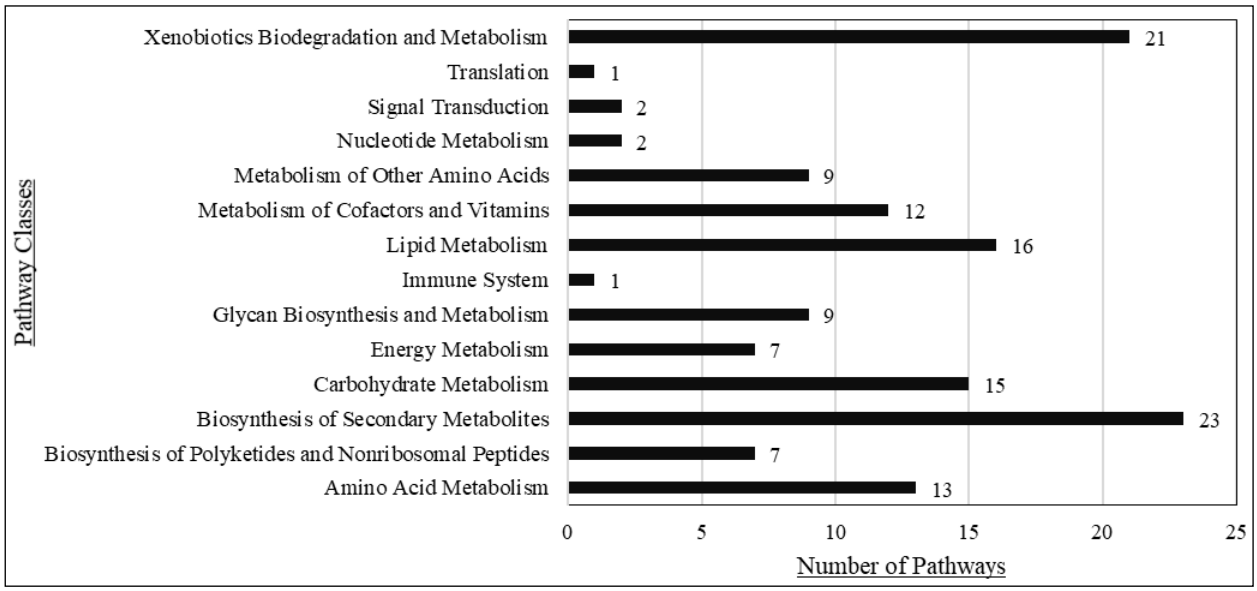

Fig. 1. 7: *Mtb* genome encoded 138 metabolic pathways classified into 14 classes according to PATRIC database analysis.

1.2.4 *M. tuberculosis* Protein Secretion Pathways

Mtb protein secretion pathways play an essential role in translocation of extracellular proteins and insertion of proteins across and into the cell membrane, respectively. The extracellular proteins include virulence factors that facilitate the adhesion and invasion of host cells, and some are essential for acquisition and absorption of nutritional elements into mycobacteria (McCann *et al.*, 2011). Therefore, protein secretion pathways are essential for growth and determine mycobacterial pathogenicity.

1.2.5 The Classical Protein Secretion Systems

Generally, *Mtb* uses two main pathways for protein secretion/export via the cytoplasmic membranes, namely, the Secretion (Sec) and the Twin-arginine translocation (Tat) pathways

(Braunstein *et al.*, 2001, Feltcher *et al.*, 2010). Successful protein secretion using both pathways is dependent on amino-terminal (N-terminal) extensions called signal peptides that direct proteins to the export machineries (Feltcher *et al.*, 2010). The Sec-pathway facilitates the transmembrane translocation of unfolded proteins that only fold into native conformation upon secretion. It also facilitates the insertion of membrane proteins into the cytoplasmic membrane (Driessen *et al.*, 2001, Natale *et al.*, 2008). In contrast, the Tat-pathway is responsible for translocation of folded proteins, sometimes with co-factors, across the membrane (Berks *et al.*, 2005).

1.2.5.1 *M. tuberculosis* Sec pathway system

The Sec pathway system is responsible for the secretion of most extracellular proteins that are involved in metabolism, transporting essential molecules, cell envelope structure, and those that participate in sensing and cell communication (Lee *et al.*, 2001). The Sec pathway is made up of cytosolic and membrane proteins that together enable protein translocation. The Sec system (Sec translocase) consists of three membrane proteins, SecY, SecE and SecG that assemble to form a complex known as the SecYEG (Natale *et al.*, 2008). The SecYEG complex forms a passage through which the synthesized proteins in the cytoplasm are transported to the extracellular environment (Feltcher *et al.*, 2010). In addition to SecYEG, an ATP-dependent motor protein, SecA, is responsible for driving translocation of secretory protein across the membrane (Tsirigotaki *et al.*, 2017). Two SecA homologs (Fig. 1.8), SecA1 and SecA2, are present in mycobacteria (Braunstein *et al.*, 2001). SecA1 protein is essential for general protein secretion, and SecA2 is required for secretion of specific proteins that are associated with *Mtb* virulence (Braunstein *et al.*, 2001, Braunstein *et al.*, 2003, Hou *et al.*, 2008, Feltcher *et al.*, 2013, Majlessi *et al.*, 2015). Interestingly, the SecA2 translocation substrates include proteins with and without N-terminal signal sequences (Braunstein *et al.*, 2003).

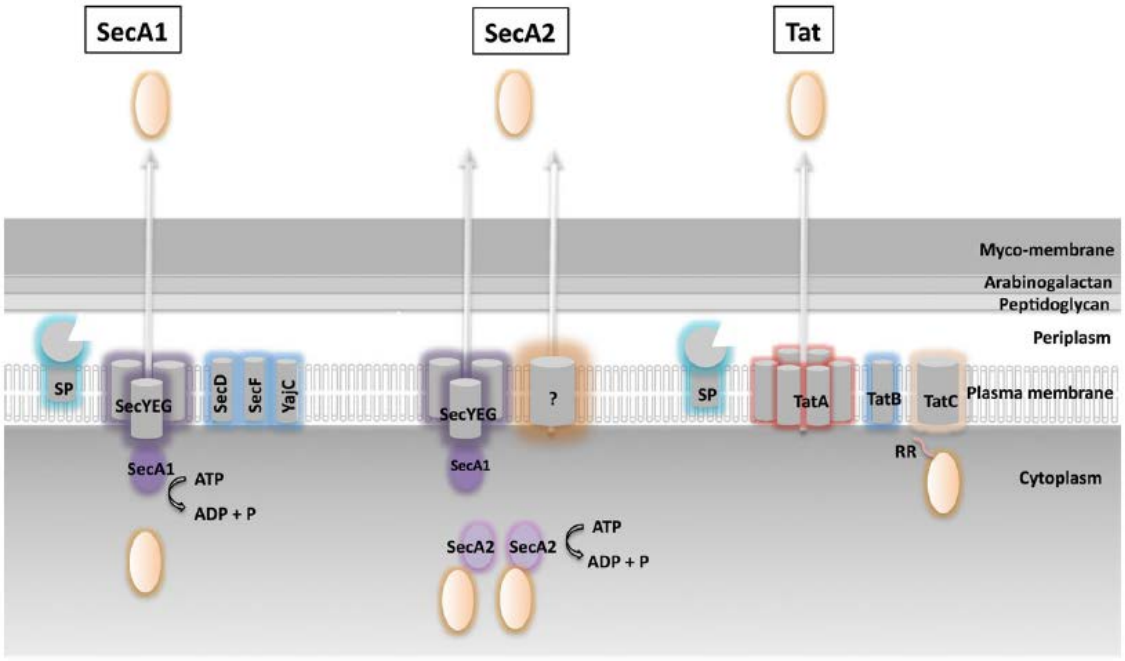

Fig. 1. 8: Schematic representation of general SecA1, alternative SecA2, and Tat protein export/secretion systems in *Mtb* (Majlessi *et al.*, 2015).

In general, bacteria use two different mechanisms to direct secretory proteins to the Sec translocase, viz. co-translational and post-translational targeting. Post-translational targeting involves specific labelling of secretory protein with an N-terminal signal peptide sequence that directs the unfolded protein to the Sec translocase after it is released from the ribosome once synthesis is complete (Natale *et al.*, 2008). In co-translational targeting, during translation, the pre-protein N-terminal signal sequence is recognized while emerging from the ribosome during protein synthesis. This results in targeting of the pre-protein complex including the ribosome, to the Sec translocase for processing. The co-translational targeting of pre-proteins facilitates the integration of integral membrane proteins into the cytoplasmic membrane (Natale *et al.*, 2008). The insertion of proteins into the plasma membrane is accomplished solely by SecYEG or by SecYEG in combination with the membrane protein insertase YidC (Dalbey *et al.*, 2014).

1.2.5.2 *M. tuberculosis* Tat pathway system

The second major protein secretion system, the Twin arginine translocation (Tat) pathway is responsible for translocation of pre-folded proteins across membranes and is less understood than the Sec pathway (Natale *et al.*, 2008, Ligon *et al.*, 2012). The Tat translocase consists of three membrane integrated proteins (Fig. 1.8); TatA (Rv2094c), TatB (Rv1224) and TatC (Rv2093c), that together form a channel called TatABC translocase complex (McDonough *et al.*, 2008, Natale *et al.*, 2008). Another protein, TatD (Rv1008) is suspected of playing some role in the Tat secretion system (Majlessi *et al.*, 2015). Most Tat substrates are translocated as substrate-cofactor complexes, thus, catering for aerobic respiration and a multitude of redox pathways for anaerobic respiration, most of which rely on the Tat system (Natale *et al.*, 2008).

1.2.5.3 Signal peptides, recognition and cleavage

Protein secretion via the Sec and Tat pathways relies on the N-terminal extension of the secretory protein substrate called signal peptide sequence (SPS). SPS is required to sort and target the protein to the correct pathway and on average, it is 16 to 30 amino acids residues in length. However, some are more than 50 amino acid residues long. Therefore, SPSs differ in sequence and length, with extensions on the n-region or hydrophobic region (Kapp *et al.*, 2000). The Sec and Tat SPS is characterized by a tripartite structure of a positively charged amino-terminal (n-region), a hydrophobic core (h-region) and a polar carboxyl terminal. In the Sec system, the charged n- and h-regions facilitate the translocation by interacting with SecA and signal recognition particle (SRP). SRP recognizes the hydrophobic signal sequence during emergence from the ribosome and thereafter, is targeted to the Sec-translocase (Natale *et al.*, 2008).

The Tat SPS possesses a highly conserved twin-arginine motif located at the junction of the n- and the h- regions (Posey *et al.,* 2006). The twin-arginines are found 2-30 residues behind the N-terminus and the SPSs are longer, with a larger and less hydrophobic h-region of 13–20 uncharged residues, thus, slightly longer than that of Sec-substrates (Goosens *et al.,* 2014). In order to avoid recognition by the Sec pathway, the Tat c-region regularly contains a positively charged amino acid residue (Bogsch *et al.,* 1997).

Successful translocation and release of Sec and Tat substrates depends on the recognition and cleavage of SPS by signal peptidase (Natale *et al.,* 2008). Signal peptidases are membrane-anchored enzymes responsible for processing proteins that are translocated across the membranes, cleaved and secreted/released into the surrounding environment or remain displayed or anchored on the bacterial surface. The site of cleavage is recognized by the amino acids with short side chains at the -1 and no charged amino acid residues at the -3 position (Ting *et al.,* 2016).

1.2.6 The Non-Classical Type VII Secretion Pathway

Mtb also possesses a non-classical, specialized type VII secretion (T7S) pathway called ESAT-6-like (ESX) pathway, which is required for *Mtb* virulence. This pathway was named after the first known secreted substrate of the ESX pathway, the 6kDa early secreted antigenic target (ESAT-6). The ESX pathway substrates lack Sec or Tat signal peptides and rely on ESX systems for secretion. The five existing ESX systems, ESX-1, ESX-2, ESX-3, ESX-4 and ESX-5 are encoded at different positions within *Mtb* genome (Fig. 1.9). Each *esx* locus contain a pair of *esx* genes that is flanked by genes coding for the secretion apparatus (Majlessi *et al.,* 2015). In addition, each *esx* locus encodes all molecules required for the assembly of a completely functioning secretory apparatus, including the regulatory proteins, membrane

transporters, substrates for secretion by the same system and the powering adenosine triphosphatase (ATPase) (Gray *et al.*, 2016).

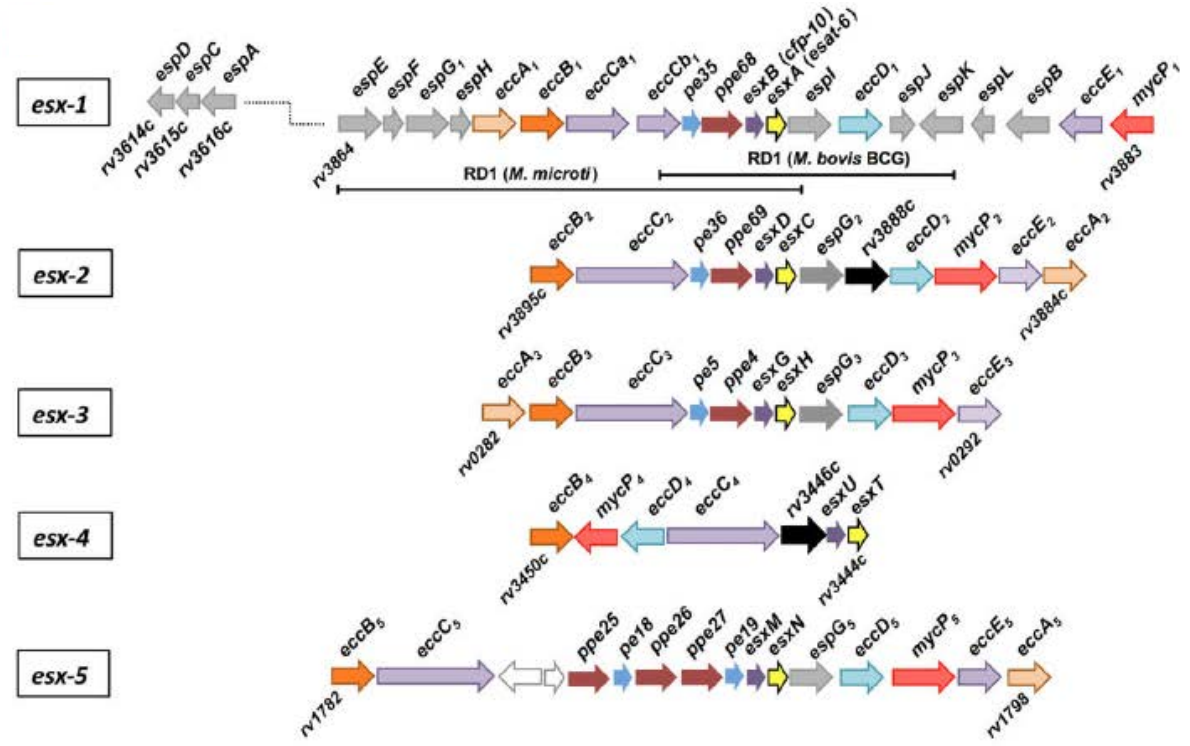

Fig. 1. 9: Schematic representation of genetic organization and gene names of the 5 ESX systems of the *Mtb*-specific type VII secretion systems (Majlessi *et al.*, 2015).

The conserved genes called the ESX-conserved component (*ecc*), the *eccB*, *eccC* and *eccD* are present in each locus. The ESX-conserved components consist of the membrane protein EccB, a protein with an ATPase domain providing energy for protein transport; EccC; and EccD, which contain multiple transmembrane domains which act as an inner membrane export channel (van Pittius *et al.*, 2001). Assembled together, the EccB, EccC and EccD probably form the core membrane channel for translocation of ESX substrates during which MycP perform some uncertain cleavage in the process (Houben *et al.*, 2014). The *mycP* gene is conserved in all *esx* locus and encodes mycosin (MycP), a membrane-anchored protein with protease activity for processing some of the ESX substrates (van Pittius *et al.*, 2001).

The other ESX pathway substrates, the *pe* and *ppe* genes, are found organized as an operon in all *esx* locus except ESX-4 system. The majority of *pe/ppe* genes are secreted by the ESX-5 system, the most active secretion system. The *pe* and *ppe* genes are also present in other positions in the *Mtb* genome (Houben *et al.*, 2014). The other common ESX system components are Esps, for ESX-specific proteins which are secreted substrates and may also be located in a separate operon such as *espACD* while others are located within the *esx* locus (Majlessi *et al.*, 2015).

The well-studied ESX-1 system encodes two major substrates, ESAT-6 (EsxA) and culture filtrate protein 10 (CFP- 10) (EsxB) that are associated with virulence and are targets of the immune response in infected individuals (McLaughlin *et al.*, 2007). The ESX-1 system core components genes, *eccA, eccB, eccCa, eccCb, eccD, eccE* and *mycP*, are required for EsxA and EsxB secretion (Feltcher *et al.*, 2010, Stoop *et al.*, 2012). EccA and EccCb are cytoplasmic ATPases, supplying energy for the secretion process and are involved in targeting proteins for ESX-1 secretion (Feltcher *et al.*, 2010). EccCa1 is an integral membrane protein that interacts with the ATPase. EccB1, EccD1 and EccE1 are transmembrane proteins that ensure smooth translocation of ESX-1 substrates across the cytoplasmic membrane. EccD acts as a channel for protein translocation while EccE1 interacts with EspD that is required for ESAT-6/CFP-10 secretion but its function remains unclear (MacGurn *et al.*, 2005). In the ESX-1 system, MycP1 is central to the functioning of the system as evident by its deletion, which completely shut-downs the ESX-1 secretion (Stoop *et al.*, 2012). The core ESX-1 components are associated with virulence, with some required for intracellular growth (EccA, EccCa, EccCb, EsxB, EsxA, EccD and mycP) and others involved in active evasion of host immune response by immunomodulation (EccCa, EccCb, EsxA and EccD) and arresting phagosomal maturation (EccCa, EccCb and EccD) as reviewed in Stoop et al. (2012).

1.2.7 *M. tuberculosis* Secretome

Mtb secreted proteins (secretome) consists of the membrane exported proteins that can be cleaved and released into the surrounding environment to perform diverse extracellular functions. Some of the secreted proteins possess a directive signal for their insertion into the outer membrane or are membrane-anchored proteins found on the cell surface. *Mtb* secretes a large number of proteins that are involved in nutrient acquisition and counter-action of targeting molecules. The secreted proteins include extracellular released proteins, transmembrane and surface membrane proteins (Ivankov *et al.*, 2013). They also include enzymes that are involved in catalysis of chemical reactions associated with bacterial metabolism, catabolism and biosynthesis. Enzymes are either released extracellularly to act on substrate for efficient uptake or could be membrane associated to facilitate transportation of molecules (Maffei *et al.*, 2017). Some of the secreted proteins are associated with virulence as they may be involved in host cell invasion, acquisition of nutritional molecules by breakdown of host cells and modulation of host immune response in order to avoid detection. Therefore, the ability of *M. tuberculosis* to cause disease entirely depends on the virulence properties of the mycobacterial secretome (Majlessi *et al.*, 2015).

Using bioinformatics tools, Rashid et al., (2007) allocated 3918 *Mtb* open reading frames into their 4 subcellular localization compartments: the cytoplasmic, secreted/excreted, integral membrane and lipid-anchored surface membrane proteins (Rashid *et al.*, 2007) (Fig. 1.10). The secreted, membrane anchored and some integral membrane proteins are involved in host-pathogen interactions and facilitate the uptake of nutrient molecules, and therefore, represent virulence factor proteins (Niederweis *et al.*, 2010). *M. tuberculosis* uses sophisticated export and secretion pathways to process most secretome proteins that enable survival and persistence in different environments and hosts (Majlessi *et al.*, 2015).

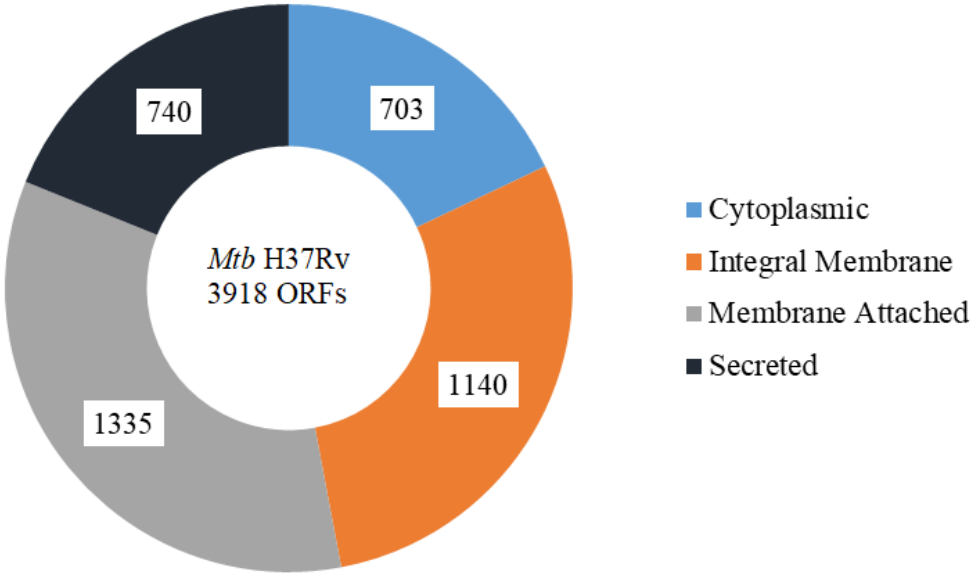

Fig. 1. 10: Subcellular localization of 3918 *Mtb* H37Rv proteins into cytoplasmic, integral membrane, membrane attached and secreted proteins.

The *Mtb* secretome plays an important role in the host-pathogen interaction and can serve as a reservoir for candidates or targets for the development of new drugs and vaccines, and diagnostics (Harth *et al.*, 1999b). The present study used the phage display technique to select and identify novel *Mtb* secreted biomarkers that can potentially be used in TB diagnosis and in the development of a new protective TB vaccine.

1.2.8 Phage Display Technology

The filamentous bacteriophages were first used as phage vectors for peptide-display by George Smith in 1985 (Smith, 1985). Filamentous phages, including M13, f1 and fd, are a group of viruses that infect gram-negative bacteria via F-pili**.** The filamentous phages do not lyse infected cells during their lifecycles. Their genomes comprise single-stranded DNA (ssDNA)

wrapped inside 2700 copies of the 50 amino acid long major coat protein pVIII (Cesareni, 1992, Rodi *et al.*, 1999).

The phage genome encodes 10 proteins (Fig. 1.11): five phage structural proteins, pIII, pVI, pVII, pVIII and pIX; two assembly and export proteins, pI and pIV; and three proteins for replication, pII, pV and pX (Wang *et al.*, 2004). There are 5 copies each of pIII and pVI at one end of the phage particle. The pIII is a 406 amino acid residue protein required for host infection and normal morphogenesis of the phage particle. Protein VI is a 113 amino acid protein required for attachment of pIII protein to the phage particle. On the other end, there are 5 copies each of pVII and pIX that are 32 and 33 amino acids in length, respectively. These proteins are responsible for initiation of assembly and for maintaining the stability of phage particles, respectively (Cesareni, 1992, Rodi *et al.*, 1999).

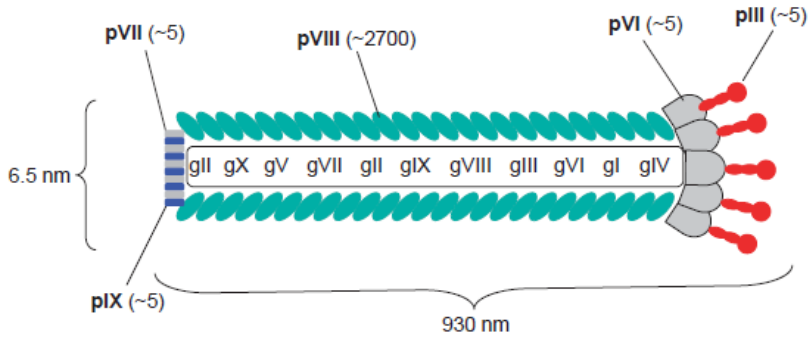

Fig. 1. 11: The main structural proteins of bacteriophages: pIII, pVI, pVII, pVIII and pIX encoded by ssDNA (Mullen *et al.*, 2006).

The phage life cycle (Fig. 1.12) begins with the binding of phage via pIII protein to the F-pilus of susceptible bacteria to initiate infection. The ssDNA genome is injected into the cell and the host polymerase is used to synthesize a complementary strand to make a double stranded phage DNA genome that is the replication form. This is followed by the transcription and translation of all 10 genes for production of proteins using host machinery. The synthesized proteins

include structural proteins (coat proteins pIII, pVI, pVII, pVIII and pIX), proteins for assembly and export (pI and pIV), and proteins for replication (pII, pV and pX) (Wang *et al.*, 2004, Mullen *et al.*, 2006).

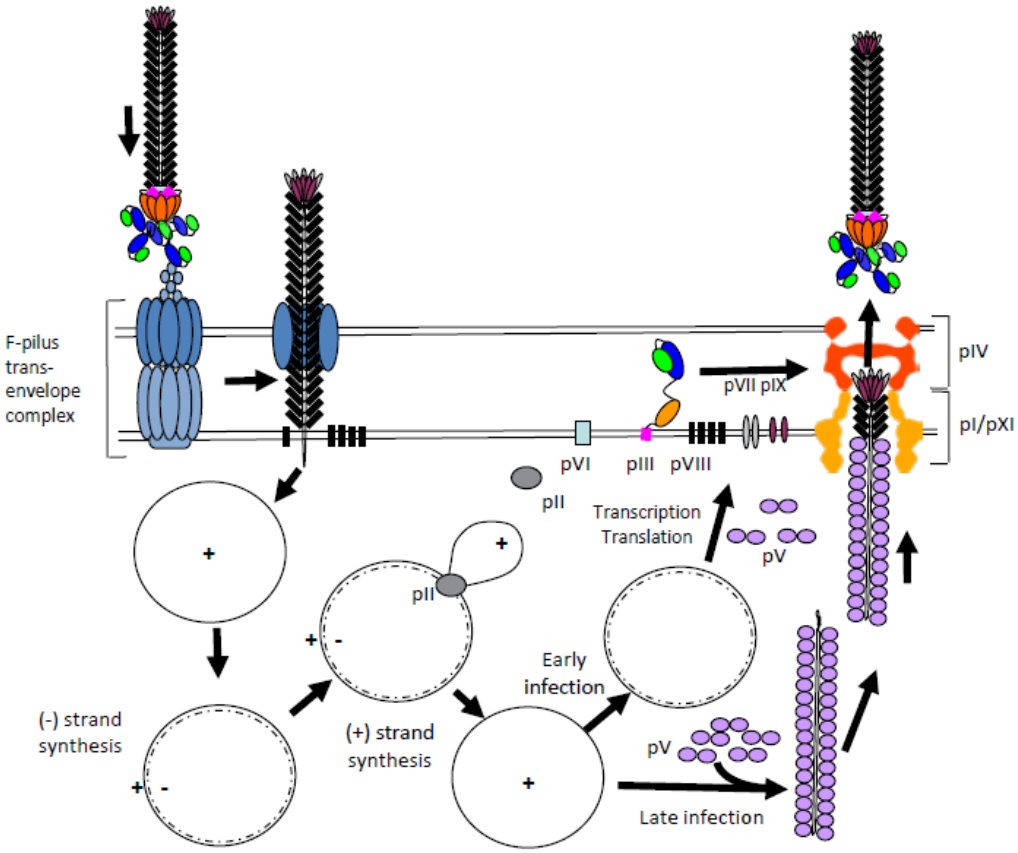

Fig. 1. 12: The life cycle begins when phage bind to the bacterial pilus and single-stranded viral genome is injected. Once inside, it is converted into double-stranded phage genome. Phage-encoded proteins are produced by host-mediated protein synthesis and virions are assembled and exported across the bacterial membranes (Rakonjac, 2012).

All capsid proteins are inserted in the bacterial outer membrane prior to phage assembly for final assembly of phage particles in the periplasmic environment. Therefore, the mature phage structure may not include any fusions that disturb the export process of the coat protein (Danner *et al.*, 2001). Unlike pVIII, pIII membrane insertion is Sec-pathway dependent, thus, it is more

likely that large fusions to pIII are exported into the periplasm more easily than large pVIII fusions (Thie *et al.*, 2008).

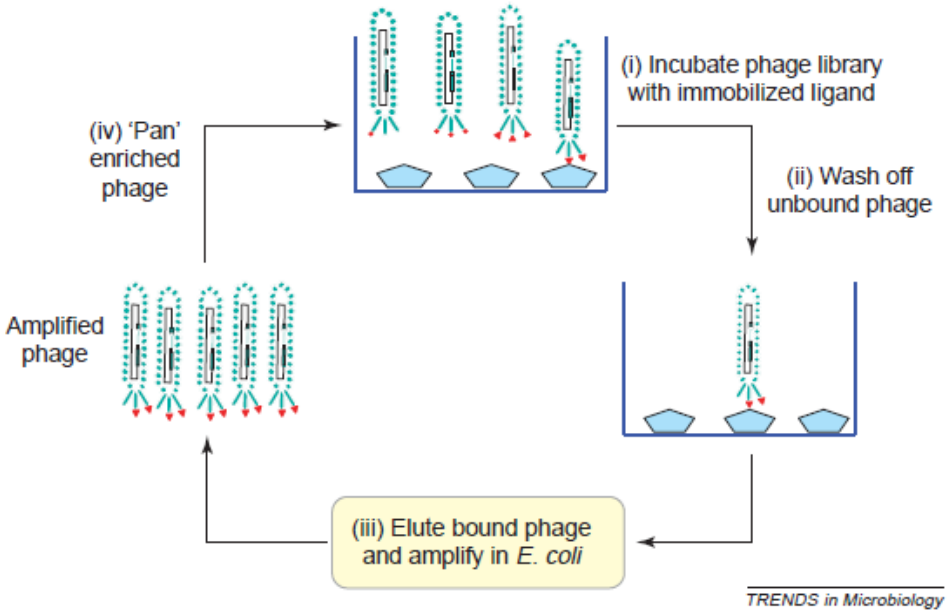

Fig. 1. 13: Affinity selection of filamentous phage display libraries to identify phage that display fusion proteins (pIII fusions) that bind to the immobilized target: (i) target molecule is immobilized and incubated with phage library, (ii) non-binding phages are washed away, (iii) bound phages are eluted and amplified in *E. coli*, and (iv) resulting in enrichment of high affinity binders (Mullen *et al.*, 2006).

The use of bacteriophages to display foreign proteins on their surface is termed phage display. Phage display technology is based on the fusion of a foreign gene into a bacteriophage coat protein gene, for co-expression and display on the surface of a phage particle to make peptide or protein libraries (Smith, 1985, McCafferty *et al.*, 1990). The foreign peptide sequences can be fused into one of several capsid proteins, namely the pIII (predominantly), pVI, pVII, pVIII or pIX (Hoogenboom *et al.*, 1998, Rodi *et al.*, 1999, Gao *et al.*, 2002). This technology allows the expression and surface display of polypeptides (phenotype) while physically linked to their coding DNA (genotype) (Mullen *et al.*, 2006). The phage display peptide library can be

screened for phage-displayed molecules that perform specific functions like desired binding specificities to a chosen immobilized target by a process called “biopanning” (Fig. 1.13). Biopanning involves surface immobilization of target molecules and incubation with the phage display library to allow binding of phages displaying cognate molecules. After the unbound phage particles are washed away, target bound particles are released/eluted at high or low pH and are amplified in *E. coli* cells. The resultant individual colonies are picked and DNA of the foreign inserted fragment in the phage genome is sequenced since the isolated phage particles bear the gene sequence of the displayed foreign molecules. Therefore, this method allows both selection and amplification of a phage particle containing the desired gene sequences (Azzazy *et al.*, 2002).

Phage display is a powerful *in vitro* selection technology that can be exploited to specifically extract proteins with novel, desired properties from large protein libraries. Liu et al., 2011, constructed a *Mtb* phage secretome library from H37Rv genomic DNA. Serum samples from 14 active-TB patients and healthy controls were used for immunoscreening of the library. The screening identified 47 proteins including immunogenic protein MPT64, secreted fibronectin-binding protein antigen 85B, PPE family proteins and three novel antigens (polyketide synthase associated protein papA2, lipoprotein LpqA and putative conserved protein CpsA) with potential as vaccine or diagnostic candidates (Liu *et al.*, 2011, Liu *et al.*, 2013).

Other pathogens studied using phage display to identify novel antigens recognized by serum antibodies from infected human or animals include *Streptococcus pneumoniae* (Beghetto *et al.*, 2006), *Taenia solium* (González *et al.*, 2010), *Mycoplasma mycoides* subsp. *mycoides* (Naseem *et al.*, 2010), *Mycobacterium avium* subsp. *paratuberculosis* (Nagata *et al.*, 2013), and *Neisseria gonorrhoea* (Connor *et al.*, 2016). The selection and identification of pathogen specific biomarkers have potential use in disease diagnosis, vaccine and drugs design.

1.2.8.1 Phage display vectors

Phage display libraries can be constructed using vectors based on the wild type filamentous bacteriophage or by using ‘phagemids’, the hybrids of phage intergenic regions and plasmid DNA vectors (O'Connell *et al.*, 2002). Phagemid vectors are preferred over phages due to higher transformation efficiencies and because of their monovalent display of the foreign peptides (Hoogenboom *et al.*, 1998). The phagemid vectors are designed to have an Ff phage and *E. coli* plasmid origins of replication (ori), *gIII* and/or *gVIII* for fusion of a foreign peptide into a coat protein of choice, and an antibiotic resistance gene. But, phagemids lack all other phage genes that are required to produce a complete phage. Therefore, a helper phage is used to supply all the structural proteins required for phage assembly during rescue and packaging of phagemid into phage particles. Notably, the wild-type pIII from the helper competes with the foreign peptide-pIII fusion for display on the phage surface. The rate of phagemid rescue differs for each member in the phage library due to the different effects exerted on the host *E. coli* by each exogenous gene and/or its gene product (Gupta *et al.*, 2013).

1.3 MOTIVATION

The emergence and spread of MDR/XDR strains of *Mtb* is a major health problem, a threat to humankind and complicate the fight against TB (WHO, 2017). Therefore, there is an urgent need for new *Mtb* derived biomarker targets for development of new drugs, vaccines and diagnostic tools. Hence, there is a need to study clinically relevant strains of *Mtb* that will provide more insight into wild-type virulent behavior (Palanisamy *et al.*, 2009, Devasundaram *et al.*, 2016). In the present study, XDR *Mtb* F15/LAM4/KZN strain was investigated for secretory proteins involved in host-pathogen interactions, immunodominant antigens, new biomarkers and new drug targets.

Mtb secretes a large number of proteins into its extracellular space and these play an important role in host-pathogen interactions and therefore, may be ideal candidates or targets for the development of new drugs and vaccines, and for use in diagnosis (Harth *et al.*, 1999a). Using a phage display approach, we investigated the suitability of phage display technology to select novel *Mtb* secreted proteins and identify biomarkers recognized by the humoral response in latent-TB and active-TB patients.

Targeting the expressed and secreted *Mtb* proteins, the phagemid vector system used lacks a leader sequence that is necessary for secretion and successful display of foreign peptide on the phage particle surface (Rosander *et al.*, 2002, Wall *et al.*, 2003, Jankovic *et al.*, 2007). Generally, phage protein pIII contains a signal peptide sequence and is widely used to successful display foreign peptide proteins. A signal sequence is essential for correct targeting of pIII to the inner membrane and incorporation into the virion (Gupta *et al.*, 2013). Therefore, only the *Mtb* protein fragments with signal peptide sequence for secretion or some membrane insertion directive will be successful displayed as fusion proteins on the phage particle surface.

1.3.1 Hypothesis

We hypothesised that novel and secretory *Mtb* protein biomarkers may be identified through the use of phage display technique and immunoscreening against serum samples from healthy individuals and TB patients.

1.3.2 Aim

To identify a set of *Mtb* secreted protein biomarkers with potential application in the development of a cost effective point-of-care TB diagnostic immunoassay and design of a new TB vaccine.

1.3.3 Objectives

- To culture *Mtb*, extract and purify genomic DNA and generate DNA fragments
- To clone DNA fragments into a phage display vector and use recombinant phage DNA to transfect a compatible *E. coli* strain
- To screen the phage library against the sera of healthy individuals, active- and latent- TB patients
- To sequence phage DNA and analyze the genetic composition of selected phages and identify novel disease state specific biomarkers
- To identify novel and useful parallel B-cell and T-cell epitopes

1.3.4 Study Design

The mycobacterial genomic DNA was extracted from *Mtb* culture and DNA fragments of between 150-to-1500 base pairs were generated, cloned into phage display vector and expressed in *E. coli*. If the *Mtb* DNA insert encodes the signal peptide sequence or transmembrane insertion motif, the expressed protein is displayed on the phage particle surface. The *Mtb* phage displayed secretome library was screened against the immobilized polyclonal sera from TB patients to select and purify phages bearing sequences with desired binding specificities from the nonbinding variants (Liu *et al.*, 2011). The selected phage particle genotype composition was analyzed, leading to the discovery of a useful and novel clinical set of biomarkers that can be used to diagnose TB and/or design a new TB vaccine.

1.3.5 Scope of Thesis

This thesis comprises of five chapters and is presented in a thesis by ‘manuscript” format. Chapter one includes the introduction and literature review. Chapter two, Identification of unique putative biomarkers from a *M. tuberculosis* F15/LAM4/KZN phage secretome library published in 2017 in the journal, Pathogens and Disease, describes the construction of the *Mtb* F15/LAM4/KZN phage secretome library and the identification of unique *Mtb* secreted proteins. In Chapter three, Immunoscreening of the *M. tuberculosis* F15/LAM4/KZN secretome library against TB patients’ sera identifies unique active- and latent-TB specific biomarkers (submitted in January, 2018 to the journal, Tuberculosis), *Mtb* F15/LAM4/KZN phage secretome library was screened against immobilized polyclonal sera from active- and latent-TB patients to identify unique active- and latent-TB specific biomarkers. Chapter four, B-cell epitope derived T-cell epitopes as vaccine candidates to confer antibody and cellular mediated immunity against *M. tuberculosis* infection (manuscript in preparation) is an *in silico*

analysis of selected proteins recognized by patients' antibodies to confirm their B-cell epitopes and also identify promiscuous B-cell epitope derived T-cell epitopes of *Mtb* proteins binding to MHC Class I and Class II molecules for the potential design of a recombinant polypeptide TB vaccine. Chapter five provides a synthesis of the 3 research chapters, the overall conclusions and recommendations for future research.

1.4 REFERENCES

Azzazy, H.M. and Highsmith, W.E. (2002). Phage display technology: clinical applications and recent innovations. *Clinical biochemistry* **35,** 425-445.

Beghetto, E., Gargano, N., Ricci, S., Garufi, G., Peppoloni, S., Montagnani, F., *et al.* (2006). Discovery of novel *Streptococcus pneumoniae* antigens by screening a whole-genome λ-display library. *FEMS microbiology letters* **262,** 14-21.

Bekmurzayeva, A., Sypabekova, M. and Kanayeva, D. (2013). Tuberculosis diagnosis using immunodominant, secreted antigens of *Mycobacterium tuberculosis*. *Tuberculosis* **93,** 381-288.

Berks, B.C., Palmer, T. and Sargent, F. (2005). Protein targeting by the bacterial twin-arginine translocation (Tat) pathway. *Current opinion in microbiology* **8,** 174-181.

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

Bogsch, E., Brink, S. and Robinson, C. (1997). Pathway specificity for a ΔpH-dependent precursor thylakoid lumen protein is governed by a 9sec-avoidance'motif in the transfer peptide and a 9sec-incompatible'mature protein. *The EMBO Journal* **16,** 3851-3859.

Braunstein, M., Brown, A.M., Kurtz, S. and Jacobs, W.R. (2001). Two nonredundant SecA homologues function in mycobacteria. *Journal of bacteriology* **183,** 6979-6990.

Braunstein, M., Espinosa, B.J., Chan, J., Belisle, J.T. and R Jacobs, W. (2003). SecA2 functions in the secretion of superoxide dismutase A and in the virulence of *Mycobacterium tuberculosis*. *Mol. Microbiol.* **48,** 453-464.

Brosch, R., Gordon, S.V., Pym, A., Eiglmeier, K., Garnier, T. and Cole, S.T. (2000). Comparative genomics of the mycobacteria. *International Journal of Medical Microbiology* **290,** 143-152.

Caminero, J.A., Sotgiu, G., Zumla, A. and Migliori, G.B. (2010). Best drug treatment for multidrug-resistant and extensively drug-resistant tuberculosis. *The Lancet Infectious Diseases* **10,** 621-629.

Camus, J., Pryor, M., Medigue, C. and Cole, S. (2002). Re-annotation of the genome sequence of *Mycobacterium tuberculosis* H37Rv. *Microbiology (Reading, England)* **148,** 2967 - 2973.

Cesareni, G. (1992). Peptide display on filamentous phage capsids An new powerful tool to study protein—ligand interaction. *FEBS letters* **307,** 66-70.

Chopra, P., Singh, B., Singh, R., Vohra, R., Koul, A., Meena, L.S., *et al.* (2003). Phosphoprotein phosphatase of *Mycobacterium tuberculosis* dephosphorylates serine–threonine kinases PknA and PknB. *Biochemical and Biophysical Research Communications* **311,** 112-120.

Cole, S.T. (1998). Comparative mycobacterial genomics. *Current Opinion in Microbiology* **1,** 567-571.

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

Connor, D.O., Zantow, J., Hust, M., Bier, F.F. and von Nickisch-Rosenegk, M. (2016). Identification of Novel Immunogenic Proteins of *Neisseria gonorrhoeae* by Phage Display. *PLoS ONE* **11,** e0148986.

Dalbey, R.E., Kuhn, A., Zhu, L. and Kiefer, D. (2014). The membrane insertase YidC. *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research* **1843,** 1489-1496.

Daniel, T.M. (2006). The history of tuberculosis. *Respiratory medicine* **100,** 1862-1870.

Danner, S. and Belasco, J.G. (2001). T7 phage display: a novel genetic selection system for cloning RNA-binding proteins from cDNA libraries. *Proceedings of the National Academy of Sciences* **98,** 12954-12959.

Dasgupta, A., Datta, P., Kundu, M. and Basu, J. (2006). The serine/threonine kinase PknB of *Mycobacterium tuberculosis* phosphorylates PBPA, a penicillin-binding protein required for cell division. *Microbiology* **152,** 493-504.

Devasundaram, S. and Raja, A. (2016). Variable transcriptional adaptation between the laboratory (H37Rv) and clinical strains (S7 and S10) of *Mycobacterium tuberculosis* under hypoxia. *Infection, Genetics and Evolution* **40,** 21-28.

Driessen, A.J., Manting, E.H. and van der Does, C. (2001). The structural basis of protein targeting and translocation in bacteria. *Nature Structural & Molecular Biology* **8,** 492.

Durán, R., Villarino, A., Bellinzoni, M., Wehenkel, A., Fernandez, P., Boitel, B., *et al.* (2005). Conserved autophosphorylation pattern in activation loops and juxtamembrane regions of *Mycobacterium tuberculosis* Ser/Thr protein kinases. *Biochemical and Biophysical Research Communications* **333,** 858-867.

Feltcher, M.E., Gibbons, H.S., Ligon, L.S. and Braunstein, M. (2013). Protein Export by the Mycobacterial SecA2 System Is Determined by the Preprotein Mature Domain. *Journal of Bacteriology* **195,** 672-681.

Feltcher, M.E., Sullivan, J.T. and Braunstein, M. (2010). Protein export systems of *Mycobacterium tuberculosis*: novel targets for drug development? *Future Microbiol* **5,** 1581-1597.

Fleischmann, R.D., Alland, D., Eisen, J.A., Carpenter, L., White, O., Peterson, J., *et al.* (2002). Whole-Genome Comparison of *Mycobacterium tuberculosis* Clinical and Laboratory Strains. *Journal of Bacteriology* **184,** 5479-5490.

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

Galagan, J.E. (2014). Genomic insights into tuberculosis. *Nature Reviews Genetics* **15,** 307.

Gandhi, N.R., Moll, A., Sturm, A.W., Pawinski, R., Govender, T., Lalloo, U., *et al.* (2006). Extensively drug-resistant tuberculosis as a cause of death in patients co-infected with tuberculosis and HIV in a rural area of South Africa. *The Lancet* **368,** 1575-1580.

Gao, C., Mao, S., Kaufmann, G., Wirsching, P., Lerner, R.A. and Janda, K.D. (2002). A method for the generation of combinatorial antibody libraries using pIX phage display. *Proceedings of the National Academy of Sciences* **99,** 12612-12616.

González, E., Robles, Y., Govezensky, T., Bobes, R.J., Gevorkian, G. and Manoutcharian, K. (2010). Isolation of neurocysticercosis-related antigens from a genomic phage display library of *Taenia solium*. *Journal of biomolecular screening* **15,** 1268-1273.

Goosens, V.J., Monteferrante, C.G. and van Dijl, J.M. (2014). The Tat system of Gram-positive bacteria. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research* **1843,** 1698-1706.

Gray, T.A., Clark, R.R., Boucher, N., Lapierre, P., Smith, C. and Derbyshire, K.M. (2016). Intercellular communication and conjugation are mediated by ESX secretion systems in mycobacteria. *Science* **354,** 347-350.

Gupta, A., Shrivastava, N., Grover, P., Singh, A., Mathur, K., Verma, V., *et al.* (2013). A Novel Helper Phage Enabling Construction of Genome-Scale ORF-Enriched Phage Display Libraries. *PLoS ONE* **8,** e75212.

Harth, G. and Horwitz, M.A. (1999a). Export of Recombinant *Mycobacterium tuberculosis* Superoxide Dismutase Is Dependent upon Both Information in the Protein and Mycobacterial Export Machinery: A MODEL FOR STUDYING EXPORT OF LEADERLESS PROTEINS BY PATHOGENIC MYCOBACTERIA. *Journal of Biological Chemistry* **274,** 4281-4292.

Harth, G. and Horwitz, M.A. (1999b). An inhibitor of exported *Mycobacterium tuberculosis* glutamine synthetase selectively blocks the growth of pathogenic mycobacteria in axenic culture and in human monocytes: extracellular proteins as potential novel drug targets. *J Exp Med* **189,** 1425-1436.

Hoogenboom, H.R., de Bruıne, A.P., Hufton, S.E., Hoet, R.M., Arends, J.-W. and Roovers, R.C. (1998). Antibody phage display technology and its applications. *Immunotechnology* **4,** 1-20.

Hou, J.M., D'Lima, N.G., Rigel, N.W., Gibbons, H.S., McCann, J.R., Braunstein, M. and Teschke, C.M. (2008). ATPase Activity of *Mycobacterium tuberculosis* SecA1 and SecA2 Proteins and Its Importance for SecA2 Function in Macrophages. *Journal of Bacteriology* **190,** 4880-4887.

Houben, E.N., Korotkov, K.V. and Bitter, W. (2014). Take five—Type VII secretion systems of Mycobacteria. *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research* **1843,** 1707-1716.

Ivankov, D.N., Payne, S.H., Galperin, M.Y., Bonissone, S., Pevzner, P.A. and Frishman, D. (2013). How many signal peptides are there in bacteria? *Environmental microbiology* **15,** 983-990.

Jankovic, D., Collett, M.A., Lubbers, M.W. and Rakonjac, J. (2007). Direct selection and phage display of a Gram-positive secretome. *Genome biology* **8,** R266.

Kapp, K., Schrempf, S., Lemberg, M.K. and Dobberstein, B. (2000). Post-targeting functions of signal peptides.

Koch, R. (1882). Die aetiologie der tuberkulose. *Berl Klin Wchschr* **19,** 221.

Kubica, G.P., Kim, T.H. and Dunbar, F.P. (1972). Designation of Strain H37Rv as the Neotype of *Mycobacterium tuberculosis*. *International Journal of Systematic and Evolutionary Microbiology* **22,** 99-106.

Lee, V.T. and Schneewind, O. (2001). Protein secretion and the pathogenesis of bacterial infections. *Genes & development* **15,** 1725-1752.

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

Lew, J.M., Mao, C., Shukla, M., Warren, A., Will, R., Kuznetsov, D., *et al.* (2013). Database resources for the Tuberculosis community. *Tuberculosis (Edinburgh, Scotland)* **93,** 12-17.

Ligon, L.S., Hayden, J.D. and Braunstein, M. (2012). The ins and outs of *Mycobacterium tuberculosis* protein export. *Tuberculosis* **92,** 121-132.

Liu, S., Han, W., Sun, C., Lei, L. and Feng, X. (2013). Identification of Two New Virulence Factors of *Mycobacterium Tuberculosis* that Induce Multifunctional CD4 T-cell Responses. *J Mycobac Dis S* **6,** 2161-1068.

Liu, S., Han, W., Sun, C., Lei, L., Feng, X., Yan, S., *et al.* (2011). Subtractive screening with the *Mycobacterium tuberculosis* surface protein phage display library. *Tuberculosis* **91,** 579-586.

MacGurn, J.A., Raghavan, S., Stanley, S.A. and Cox, J.S. (2005). A non-RD1 gene cluster is required for Snm secretion in *Mycobacterium tuberculosis*. *Mol. Microbiol.* **57,** 1653-1663.

Maffei, B., Francetic, O. and Subtil, A. (2017). Tracking Proteins Secreted by Bacteria: What's in the Toolbox? *Frontiers in Cellular and Infection Microbiology* **7**.

Majlessi, L., Prados-Rosales, R., Casadevall, A. and Brosch, R. (2015). Release of mycobacterial antigens. *Immunological Reviews* **264,** 25-45.

McCafferty, J., Griffiths, A., Winter, G. and Chiswell, D. (1990). Phage antibodies: filamentous phage displaying antibody variable domains. *Nature* **348,** 552 - 554.

McCann, J.R., McDonough, J.A., Sullivan, J.T., Feltcher, M.E. and Braunstein, M. (2011). Genome-Wide Identification of *Mycobacterium tuberculosis* Exported Proteins with Roles in Intracellular Growth. *Journal of Bacteriology* **193,** 854-861.

McDonough , J.A., McCann, J.R., Tekippe, E.M., Silverman, J.S., Rigel, N.W. and Braunstein, M. (2008). Identification of Functional Tat Signal Sequences in *Mycobacterium tuberculosis* Proteins. *Journal of Bacteriology* **190,** 6428-6438.

McEvoy, C.R., Cloete, R., Muller, B., Schurch, A.C., van Helden, P.D., Gagneux, S., *et al.* (2012). Comparative analysis of *Mycobacterium tuberculosis* pe and ppe genes reveals high sequence variation and an apparent absence of selective constraints. *PLoS One* **7,** e30593.

McLaughlin, B., Chon, J.S., MacGurn, J.A., Carlsson, F., Cheng, T.L., Cox, J.S. and Brown, E.J. (2007). A mycobacterium ESX-1-Secreted virulence factor with unique requirements for export. *Plos Pathogens* **3,** 1051-1061.

Mullen, L.M., Nair, S.P., Ward, J.M., Rycroft, A.N. and Henderson, B. (2006). Phage display in the study of infectious diseases. *Trends in Microbiology* **14,** 141-147.

Nagata, R., Kawaji, S. and Mori, Y. (2013). Use of enoyl coenzyme A hydratase of *Mycobacterium avium* subsp. paratuberculosis for the serological diagnosis of Johne's disease. *Veterinary immunology and immunopathology* **155,** 253-258.

Naseem, S., Meens, J., Jores, J., Heller, M., Dubel, S., Hust, M. and Gerlach, G.-F. (2010). Phage display-based identification and potential diagnostic application of novel antigens from *Mycoplasma mycoides* subsp. mycoides small colony type. *Vet Microbiol* **142,** 285 - 292.

Natale, P., Brüser, T. and Driessen, A.J.M. (2008). Sec- and Tat-mediated protein secretion across the bacterial cytoplasmic membrane—Distinct translocases and mechanisms. *Biochimica et Biophysica Acta (BBA) - Biomembranes* **1778,** 1735-1756.

Nguyen, L. (2016). Antibiotic resistance mechanisms in *M. tuberculosis*: an update. *Archives of Toxicology***,** 1-20.

NHLS (2009). National Institute for Communicable Diseases Annual Report 2009: National Health Laboratory Services. *http://www.nicd.ac.za/assets/files/Annual_report_2009.pdf*.

Niederweis, M., Danilchanka, O., Huff, J., Hoffmann, C. and Engelhardt, H. (2010). Mycobacterial outer membranes: in search of proteins. *Trends in Microbiology* **18,** 109-116.

O'Connell, D., Becerril, B., Roy-Burman, A., Daws, M. and Marks, J.D. (2002). Phage versus phagemid libraries for generation of human monoclonal antibodies. *Journal of molecular biology* **321,** 49-56.

Okumura, K., Kato, M., Kirikae, T., Kayano, M. and Miyoshi-Akiyama, T. (2015). Construction of a virtual *Mycobacterium tuberculosis* consensus genome and its application to data from a next generation sequencer. *BMC Genomics* **16,** 218.

Palanisamy, G.S., DuTeau, N., Eisenach, K.D., Cave, D.M., Theus, S.A., Kreiswirth, B.N., *et al.* (2009). Clinical strains of *Mycobacterium tuberculosis* display a wide range of virulence in guinea pigs. *Tuberculosis* **89,** 203-209.

Peters, J., Calder, B., Gonnelli, G., Degroeve, S., Rajaonarifara, E., Mulder, N., *et al.* (2016). Identification of quantitative proteomic differences between *Mycobacterium tuberculosis* lineages with altered virulence. *Frontiers in Microbiology* **7**.

Pillay, M. and Sturm, A.W. (2007). Evolution of the extensively drug-resistant F15/LAM4/KZN strain of *Mycobacterium tuberculosis* in KwaZulu-Natal, South Africa. *Clinical infectious diseases* **45,** 1409-1414.

Posey, J.E., Shinnick, T.A. and Quinn, F.D. (2006). Characterization of the twin-arginine translocase secretion system of *Mycobacterium smegmatis*. *Journal of Bacteriology* **188,** 1332-1340.

Rakonjac, J. (2012). Filamentous bacteriophages: biology and applications. *eLS*.

Rashid, M., Saha, S. and Raghava, G.P. (2007). Support Vector Machine-based method for predicting subcellular localization of mycobacterial proteins using evolutionary information and motifs. *BMC bioinformatics* **8,** 1.

Rodi, D.J. and Makowski, L. (1999). Phage-display technology–finding a needle in a vast molecular haystack. *Current opinion in biotechnology* **10,** 87-93.

Rosander, A., Bjerketorp, J., Frykberg, L. and Jacobsson, K. (2002). Phage display as a novel screening method to identify extracellular proteins. *Journal of Microbiological Methods* **51,** 43-55.

Sakula, A. (1983). BCG: who were Calmette and Guerin? *Thorax* **38,** 806.

Smith, G. (1985). Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface. *Science* **228,** 1315 - 1317.

Stoop, E.J.M., Bitter, W. and van der Sar, A.M. (2012). Tubercle bacilli rely on a type VII army for pathogenicity. *Trends in Microbiology* **20,** 477-484.

Thie, H., Schirrmann, T., Paschke, M., Dübel, S. and Hust, M. (2008). SRP and Sec pathway leader peptides for antibody phage display and antibody fragment production in E. coli. *New biotechnology* **25,** 49-54.

Ting, Y.T., Harris, P.W., Batot, G., Brimble, M.A., Baker, E.N. and Young, P.G. (2016). Peptide binding to a bacterial signal peptidase visualized by peptide tethering and carrier-driven crystallization. *IUCrJ* **3,** 10-19.

Tsirigotaki, A., De Geyter, J., Šoštaric, N., Economou, A. and Karamanou, S. (2017). Protein export through the bacterial Sec pathway. *Nature Reviews Microbiology* **15,** 21-36.

van Pittius, N.C.G., Gamieldien, J., Hide, W., Brown, G.D., Siezen, R.J. and Beyers, A.D. (2001). The ESAT-6 gene cluster of *Mycobacterium tuberculosis* and other high G+ C Gram-positive bacteria. *Genome biology* **2,** research0044. 0041.

Velayati, A.A., Masjedi, M.R., Farnia, P., Tabarsi, P., Ghanavi, J., ZiaZarifi, A.H. and Hoffner, S.E. (2009). Emergence of new forms of totally drug-resistant tuberculosis bacilli: super extensively drug-resistant tuberculosis or totally drug-resistant strains in iran. *Chest* **136,** 420-425.

Von Pirquet, C. (1909). Frequency of tuberculosis in childhood. *Journal of the American Medical Association* **LII,** 675-678.

Wall, T., Roos, S., Jacobsson, K., Rosander, A. and Jonsson, H. (2003). Phage display reveals 52 novel extracellular and transmembrane proteins from *Lactobacillus reuteri* DSM 20016T. *Microbiology* **149,** 3493-3505.

Wang, L.-F. and Yu, M. (2004). Epitope identification and discovery using phage display libraries: applications in vaccine development and diagnostics. *Current drug targets* **5,** 1-15.

WHO (2012). Global Tuberculosis Report 2012. *World Health Organization*.

WHO (2017). Global Tuberculosis Report 2017 *World Health Organization*.

CHAPTER 2: Identification of unique essential proteins from a *M. tuberculosis* F15/LAM4/KZN phage secretome library

Thamsanqa Emmanuel Chiliza[1], Manormoney Pillay[2], and Balakrishna Pillay[1*]

[1]Discipline of Microbiology, School of Life Sciences, College of Agriculture, Engineering and Science, University of KwaZulu-Natal, Westville Campus, South Africa;

[2]Medical Microbiology and Infection Control, School of Laboratory Medicine and Medical Sciences, College of Health Sciences, University of KwaZulu-Natal, Durban, South Africa

Original article published in: Pathogens and Disease 2017; 75(1): ftx001-ftx001

2.1 ABSTRACT

Mycobacterium tuberculosis (*Mtb*) is the causative agent of tuberculosis disease (TB), the leading cause of death from bacterial infection worldwide. Although treatable, the resurgence of multi- and extensively-drug resistant TB is a major setback for the fight against TB globally. Consequently, there is an urgent need for new *Mtb* derived biomarkers for use in the design of new drugs and rapid point-of-care diagnostic or prognostic tools for management of TB transmission. Therefore, the present study aimed to identify unique *Mtb* secreted proteins from the extensively-drug resistant *Mtb* F15/LAM4/KZN phage secretome library. A whole genome library was constructed using genomic DNA fragments of *Mtb* F15/LAM4/KZN strain. A phage secretome sub-library of $8x10^3$ clones was prepared and phage DNA was sequenced from 120 randomly selected clones. DNA sequence BLAST analysis identified 86 open reading frames. Using bioinformatics tools and databases, ten proteins essential for *in vivo* growth and survival of *Mtb* (Nrp, PssA, MmpL5, SirA, GatB, EspA, TopA, EccCa1, Rv1634 and Rv3103c) were identified. Proteins essential for growth and survival of *Mtb* during infection have potential application in the development of diagnostic tools, new drugs and vaccines. Further studies will be conducted to evaluate their potential application in the fight against TB.

Keywords

M. tuberculosis, biomarkers, phage display, secreted proteins, diagnosis, vaccine

2.2 INTRODUCTION

Mycobacterium tuberculosis, the causative agent of tuberculosis (TB), is harboured asymptomatically by one-third of the world's population. In 2014, approximately 10.4 million new infections and 1.8 million deaths were recorded globally (WHO, 2016). The emergence and spread of multi- and extensively drug resistant (MDR/XDR) strains of *M. tuberculosis* as well as high human immunodeficiency virus (HIV) co-infection rates (WHO, 2016), have complicated the fight against TB. Improved management of TB transmission can be achieved by early and rapid detection, and timely administration of effective TB treatment. The slow progress in the development of rapid point-of-care diagnostic assays, new drugs, vaccines and immunotherapeutic agents has largely been due to the availability of only a few, novel *M. tuberculosis* derived biomarkers (Wallis *et al.*, 2013). *M. tuberculosis* proteins secreted into its surrounding environment play an important role in host-pathogen interaction and can be ideal candidate biomarkers for the development of new drugs; vaccines; and for use in diagnosis (Andersen *et al.*, 1991, Harth & Horwitz, 1999). Preferably, a biomarker should be easily accessible as targets for effective, less invasive sampling for diagnosis and/or therapeutic interventions.

M. tuberculosis secretes a variety of extracellular virulence factors that play a role in adhesion and invasion of host target cells using various secretion pathways (Harth & Horwitz, 1999, Bendtsen *et al.*, 2005, Abdallah *et al.*, 2006). The unique cell wall and the associated extracellular virulence factors have been credited for the success of this pathogen (Champion & Cox, 2007) and may represent ideal candidates for the development of new drugs, vaccines, and TB diagnostics. The main *M. tuberculosis* protein secretory pathways are the Sec-dependent and Twin-arginine translocation (Tat) pathways (Ligon *et al.*, 2012). Two specialized protein secretion systems, the SecA2-dependent and the type VII secretion system or ESX, are also present in *M. tuberculosis* (Bendtsen *et al.*, 2005, Thakur *et al.*, 2016). The

ESX system is responsible for secretion of small virulence proteins, such as early secreted antigenic target of 6 kDa (ESAT-6) and culture filtrate protein of 10 kDa (CFP-10). *M. tuberculosis* contains the genetic information for five type VII secretion machineries [ESX-1, ESX-2, ESX-3, ESX-4 and ESX-5] (Gey van Pittius *et al.*, 2006), suggesting the importance of Sec-independent protein secretion for this pathogen.

Generally, mass spectrometry based methods are used to study *M. tuberculosis* secreted proteins in culture filtrate (Ge *et al.*, 2003, de Souza *et al.*, 2011, Zheng *et al.*, 2013). While these methods have been highly effective, their main limitation is inconsistent protein expression owing to varying environmental growth conditions and the inability to detect proteins expressed at low concentrations (Forrellad *et al.*, 2013). Therefore, the phage display method offers a good alternative for bacterial secretome repertoire analysis (Jacobsson *et al.*, 2003, Rosander *et al.*, 2003, Wall *et al.*, 2003, Karlström *et al.*, 2004, Jankovic *et al.*, 2007, Rosander *et al.*, 2011, Gagic *et al.*, 2013). In 2011, *M. tuberculosis* H37Rv phage secretome library was used to identify six immunogenic proteins [MPT64 (Rv1980c), Ag85B (Rv1886c), cpsA (Rv3484), LpqA (Rv3016), PapA2 (Rv3820c) and EsxO (Rv2346c)] with potential as vaccine or diagnostic candidates (Liu *et al.*, 2011, Liu *et al.*, 2013).

Since the availability of the *M. tuberculosis* H37Rv complete genome sequence (Cole *et al.*, 1998), many *in vitro* and *in silico* studies have been conducted to analyse expression, identify open reading frames (ORFs) and predict their potential functions. However, the laboratory strain H37Rv has been reported to accumulate adaptation changes during repeated passage in culture, resulting in phenotypic alterations and partial attenuation (Ioerger *et al.*, 2010, Mehaffy *et al.*, 2010). Thus, the varying degree of pathogenicity among clinical *M. tuberculosis* strains (Palanisamy *et al.*, 2009), the phenotypic changes observed in H37Rv strain (Devasundaram & Raja, 2016), and the threat posed by emergence of highly transmissible MDR/XDR strains (Pillay & Sturm, 2007, Peters *et al.*, 2016) indicate a need to study clinically relevant strains

that may provide more insight into wild-type virulent behavior (Palanisamy *et al.*, 2009, Devasundaram & Raja, 2016).

The current study aimed to identify appropriate *M. tuberculosis* derived diagnostic and drug target candidate protein biomarkers. The phage display method was used to construct a whole genome phage library of XDR *M. tuberculosis* F15/LAM4/KZN genomic DNA fragments. A phage secretome sub-library of $8x10^3$ clones was generated and sequence analysis was performed on randomly selected clones. The encoded ORFs were identified from the *M. tuberculosis* comparative database. Different bioinformatic tools, gene enrichment analysis and other databases were used to deduce essential and virulence associated *M. tuberculosis* protein biomarkers. The identified secretory protein biomarkers may have potential as diagnostic biomarkers, vaccine candidates and target for immunotherapeutic agents and drug discovery.

2.3 MATERIALS AND METHODS

2.3.1 Bacterial Strains, Phage Display Vector and Helper Phage

The XDR *M. tuberculosis* F15/LAM4/KZN (KZN605) strain was obtained from the archived collection at Medical Microbiology and Infection Control, University of KwaZulu-Natal. The XDR-KZN605 strain (GenBank accession: NC_018078) is resistant to isoniazid, rifampicin, ethambutol, ofloxacin, kanamycin, capreomycin and niacinamide. IS6110-restriction fragment length polymorphism analysis identified KZN605 to belong to the F15/LAM4/KZN strain family, while spoligotyping categorised it as the shared type (ST) 60 based on the absence of spacers 21-24, 33-36, 40 (Naidoo & Pillay, 2014). The strain was cultured aerobically in Middlebrook 7H9 broth (Difco), supplemented with 10 % oleic acid, albumin, dextrose, catalase (OADC), 0.05 % Tween 80 and 0.5 % glycerol at 37^{o}C with shaking to an OD_{600} 1. The pDJ01 phagemid vector, helper phage VCSM13d3 and pJARA plasmid DNA were a gift

from Dr Jasna Rakonjac of Massey University. The *E. coli* TG1 cells were from Lucigen Corporation.

2.3.2 *M. tuberculosis* Whole Genome Library Construction

Genomic DNA fragments of XDR *M. tuberculosis* F15/LAM4/KZN strain were used to construct a phage secretory protein repertoire library in the pDJ01 phage display system (Jankovic *et al.,* 2007). The minimum size of the library required to represent the whole *M. tuberculosis* genome was calculated using the following formula (Jacobsson *et al.,* 2003):

$$N = \ln (1-P)/\ln (1-f)$$

Where: P is the desired probability

f is the fractional proportion of the genome in a single recombinant

N is the necessary number of recombinants

Since the *M. tuberculosis* genome size is 4.41 $x10^6$ base pairs and the average fragment size is estimated at 500 base pairs,

$$N = \ln (1 - 0.99)/\ln [1-(500bp/4.41x10^6bp)]$$

$$N= 4x10^4 \text{ clones}$$

Genomic DNA was extracted using a modified sodium chloride-cetyl trimethyl-ammonium bromide (CTAB-NaCl) method (van Soolingen *et al.*, 1994). Genomic DNA (200ng/uL) was fragmented by sonication on ice for 8 min using the Sonic Ruptor 400 (OMNI International). Fragments of 150 to 1500 base pair sizes were purified using a PCR clean-up kit (Macherey-Nagel GmbH & Co. KG), followed by repair of blunt ends by End Repair Enzyme Mix (Thermo Fisher Scientific Inc.). The phagemid pDJ01 vector DNA was digested with *SmaI* (Thermo Fisher Scientific) and dephosphorylated with FastAP (Thermo Fisher Scientific Inc.)

in the same reaction at 37ºC to generate blunt ends. DNA fragments (~8µg) were ligated into the pDJ01 vector (~8µg) in 1:1 ratio using a Rapid Ligation Kit (Thermo Fisher Scientific Inc.). After desalting, the ligation reaction mixtures were electroporated into electro-competent *E. coli* TG1 cells (Lucigen Corporation), and incubated in 2xTY broth at 37ºC with aeration for 1hr. Serial dilutions were prepared for library size determination and the remaining cells were plated on TYE agar supplemented with 20 µg/mL chloramphenicol (Cm) and incubated overnight at 30ºC. These cultures were used to prepare 1 mL aliquots of 15% glycerol stock culture of the whole genome library and stored at -70ºC.

2.3.3 *M. tuberculosis* Phage Secretome Sub-library Preparation

The phage display secretome library was prepared as follows: 1 mL of whole genome library stock culture was inoculated into 25 mL of 2xTY-Cm. The exponentially growing culture (OD_{600} 0.2) was infected with helper phage VCSM13d3 [phage to bacterium MOI = 50:1] for 1hr at 37ºC. Cells were centrifuged and the pellet re-suspended in 250 ml of 2xYT-Cm with 50 µg/mL kanamycin and incubated for 4hrs at 37ºC. After centrifugation at 10,000xg for 20 min, phagemid particles in the supernatant were precipitated with 5% (w/v) PEG/0.5 M NaCl overnight at 4ºC. The phagemid particles were centrifuged and pellet re-suspended in TN buffer (10 mmol/L Tris-HCl pH 7.6; 50 mmol/L NaCl). Defective phagemid particles were eliminated by treatment of the phagemid suspension ($1x10^{12}$ CFU/mL) with sarcosyl at a final concentration of 0.1% (w/v), followed by DNase I (100 U) in the presence of 5 mM $MgCl_2$, and then inactivated by EDTA (20 mM). The remaining sarcosyl-resistant phage particles were precipitated with PEG/NaCl solution as above. For preparation of secretome sub-library, the ssDNA was extracted from sarcosyl resistant phagemid particles by first incubating at 70ºC for 10 min in the presence of 1.2% (w/v) SDS. The phenol-chloroform DNA isolation method was

used for further purification of ssDNA (Su *et al.*, 1998). The surface protein library was amplified by transforming the ssDNA into *E. coli* TG1 that was grown on 2xYT-Cm plates overnight at 37ºC.

2.3.4 DNA Sequencing

The recombinant phage DNA of randomly selected clones was sequenced at Inqaba Biotechnical Industries (Pty) Ltd (South Africa) using the primer set flanking the vector cloning site, pDJ01R02 (5'-CCGGAAACGTCACCAATGAA-3') and pDJF03 (5'-ATGTTGCTGTTGATTCTTCA-3'). The DNA sequences were analysed using the CLCBio Workbench (v. 2.0).

2.3.5 Prediction of Functional Proteins, Signal Peptides and Transmembrane Proteins

BLAST analysis of the nucleotide sequences against the *M. tuberculosis* comparative database was performed to retrieve encoded open reading frames (ORFs). Tuberculist database (http://tuberculist.epfl.ch/) was used to determine functional categories of identified proteins (Lew *et al.*, 2011). Unknown proteins not documented on Tuberculist database were placed in the unknown functional category.

For prediction of Sec-dependent and Tat-dependent amino terminal signal peptide for secretion, SignalP 4.1 and TatP 1.0 were used (Bendtsen *et al.*, 2005a, Petersen *et al.*, 2011). SecretomeP 2.0 was used for prediction of non-classical secreted proteins (Bendtsen *et al.*, 2005b), and transmembrane proteins were predicted using TMHMM 2.0 (Krogh *et al.*, 2001). LipoP 1.0 was used for prediction of lipoprotein signal peptide (Rahman *et al.*, 2008). All software are freely available from the Centre for Biological Sequence Analysis at the Technical University of Denmark (http://www.cbs.dtu.dk/services). Alternative software, PRED-SIGNAL and PRED-TAT were used for supplementary analysis to predict Sec-dependent and

Tat-dependent amino terminal signal peptide, respectively (Bagos *et al.*, 2009, Bagos *et al.*, 2010). PRED-LIPO was used for prediction of lipoprotein signal peptide (Bagos *et al.*, 2008). The alternative software were accessed via the Computational Genetics Research Group at the Department of Computer Science and Biomedical Informatics (http://www.compgen.org/tools) at the University of Thessaly. The subcellular localization of uncharacterized proteins was predicted with TBPred designed for analysis of mycobacterial proteins (http://www.imtech.res.in/raghava/tbpred/) (Rashid *et al.*, 2007).

2.3.6 Gene Enrichment Analysis and Functional Annotation

The Universal Protein Resource (UniProt) database (http://www.uniprot.org/blast/) accession numbers were retrieved by BLAST analysis of protein sequences against UniProt database (Suzek *et al.*, 2015). Gene enrichment analysis was performed using the database for annotation, visualization and integrated discovery (DAVID) (https://david.ncifcrf.gov/) for gene ontology (GO) terms and functional annotation clusters (Huang *et al.*, 2009). The significantly enriched ($p < 0.05$) GO biological process (BP), cellular component (CC) and molecular function (MF) were identified. Pathosystems Research Intergrated Centre (PATRIC) database (https://www.patricbrc.org/) was used to identify essential and virulence genes as well as genes required for *M. tuberculosis* growth and survival within host (Wattam *et al.*, 2014).

2.4 RESULTS AND DISCUSSION

2.4.1 *M. tuberculosis* Whole Genome Library

A whole genome XDR *M. tuberculosis* phage library was successfully constructed by cloning DNA fragment inserts into pDJ01 phage display vector. The library size was ~1.76×10^6 clones, exceeding the desired calculated size. Library diversity was confirmed on randomly selected

clones by colony PCR (Fig. 2.1). More than 90% of the library contained the *M. tuberculosis* DNA fragments. The library size was approximately ten fold larger than that obtained by Liu et al. (2011). This is probably due to the total genomic DNA and fragments sizes (150bp to 1500bp) used as compared to 300bp to 1500bp fragments in Liu et al. (2011).

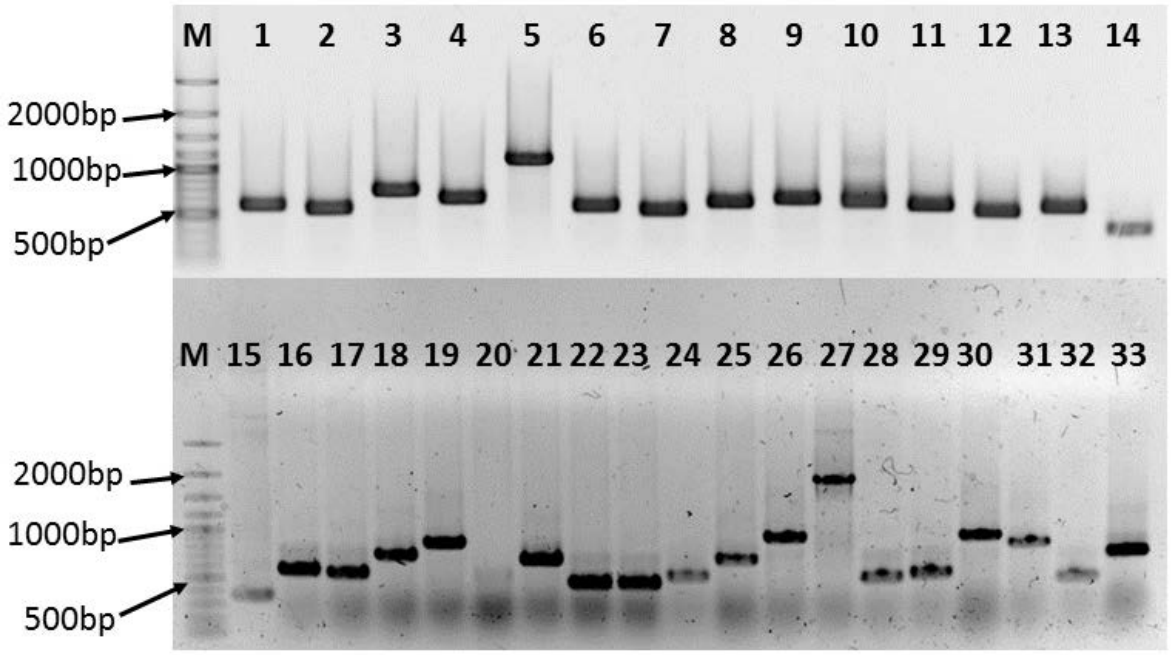

Fig. 2. 1: Colony PCR of randomly selected clones of the whole genome *M. tuberculosis* phage library. Lane M is 100bp ladder and 1-33 are PCR amplicons of library clones.

2.4.2 *M. tuberculosis* Phage Secretome

M. tuberculosis secretes proteins via classical (Sec- and Tat-dependent) and non-classical (Sec-independent) pathways. The classical pathway secreted proteins must have the signal peptide sequence for successful secretion (Ligon *et al.*, 2012). In the present study, the signal peptide sequence and transmembrane helices are responsible for guiding secretome to secretion and anchoring into membrane for display on phage surface, respectively (Gagic *et al.*, 2016). Therefore, successful display of *M. tuberculosis* peptide protein will require the DNA insert to

encode signal sequence or directive transmembrane helices peptide and be cloned in-frame to the vector resulting in a stable phage particle.

Some *M. tuberculosis* virulence factors are secreted through the non-classical type VII secretion pathway. Usually, the type VII secretion pathway substrates are secreted as a complex of two or more proteins as they need each other for successful translocation through membrane channels (Chen *et al.*, 2013). Even though some of the type VII secreted proteins will be in-frame, these might not be displayed on the phage surface as their secretion partners required to form a complex will be missing. The surrogate *E. coli* used for phage display is not expected to mimic the *M. tuberculosis* type VII secretion pathway. This should limit the discovery of most type VII/ESX secretion substrates. However, the type VII secreted membrane proteins may find their way into the phage secretome via transmembrane helices directing membrane localization.

Generally, it is expected that approximately 1-2% of DNA inserts will be in-frame and result in sarcosyl-resistant phagemid particles (Jacobsson & Frykberg, 2001). Furthermore, some *M. tuberculosis* signal sequences may not function efficiently when expressed in *E. coli* (Smith, 1985), perhaps due to different codon usage and GC content (Poquet *et al.*, 1998). Therefore, the majority of the *M. tuberculosis* DNA inserts will be packaged into sarcosyl-sensitive phagemid as they lack an in-frame signal sequence or encode some type VII secretion pathways proteins. In this study, a *M. tuberculosis* phage secretome sub-library of ~$8x10^3$ clones was obtained from the DNA of sarcosyl resistant phages [Fig. 2.2a]. Colony PCR confirmed the presence of insert DNA in randomly selected clones [Fig. 2.2b].

Phage DNA from 120 randomly selected sub-library clones were sequenced using primers flanking the pDJ01 vector cloning site. Analysis of DNA sequences using CLC Workbench software demonstrated that 18% of sequences comprised repeat clones and 98 distinct DNA sequences were identified (Appendix A, Table S2.1). BLAST analysis of distinct DNA

sequences against the *M. tuberculosis* comparative database identified 86 corresponding ORFs and their protein sequences were retrieved for further analysis.

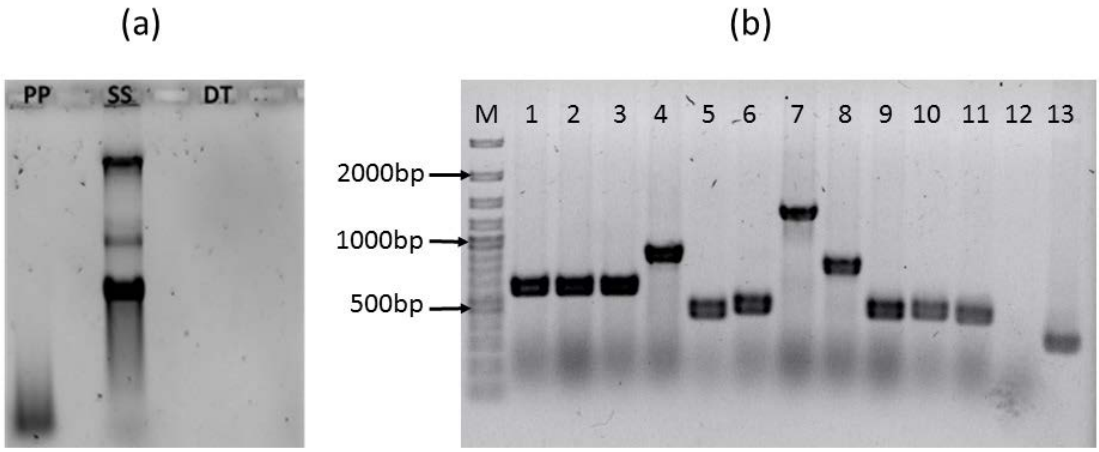

Fig. 2. 2: (a) Selective disassembly of phage particles display no or non-secretory protein. PP: phage particles prepared from *M. tuberculosis* phage library; SS: DNA released from sarcosyl sensitive phage particles; DT: DNase I treatment results showing almost complete removal or digestion of phage DNA released during SS step. (b) PCR amplicons of randomly selected clones of the phage secretome sub-library.

2.4.3 Functional Categories of Identified Proteins

All 86 proteins were assigned a functional category as per Tuberculist database (Appendix A, Table S2.2). The distribution of proteins to their functional category is presented in Fig. 2.3. The majority of proteins (34/86) belong to cell wall and cell processes category. The cell wall and cell processes proteins include two immunogenic proteins Mpt63 (Rv1926c) and Mpt64 (Rv1980c), two lipoproteins, LpqX (Rv1228) and LprO (Rv0179c), metal cation transporter CtpH (Rv0425c), MmpL12 (Rv1522c) involved in fatty acid transport, EmbC (Rv3793), alanine-leucine rich (Rv2693c) and alanine-valine-leucine rich (Rv2729c) suspected to be involved in the active evasion of the host immune response. Two ESX-1 secretion system

proteins EspA (Rv3616c) and the conserved component protein EccCa1 (Rv3870) were also associated with cell wall and cell processes.

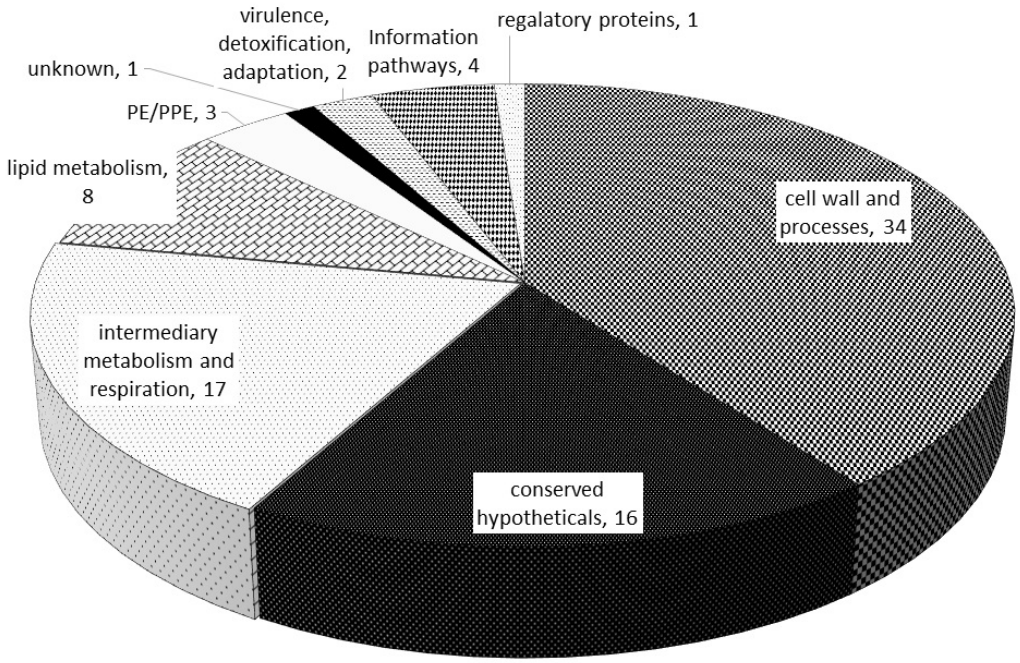

Fig. 2. 3: Distribution of the functional categories of 86 identified ORFs from F15/LAM4/KZN phage displayed secretome. The number of proteins in the different functional categories included: cell wall and cell processes (34), conserved hypothetical proteins (16), intermediary metabolism and respiration (17), lipid metabolism (8), information pathways (4), PE/PPE family proteins (3), virulence, detoxification, and adaptation (2), regulatory proteins (1) and unknown (1) function according to UniProt. Functional group codes were obtained from the Tuberculist database web server except for the unknown protein that was confirmed by the UniProt server.

There were seventeen intermediary metabolism and respiration category proteins that included serine protease PepA (Rv0125), acyltransferase (Rv1254), putative ligase (Rv3712), and 4-hydroxy-2-oxovalerate aldolase (Rv3469c). Sixteen were conserved hypothetical proteins, including ala-, pro-rich protein (Rv1157c) and secreted protein Rv1268c. Eight lipid metabolism category proteins identified included the secreted Ag85B (Rv1886c, Ag85C

(Rv0129c), peptide synthetase Nrp (Rv0101) and triacylglycerol synthase Tgs2 (Rv3734c). Other functional categorized proteins included four information pathways (Rv1700, Rv1981c, Rv3009c, and Rv3646c); three PPE family proteins (PPE32, PPE43 and PPE54); two virulence, detoxification and adaptation proteins, Rv1478 and Rv1566c associated with host invasion; and one regulatory protein (Rv1267c). One un-annotated protein, MT3042 was classified as a protein of unknown functional category.

2.4.4 Secretion Signal Peptide and Cellular Localization

Using different bioinformatics tools, the analysis of ORF protein sequences revealed Sec- and Tat- secreted proteins, membrane proteins and the secreted leaderless proteins. Twenty-seven proteins including 12 membrane proteins were predicted to harbor Sec-dependent N-terminal signal peptides by SignalP 4.1 and PRED-SIGNAL. Twenty-one proteins including 3 membrane proteins Rv1230c, Rv3162c and Rv3395A were predicted to contain a Tat signal peptide by TatP 1.0 and PRED-TAT. Twenty-one proteins were predicted trans-membrane proteins with 2 to 14 transmembrane helical structures by TMHMM 2.0 (Table 2.1).

In this study, 32 identified leaderless proteins were analyzed using SecretomeP 2.0 and TBPred databases to predict their subcellular localization (Table 2.2). Nine secreted proteins including two PPE proteins were identified by SecretomeP 2.0, whilst 21 of the 23 remaining proteins were confirmed as either secreted, membrane or membrane attached proteins by TBPred. Interestingly, 2 proteins Rv3616c and Rv3703c were predicted to be cytoplasmic proteins by TBPred database (Table 2.2). Rv3616c encodes the indirect ESX-1 pathway substrate, EspA, which is associated with *M. tuberculosis* virulence (Garces et al., 2010), whilst the Rv3703c gene is a member of gene cluster or operon (Rv3704c–Rv3700c) responsible for ergothioneine (EGT) biosynthesis in *M. tuberculosis* (Glökler et al., 2010). EGT is a low-molecular weight

protein essential for survival of *M. tuberculosis* in macrophages and is associated with antimycobacterial drug resistance (Glökler et al., 2010; Xu et al., 2015).

Table 2. 1**:** List of Sec- and Tat-dependent N-terminal signal harbouring and transmembrane proteins

Sec-dependent		**Tat-dependent**		**TMHMM Predicted**	
Rv no.	**Gene name**	**Rv no.**	**Gene name**	**Rv no.**	**Gene name**
Rv0116c	*ldtA*	-	*MT3042*	[a]Rv0236c	*aftD*
Rv0320	-	Rv0125	*pepA*	[a]Rv0425c	*ctpH*
Rv0559c	-	Rv0129c	*fbpC*	[a]Rv0436c	*pssA*
Rv0603	-	Rv0179c	*lprO*	Rv0676c	*mmpL5*
Rv0675	*echA5*	Rv0203	-	[a]Rv0842	-
Rv1157c	-	Rv0455c	-	[a]Rv1029	*kdpA*
Rv1228	*lpqX*	Rv1268c	-	Rv1200	-
Rv1478	-	Rv1291c	-	[b]Rv1230c	-
Rv1804	-	Rv1435c	-	[a]Rv1254	-
Rv1808	*ppe32*	Rv1566c	-	[a]Rv1522c	*mmpL12*
Rv1926c	*mpt63*	Rv1813c	-	[a]Rv1621c	*cydD*
Rv1980c	*mpt64*	Rv1886c	*fbpB*	Rv1634	-
Rv2376c	*cfp2*	Rv2301	*cut2*	Rv1733c	-
Rv2878c	*mpt53*	Rv2391	*sirA*	[a]Rv2693c	-
Rv3310	*sapM*	Rv3222c	-	[a]Rv2729c	-
		Rv3646c	*topA*	[b]Rv3162c	-
		Rv3712	-	[a]Rv3193c	-
		Rv3835	-	[a]Rv3365c	-
				[b]Rv3395A	-
				Rv3793	*embC*
				Rv3870	*eccCa1*

[a]Sec-dependent N-terminal signal
[b]Tat-dependent N-terminal signal

The EGT synthesis gene cluster consists of *EgtA*, *EgtB*, *EgtC*, *EgtD* and *EgtE* respectively. The involuntary prediction of the well-known ESX-1 substrate (Rv3616c) as a cytoplasmic protein suggests that Rv3703c may also be a secreted protein. Similarly, Jankovic et al., (2007) identified two peptides without membrane or signal peptide sequence but were shown to possess undetermined secretion directive. Therefore, further studies should be undertaken in order to establish subcellular localization of all EGT synthesis gene cluster members and determine their possible leakage or secretory pathway.

Table 2. 2: Subcellular localization prediction of 32 proteins using SecretomeP 2.0a and TBPred.

Rv no.	Gene name	Subcellular location	Rv no.	Gene name	Subcellular location
Rv0101	*nrp*	secreted	Rv1981c	*nrdF1*	secreted
[c]Rv0192	-	-	Rv1988	*erm(37)*	transmembrane
Rv0255c	*cobQ1*	membrane attached	Rv2239c	-	secreted
[c]Rv0822c	-	-	[c]Rv2768c	*ppe43*	-
Rv0824c	*desA1*	secreted	[c]Rv2922A	*acyP*	-
[c]Rv0983	*pepD*	-	Rv3009c	*gatB*	membrane attached
[c]Rv1118c	-	-	[c]Rv3103c	-	-
Rv1156	-	membrane attached	Rv3150	*nuoF*	secreted
Rv1161	*narG*	Secreted	Rv3280	*accD5*	secreted
Rv1267c	*embR*	membrane attached	[c]Rv3343c	*ppe54*	-
Rv1366	-	Secreted	Rv3410c	*guaB3*	membrane attached
Rv1447c	*zwf2*	Transmembrane	Rv3469c	*mhpE*	secreted
Rv1613	*trpA*	membrane attached	Rv3616c	*espA*	Cytoplasmic
[c]Rv1638	*uvrA*	-	Rv3703c	*egtB*	Cytoplasmic
Rv1700	-	membrane attached	Rv3734c	*tgs2*	transmembrane
Rv1916	*aceAb*	Secreted	Rv3859c	*gltB*	membrane attached

[c]SecretomeP 2.0a predicted proteins

2.4.5 ESX Proteins

The ESX-1 secretion system, that includes the secreted substrates ESAT-6 and CFP-10 virulence effectors, plays a critical role in *M. tuberculosis* pathogenicity (Simeone et al., 2009). Two ESX-1 associated proteins, EccCa1 and EspA, were identified in our study, but were not detected by Liu et al. (2011). EccCa1 is a conserved membrane component of the ESX-1 secretion system that is required for substrate exportation (Abdallah et al., 2007). EspA lacks a signal peptide sequence and was determined to be a cytoplasmic protein (Rashid et al., 2007). The gene (*espA*) encoding this protein is located within the *esx-1* gene cluster (Abdallah et al., 2007). This protein is a member of ESX-1 substrates: EsxA (ESAT-6), EsxB (CFP-10), and EspA and EspB that require each other for successful secretion (Abdallah et al., 2007; Garces et al., 2010). Upon secretion, EspA forms a disulfide bonded homodimer that is important for functional integrity of the *M. tuberculosis* cell wall. EspA is the most important determinant of

ESX-1 mediated virulence as its disruption resulted in significant attenuation of *M. tuberculosis* virulence *in vivo*. Not only does EspA "guide" known virulence factors (ESAT-6 and CFP10) through ESX-1 apparatus system proteins such as EccCa, but it also appears to regulate activity of proteins interacting with the *M. tuberculosis* cell wall (Chen et al., 2013; Garces et al., 2010). ESX-1 substrates including ESAT-6 are responsible for the translocation of mycobacteria from the phagolysosome to the cytosol (Peng and Sun, 2016). Therefore, ESX-1 system proteins like EccCa1 and EspA are potential biomarkers that can be targeted for development of therapeutic interventions (Bottai et al., 2014) and diagnostic tools.

2.4.6 PE/PPE Proteins

About 10% of the *M. tuberculosis* genome encodes two gene families, the *pe* and *ppe* genes (Cole, 1998), totaling 99 and 69 respectively, in this pathogen (Fishbein et al., 2015; Gey van Pittius et al., 2006). PE and PPE proteins are named after the Proline (Pro) and Glutamic acid (Glu), and Pro–Pro–Glu motifs near the N terminus, respectively (Cole, 1998). These proteins are involved in host-pathogen interactions and may be required for survival *in vivo* (Abdallah et al., 2006; Fishbein et al., 2015). Species within the *M. tuberculosis* complex (*MTB*C), as well as the non-tuberculosis mycobacteria such as *M. leprae*, *M. marinum*, *M. ulcerans* and *M. avium* harbor the most number of *pe/ppe* genes. Fewer *pe*/*ppe* genes are found in nonpathogenic mycobacteria (Fishbein et al., 2015). Nearly all PPE proteins are secreted through the specialized type VII (ESAT-6 like) secretion system (Abdallah et al., 2006). ESAT-6 like specialized secretion systems [ESX-1, ESX-2, ESX-3 and ESX-5] have been associated with secretion of PE and PPE proteins (Gey van Pittius et al., 2006). However, some PE and PPE proteins possess N-terminal signal peptide sequences for secretion (Forrellad et al., 2013).

In this study, three PPE proteins PPE32, PPE43 and PPE54 were identified. PE/PPE proteins are categorized into five sublineages (I to V) according to their evolutional relationship (Gey van Pittius et al., 2006). PPE32 and PPE43 belong to sublineage IV (Gey van Pittius et al., 2006) known to be secreted through ESX-5 secretion system (Fishbein et al., 2015). Interestingly, PPE32 have an N-terminal signal peptide for the Sec-dependent pathway. PPE43 and PPE54 lack a signal peptide sequence and were predicted to be extracellular proteins by SecretomeP. PPE54 belongs to the sublineage V and to the PPE_MPTR (major polymorphic tandem repeats) subfamily of PPE proteins. Sublineage V is reportedly highly expressed during *in vivo* infection (Fishbein et al., 2015), whilst PPE54 is expressed in guinea pig lungs (Kruh et al., 2010) and reported to be essential for *in vitro* growth (Tuberculist). Most PPE protein functions remain unknown, however, establishing localization of these proteins could provide some important clues to their function. Therefore, based on our findings PPE32, PPE43 and PPE54 are extracellular proteins that are either surface membrane attached or released proteins, and thus, may be directly involved in host-pathogen interaction.

2.4.7 Gene Enrichment Analysis

Gene ontology terms and functional annotation clusters with p-value < 0.05 were enriched using DAVID database. The BP term significantly enriched (p = 0.04) for five proteins (Erm(37), EmbC, Ag85B, Ag85C and Rv1634) was response to antibiotic. The CC significantly enriched term (p = 0.00) for 28 genes was extracellular region. The CC enriched genes encoded hypothetical proteins, immunogenic proteins such as MPT63 and MPT64, antigenic proteins MPT53 and CFP2, and proteins with some enzyme activity. The MF significantly enriched (p = 0.01) term of four genes (Ag85B, Ag85C, Rv0192 and LdtA) was transferase activity (Appendix A, Table S2.2).

DAVID functional annotation clustering of the gene list resulted in two significantly enriched annotation clusters. Annotation clusters provide an overview of functions associated related proteins within gene list. Cluster one contained 26 transmembrane/membrane enriched proteins (*p* -value of 0.02 to 0.04) and cluster two contained 8 proteins (NuoF, DesA1, GltB, EtgB, NarG, NrdF1, SirA and Rv0203) involved in iron binding with *p*-value of 0.01. Therefore, enrichment of GO terms and functional annotation clustering suggest that our *M. tuberculosis* phage secretome library is rich in transmembrane, surface membrane and extracellular proteins. These membrane and extracellular proteins include essential proteins required for *in vivo* growth and virulence behaviour by assisting the pathogen to acquire nutrients and navigate therapeutic interventions.

It is worth noting that only 85 of 86 proteins were analyzed by DAVID database. MT3042 was excluded from analysis and labelled as unknown protein. This could be due to fact that DAVID analysis was based on background *M. tuberculosis* H37Rv.

2.4.8 Essential Mycobacterial Proteins

M. tuberculosis secreted proteins play an important role in host-pathogen interaction and facilitate nutrient acquisition, navigate the host immune response and interfere with therapeutic intervention. Therefore, *M. tuberculosis* secretome consist of proteins essential for successful invasion and *in vivo* growth during host infection. PATRIC is the bacterial bioinformatics resource centre with curated and consistently annotated literature-based data with 7941 *M. tuberculosis* genome sequences (Wattam et al., 2014). In the search for unique *M. tuberculosis* secreted protein biomarkers, the PATRIC database was used to identify virulence factors and essential proteins.

Of 86 proteins in our list, 19 proteins were identified virulence factors (Table 2.3). The virulence factors included two cell wall (Ag85B and Ag85C), three intracellular survival (Rv1478, UvrA and AceAb), three proteins involve in modulation of host immune response (Rv1813c, MPT64 and NrdF1), seven virulence associated proteins (Nrp, SirA, Rv2693c, Rv3103c, SapM, EspA and EccCa1). The others virulence factor proteins were the PepD, a chaperone, NarG for anaerobic respiration, and PPE54 which affects the phagosome.

Table 2. 3: List of virulence factor *M. tuberculosis* secretory proteins identified using PATRIC database.

Rv no.	**Gene**	**Classification**	**References**
Rv0101	*nrp*	Virulence	(Sassetti & Rubin, 2003)
Rv0129c	*fbpC*	Cell wall	(Puech *et al.*, 2002)
Rv0983	*pepD*	Chaperone,protease	(MohamedMohaideen *et al.*, 2008)
Rv1161	*narG*	Cellular metabolism,Anaerobic respiration	PATRIC
Rv1478		Invasion,intracellular survival and replication	(Gao *et al.*, 2006)
Rv1638	*uvrA*	Intracellular survival and replication	(Houghton *et al.*, 2012)
Rv1813c		modulate host immune response	(Bretl *et al.*, 2012)
Rv1886c	*fbpB*	No evidence of virulence,cell wall	(Armitige *et al.*, 2000, Puech *et al.*, 2002)
Rv1916	*aceAb*	Intracellular survival and replication	(Muñoz-Elías & McKinney, 2005)
Rv1980c	*mpt64*	Modulate host immune response	(Kozak *et al.*, 2011)
Rv1981c	*nrdF1*	Modulate host immune response	(Kozak *et al.*, 2011)
Rv2301	*cut2*	Invasion	(Ocampo *et al.*, 2012)
Rv2391	*sirA*	Virulence factor	(Sassetti & Rubin, 2003)
Rv2693c		Virulence factor	(MacGurn & Cox, 2007)
Rv3103c		Virulence factor	(Sassetti & Rubin, 2003)
Rv3310	*sapM*	Virulence	(Chauhan *et al.*, 2013)
Rv3343c	*PPE54*	Affect phagosome	(Brodin *et al.*, 2010)
Rv3616c	*espA*	Virulence associated secretion systems, Type VII secretion	(Garces *et al.*, 2010, Chen *et al.*, 2013)
Rv3870	*eccCa1*	Virulence, Type VII secretion	(Guinn *et al.*, 2004, Champion *et al.*, 2006)

M. tuberculosis essential proteins are ideal targets for the development of diagnostic tools and new drugs because of their key role in *in vivo* bacterial survival and growth. Therefore, identifying essential *M. tuberculosis* proteins required for growth and survival in infected host

could lead to discovery of potential useful biomarkers. Using the PATRIC database, ten essential proteins were identified, five membrane proteins (PssA, MmpL5, GatB, EccCa1 and Rv1634) and five secreted proteins (Nrp, SirA, EspA, TopA and Rv3103c). Two proteins are responsible for antibiotic resistance (MmpL5 and Rv1634) and one (PssA) is a drug target (Table 2.4).

Table 2. 4: List of essential *M. tuberculosis* secretory proteins identified using PATRIC database.

Rv no.	Gene	Product	References
*Rv0101	*nrp*	Peptide synthetase Nrp	(Sassetti & Rubin, 2003)
Rv0436c	*pssA*	Phosphatidylserine synthase	PATRIC
Rv0676c	*mmpL5*	Siderophore exporter MmpL5	PATRIC
Rv1634		Probable multidrug-efflux transporter	PATRIC
*Rv2391	*sirA*	inorganic ion transport and metobolism	(Sassetti & Rubin, 2003)
Rv3009c	*gatB*	Asn/Gln amidotransferase subunit B	PATRIC
*Rv3103c		Hypothetical proline-rich protein	(Sassetti & Rubin, 2003)
**Rv3616c	*espA*	ESX-1 secretion-associated protein EspA	(Sassetti & Rubin, 2003, Fortune *et al.*, 2005)
Rv3646c	*topA*	DNA topoisomerase 1	PATRIC
**Rv3870	*eccCa1*	ESX-1 secretion system protein EccCa1	(Sassetti & Rubin, 2003, Guinn *et al.*, 2004)

*= Sassetti & Rubin, 2003

** = identified by PATRIC and Sassetti & Rubin, 2003

Generally, bacteria release their membrane proteins into the external environment as a means of membrane surface maintenance (Antelmann et al., 2001). Therefore, membrane proteins PssA, MmpL5, GatB, EccCa1 and Rv1634 will be valuable potential diagnostic biomarkers, since they could be found present in the body fluid of TB infected patients. Hence, these biomarkers have the potential to determine disease stages as it is expected that membrane proteins could be released from the bacterial cell surface at an advanced stage of growth.

There is limited literature on MT3042, the 82 amino acid protein. According to PATRIC database, MT3042 is a hypothetical protein annotated in *M. tuberculosis* CDC1551 genome

sequence. Further studies are required to investigate the prevalence and expression of MT3042 among the *M. tuberculosis* clinical strains.

2.5 CONCLUSIONS

In this study, we report an elegant and efficient *in vitro* approach, for the selective extraction of secretome genetic information from an extensively drug-resistant *M. tuberculosis* strain. This genetic information can be used to identify secreted protein ORFs and allow further secretome specific *in silico* characterization. Using this approach, more than 95% of identified ORFs were confirmed as *M. tuberculosis* secretory or surface membrane proteins using different bioinformatics tools. Furthermore, *M. tuberculosis* virulence factors and essential proteins that are a prerequisite for growth and survival during infection were identified. We identified ten essential proteins, Nrp (Rv0101), PssA (Rv0436c), MmpL5 (Rv0676c), SirA (Rv2391), GatB (Rv3009c), EspA (Rv3616c), TopA (Rv3646c), EccCa1 (Rv3870), Rv1634 and Rv3103c. The essential proteins have potential application in the development of diagnostic tools, new drugs and vaccines. Future studies should investigate the suitability of all ten essential proteins as potential *M. tuberculosis* infection specific biomarkers for the development of rapid point-of-care antigen detection test.

Our findings complement the study by Liu et al. (2011) and support the use of phage display in the study of *M. tuberculosis* clinical strains secretome for biomarker discovery. Future studies using *M. tuberculosis* phage secretome libraries should aim to identify useful protein epitopes unique to MDR and XDR strains for use in design of vaccines, drugs and diagnostic tools.

2.6 REFERENCES

Abdallah AM, Verboom T, Hannes F, *et al.* (2006) A specific secretion system mediates PPE41 transport in pathogenic mycobacteria. *Mol Microbiol* **62**: 667-679.

Andersen P, Askgaard D, Ljungqvist L, Bennedsen J & Heron I (1991) Proteins released from *Mycobacterium tuberculosis* during growth. *Infection and Immunity* **59**: 1905-1910.

Abdallah AM, van Pittius NCG, Champion PADG, Cox J, Luirink J, Vandenbroucke-Grauls CMJE, Appelmelk BJ & Bitter W (2007) Type VII secretion—mycobacteria show the way. *Nature Reviews Microbiology* **5**: 883-891.

Antelmann H, Tjalsma H, Voigt B, Ohlmeier S, Bron S, van Dijl JM & Hecker M (2001) A proteomic view on genome-based signal peptide predictions. *Genome Research* **11**: 1484-1502.

Armitige LY, Jagannath C, Wanger AR & Norris SJ (2000) Disruption of the genes encoding antigen 85A and antigen 85B of *Mycobacterium tuberculosis* H37Rv: effect on growth in culture and in macrophages. *Infection and immunity* **68**: 767-778.

Bagos PG, Tsirigos KD, Liakopoulos TD & Hamodrakas SJ (2008) Prediction of lipoprotein signal peptides in Gram-positive bacteria with a Hidden Markov Model. *Journal of Proteome Research* **7**: 5082-5093.

Bagos P, Tsirigos K, Plessas S, Liakopoulos T & Hamodrakas S (2009) Prediction of signal peptides in archaea. *Protein Engineering Design and Selection* **22**: 27-35.

Bagos PG, Nikolaou EP, Liakopoulos TD & Tsirigos KD (2010) Combined prediction of Tat and Sec signal peptides with hidden Markov models. *Bioinformatics* **26**: 2811-2817.

Bendtsen JD, Nielsen H, Widdick D, Palmer T & Brunak S (2005a) Prediction of twin-arginine signal peptides. *BMC Bioinformatics* **6**: 167.

Bendtsen J, Kiemer L, Fausboll A & Brunak S (2005b) Non-classical protein secretion in bacteria. *BMC Microbiology* **5**: 58.

Bottai D, Serafini A, Cascioferro A, Brosch R & Manganelli R (2014) Targeting type VII/ESX secretion systems for development of novel antimycobacterial drugs. *Current pharmaceutical design* **20**: 4346-4356.

Bretl DJ, He H, Demetriadou C, White MJ, Penoske RM, Salzman NH & Zahrt TC (2012) MprA and DosR coregulate a *Mycobacterium tuberculosis* virulence operon encoding Rv1813c and Rv1812c. *Infection and immunity* **80**: 3018-3033.

Brodin P, Poquet Y, Levillain F, Peguillet I, Larrouy-Maumus G, Gilleron M, Ewann F, Christophe T, Fenistein D & Jang J (2010) High content phenotypic cell-based visual screen identifies *Mycobacterium tuberculosis* acyltrehalose-containing glycolipids involved in phagosome remodeling. *PLoS Pathog* **6**: e1001100.

Champion PAD, Stanley SA, Champion MM, Brown EJ & Cox JS (2006) C-terminal signal sequence promotes virulence factor secretion in *Mycobacterium tuberculosis*. *Science* **313**: 1632-1636.

Champion PAD & Cox JS (2007) Protein secretion systems in Mycobacteria. *Cellular Microbiology* **9**: 1376-1384.

Chauhan P, Reddy PV, Singh R, Jaisinghani N, Gandotra S & Tyagi AK (2013) Secretory phosphatases deficient mutant of *Mycobacterium tuberculosis* imparts protection at the primary site of infection in guinea pigs. *PloS one* **8**: e77930.

Chen JM, Zhang M, Rybniker J, Basterra L, Dhar N, Tischler AD, Pojer F & Cole ST (2013) Phenotypic Profiling of *Mycobacterium tuberculosis* EspA Point Mutants Reveals that Blockage of ESAT-6 and CFP-10 Secretion *In Vitro* Does Not Always Correlate with Attenuation of Virulence. *Journal of Bacteriology* **195**: 5421-5430.

Cole ST, Brosch R, Parkhill J, *et al.* (1998) Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* **393**: 537-544.

de Souza GA, Leversen NA, Målen H & Wiker HG (2011) Bacterial proteins with cleaved or uncleaved signal peptides of the general secretory pathway. *Journal of Proteomics* **75**: 502-510.

Devasundaram S & Raja A (2016) Variable transcriptional adaptation between the laboratory (H37Rv) and clinical strains (S7 and S10) of *Mycobacterium tuberculosis* under hypoxia. *Infection, Genetics and Evolution* **40**: 21-28.

Fishbein S, van Wyk N, Warren RM & Sampson SL (2015) Phylogeny to function: PE/PPE protein evolution and impact on *Mycobacterium tuberculosis* pathogenicity. *Mol Microbiol* **96**: 901-916.

Forrellad MA, Klepp LI, Gioffré A, Sabio y García J, Morbidoni HR, Santangelo MdlP, Cataldi AA & Bigi F (2013) Virulence factors of the *Mycobacterium tuberculosis* complex. *Virulence* **4**: 3-66.

Fortune S, Jaeger A, Sarracino D, Chase M, Sassetti C, Sherman D, Bloom B & Rubin E (2005) Mutually dependent secretion of proteins required for mycobacterial virulence. *Proceedings of the National Academy of Sciences of the United States of America* **102**: 10676-10681.

Gagic D, Wen W, Collett MA & Rakonjac J (2013) Unique secreted–surface protein complex of *Lactobacillus rhamnosus*, identified by phage display. *MicrobiologyOpen* **2**: 1-17.

Gagic D, Ciric M, Wen WX, Ng F & Rakonjac J (2016) Exploring the Secretomes of Microbes and Microbial Communities Using Filamentous Phage Display. *Frontiers in Microbiology* **7**: 429

Gao L-Y, Pak M, Kish R, Kajihara K & Brown EJ (2006) A mycobacterial operon essential for virulence *in vivo* and invasion and intracellular persistence in macrophages. *Infection and immunity* **74**: 1757-1767.

Garces A, Atmakuri K, Chase MR, *et al.* (2010) EspA acts as a critical mediator of ESX1-dependent virulence in *Mycobacterium tuberculosis* by affecting bacterial cell wall integrity. *PLoS Pathog* **6**: e1000957.

Ge Y, ElNaggar M, Sze SK, Oh HB, Begley TP, McLafferty FW, Boshoff H & Barry Iii CE (2003) Top down characterization of secreted proteins from *Mycobacterium tuberculosis* by electron capture dissociation mass spectrometry. *Journal of the American Society for Mass Spectrometry* **14**: 253-261.

Gey van Pittius NC, Sampson SL, Lee H, Kim Y, van Helden PD & Warren RM (2006) Evolution and expansion of the *Mycobacterium tuberculosis* PE and PPE multigene families and their association with the duplication of the ESAT-6 (esx) gene cluster regions. *BMC Evolutionary Biology* **6**.

Glökler J, Schütze T & Konthur Z (2010) Automation in the high-throughput selection of random combinatorial libraries—Different approaches for select applications. *Molecules* **15**: 2478-2490.

Guinn KM, Hickey MJ, Mathur SK, Zakel KL, Grotzke JE, Lewinsohn DM, Smith S & Sherman DR (2004) Individual RD1-region genes are required for export of ESAT-6/CFP-10 and for virulence of *Mycobacterium tuberculosis*. *Mol Microbiol* **51**: 359-370.

Harth G & Horwitz MA (1999) An inhibitor of exported *Mycobacterium tuberculosis* glutamine synthetase selectively blocks the growth of pathogenic mycobacteria in axenic culture and in human monocytes: extracellular proteins as potential novel drug targets. *J Exp Med* **189**: 1425-1436.

Houghton J, Townsend C, Williams AR, Rodgers A, Rand L, Walker KB, Böttger EC, Springer B & Davis EO (2012) Important role for *Mycobacterium tuberculosis* UvrD1 in pathogenesis and persistence apart from its function in nucleotide excision repair. *Journal of Bacteriology* **194**: 2916-2923.

Huang DW, Sherman BT & Lempicki RA (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols* **4**: 44-57.

Ioerger TR, Feng Y, Ganesula K, Chen X, Dobos KM, Fortune S, Jacobs WR, Mizrahi V, Parish T & Rubin E (2010) Variation among genome sequences of H37Rv strains of *Mycobacterium tuberculosis* from multiple laboratories. *Journal of Bacteriology* **192**: 3645-3653.

Jacobsson K & Frykberg L (2001) Shotgun phage display cloning. *Comb Chem High Throughput Screen* **4**: 135 - 143.

Jacobsson K, Rosander A, Bjerketorp J & Frykberg L (2003) Shotgun Phage Display - Selection for Bacterial Receptins or other Exported Proteins. *Biol Proced Online* **5**: 123 - 135.

Jankovic D, Collett MA, Lubbers MW & Rakonjac J (2007) Direct selection and phage display of a Gram-positive secretome. *Genome Biology* **8**: R266.

Karlström Å, Jacobsson K, Flock M, Flock J-I & Guss B (2004) Identification of a novel collagen-like protein, SclC, in *Streptococcus equi* using signal sequence phage display. *Veterinary Microbiology* **104**: 179-188.

Kozak RA, Alexander DC, Liao R, Sherman DR & Behr MA (2011) Region of difference 2 contributes to virulence of *Mycobacterium tuberculosis*. *Infection and Immunity* **79**: 59-66.

Krogh A, Larsson B, Von Heijne G & Sonnhammer EL (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *Journal of Molecular Biology* **305**: 567-580.

Kruh NA, Troudt J, Izzo A, Prenni J & Dobos KM (2010) Portrait of a pathogen: the *Mycobacterium tuberculosis* proteome *in vivo*. *PLoS One* **5**: e13938.

Lew JM, Kapopoulou A, Jones LM & Cole ST (2011) TubercuList - 10 years after. *Tuberculosis* **91**: 1-7.

Ligon LS, Hayden JD & Braunstein M (2012) The ins and outs of *Mycobacterium tuberculosis* protein export. *Tuberculosis* **92**: 121-132.

Liu S, Han W, Sun C, Lei L & Feng X (2013) Identification of Two New Virulence Factors of *Mycobacterium Tuberculosis* that Induce Multifunctional CD4 T-cell Responses. *J Mycobac Dis S* **6**: 2161-1068.

Liu S, Han W, Sun C, Lei L, Feng X, Yan S, Diao Y, Gao Y, Zhao HL & Liu Q (2011) Subtractive screening with the *Mycobacterium tuberculosis* surface protein phage display library. *Tuberculosis* **91**: 579-586.

MacGurn JA & Cox JS (2007) A genetic screen for *Mycobacterium tuberculosis* mutants defective for phagosome maturation arrest identifies components of the ESX-1 secretion system. *Infection and immunity* **75**: 2668-2678.

Mehaffy C, Hess A, Prenni JE, Mathema B, Kreiswirth B & Dobos KM (2010) Descriptive proteomic analysis shows protein variability between closely related clinical isolates of *Mycobacterium tuberculosis*. *Proteomics* **10**: 1966-1984.

MohamedMohaideen NN, Palaninathan SK, Morin PM, Williams BJ, Braunstein M, Tichy SE, Locker J, Russell DH, Jacobs Jr WR & Sacchettini JC (2008) Structure and Function of the Virulence-Associated High-Temperature Requirement A of *Mycobacterium tuberculosis*. *Biochemistry* **47**: 6092-6102.

Muñoz-Elías EJ & McKinney JD (2005) *Mycobacterium tuberculosis* isocitrate lyases 1 and 2 are jointly required for *in vivo* growth and virulence. *Nature Medicine* **11**: 638-644.

Naidoo C & Pillay M (2014) Increased *in vitro* fitness of multi-and extensively drug-resistant F15/LAM4/KZN strains of *Mycobacterium tuberculosis*. *Clinical Microbiology and Infection* **20**: O361-O369.

Ocampo M, Rodriguez D, Curtidor H, Vanegas M, Patarroyo M & Patarroyo M (2012) Peptides derived from *Mycobacterium tuberculosis* Rv2301 protein are involved in invasion to human epithelial cells and macrophages. *Amino Acids* **42**: 2067-2077.

Palanisamy GS, DuTeau N, Eisenach KD, Cave DM, Theus SA, Kreiswirth BN, Basaraba RJ & Orme IM (2009) Clinical strains of *Mycobacterium tuberculosis* display a wide range of virulence in guinea pigs. *Tuberculosis* **89**: 203-209.

Peng X & Sun J (2016) Mechanism of ESAT-6 membrane interaction and its roles in pathogenesis of *Mycobacterium tuberculosis*. *Toxicon* **116**: 29-34.

Peters J, Calder B, Gonnelli G, Degroeve S, Rajaonarifara E, Mulder N, Soares NC, Martens L & Blackburn JM (2016) Identification of quantitative proteomic differences between *Mycobacterium tuberculosis* lineages with altered virulence. *Frontiers in Microbiology* **7**.

Petersen TN, Brunak S, von Heijne G & Nielsen H (2011) SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nature Methods* **8**: 785-786.

Pillay M & Sturm AW (2007) Evolution of the extensively drug-resistant F15/LAM4/KZN strain of *Mycobacterium tuberculosis* in KwaZulu-Natal, South Africa. *Clinical Infectious Diseases* **45**: 1409-1414.

Poquet I, Ehrlich SD & Gruss A (1998) An export-specific reporter designed for gram-positive bacteria: application to *Lactococcus lactis*. *Journal of bacteriology* **180**: 1904-1912.

Rahman O, Cummings SP, Harrington DJ & Sutcliffe IC (2008) Methods for the bioinformatic identification of bacterial lipoproteins encoded in the genomes of Gram-positive bacteria. *World Journal of Microbiology and Biotechnology* **24**: 2377-2382.

Rashid M, Saha S & Raghava GP (2007) Support Vector Machine-based method for predicting subcellular localization of mycobacterial proteins using evolutionary information and motifs. *BMC Bioinformatics* **8**: 1.

Rosander A, Frykberg L, Ausmees N & Müller P (2003) Identification of Extracytoplasmic Proteins in *Bradyrhizobium japonicum* Using Phage Display. *Molecular Plant-Microbe Interactions* **16**: 727-737.

Rosander A, Guss B, Frykberg L, Björkman C, Näslund K & Pringle M (2011) Identification of immunogenic proteins in *Treponema phagedenis*-like strain V1 from digital dermatitis lesions by phage display. *Veterinary Microbiology* **153**: 315-322.

Sassetti CM & Rubin EJ (2003) Genetic requirements for mycobacterial survival during infection. *Proceedings of the National Academy of Sciences* **100**: 12989-12994.

Simeone R, Bottai D & Brosch R (2009) ESX/type VII secretion systems and their role in host–pathogen interaction. *Current Opinion in Microbiology* **12**: 4-10.

Smith G (1985) Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface. *Science* **228**: 1315 - 1317.

Su M, Venkatesh T & Bodmer R (1998) Large-and small-scale preparation of bacteriophage lambda lysate and DNA. *BioTechniques* **25**: 44-46.

Suzek BE, Wang Y, Huang H, McGarvey PB, Wu CH & Consortium tU (2015) UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* **31**: 926-932.

Thakur P, Gantasala NP, Choudhary E, Singh N, Abdin MZ & Agarwal N (2016) The preprotein translocase YidC controls respiratory metabolism in *Mycobacterium tuberculosis*. *Sci Rep* **6**: 24998.

van Soolingen D, de Haas PE, Hermans PW & van Embden JD (1994) DNA fingerprinting of *Mycobacterium tuberculosis*. *Methods Enzymol* **235**: 196-205.

Wall T, Roos S, Jacobsson K, Rosander A & Jonsson H (2003) Phage display reveals 52 novel extracellular and transmembrane proteins from *Lactobacillus reuteri* DSM 20016T. *Microbiology* **149**: 3493-3505.

Wallis RS, Kim P, Cole S, Hanna D, Andrade BB, Maeurer M, Schito M & Zumla A (2013) Tuberculosis biomarkers discovery: developments, needs and challenges. *Lancet Infect Dis* **13**: 362-372.

Wattam AR, Gabbard JL, Shukla M & Sobral BW (2014) Comparative genomic analysis at the PATRIC, a bioinformatic resource center. *Host-Bacteria Interactions: Methods and Protocols* 287-308.

WHO (2016) Global Tuberculosis Report 2016. *World Health Organization.*

Xu Y, Yang E, Huang Q, Ni W, Kong C, Liu G, Li G, Su H & Wang H (2015) PPE57 induces activation of macrophages and drives Th1-type immune responses through TLR2. *Journal of Molecular Medicine* **93**: 645-662.

Zheng J, Ren X, Wei C, Yang J, Hu Y, Liu L, Xu X, Wang J & Jin Q (2013) Analysis of the Secretome and Identification of Novel Constituents from Culture Filtrate of Bacillus Calmette-Guérin Using High-resolution Mass Spectrometry. *Molecular & Cellular Proteomics* **12**: 2081-2095.

CHAPTER 3: Immunoscreening of the *M. tuberculosis* F15/LAM4/KZN secretome library against TB patients' sera identifies unique active- and latent-TB specific biomarkers

Thamsanqa E. Chiliza[1], Manormoney Pillay[2], Kogieleum Naidoo[3] and Balakrishna Pillay[1*]

[1]Discipline of Microbiology, School of Life Sciences, College of Agriculture, Engineering and Science, University of KwaZulu-Natal, Westville Campus, South Africa;

[2]Medical Microbiology, School of Laboratory Medicine and Medical Sciences, College of Health Sciences, University of KwaZulu-Natal, Durban, South Africa;

[3]Centre for the AIDS Programme of Research in South Africa (CAPRISA), University of KwaZulu-Natal, Durban, South Africa

Under review for publication in Tuberculosis journal

3.1 ABSTRACT

Tuberculosis (TB) protein biomarkers are urgently needed for the development of point-of-care diagnostics, new drugs and vaccines. *Mycobacterium tuberculosis* (*Mtb*) extracellular and secreted proteins play an important role in host-pathogen interactions. Antibodies produced against *Mtb* proteins before the onset of clinical symptoms can be used in proteomic studies to identify their target proteins. In this study, *Mtb* F15/LAM4/KZN strain phage secretome library was screened against immobilized polyclonal sera from active-TB patients (*n*=20), individuals with latent-TB (*n*=15) infection (LTBI) and TB negative (*n*=20) individuals to select and identify proteins recognized by patients' antibodies. DNA sequence analysis from randomly selected latent-TB and active-TB specific phage clones revealed 118 and 96 ORFs, respectively. Proteins essential for growth, virulence and metabolic pathways were identified using different TB databases. The identified active-TB specific biomarkers included five proteins, namely, TrpG, Alr, TreY, BfrA and EspR, with no human homologs, whilst latent-TB specific biomarkers included NarG, PonA1, PonA2 and HspR. Future studies will assess potential applications of identified protein biomarkers in the development of new drugs, vaccines and diagnostic tools with the ability to discriminate LTBI from active-TB.

KEYWORDS: tuberculosis, phage display, biomarkers, active-TB, latent-TB, diagnostics

3.2 INTRODUCTION

Nearly 2 million deaths and more than 10.4 million cases of tuberculosis (TB) were reported globally in 2016 (WHO, 2017). Approximately half of the people exposed to TB develop clinical symptoms within a year (Rustad *et al.*, 2009). Individuals with latent-TB infection (LTBI) have a 10% chance of developing active-TB (ATB) during their lifetime (Corbett *et al.*, 2003). Approximately, one-third of the world's population is latently infected with TB (WHO, 2017). Co-infection with HIV increases the risk of progression to ATB by up to 10% per year, making TB the leading cause of death from an opportunistic infection in HIV/AIDS patients (Corbett *et al.*, 2003, McShane, 2005). Therefore, LTBI plays a critical role in the current TB pandemic and is a constant source of new cases of TB disease especially among HIV/AIDS patients (Rustad *et al.*, 2009).

During infection, *Mycobacterium tuberculosis* secretes proteins into the surrounding host environment, playing an important role in host-pathogen interactions. Antibodies against mycobacterial proteins (antigens) are produced before the onset of clinical symptoms (Laal *et al.*, 1997, Lyashchenko *et al.*, 1998, Samanich *et al.*, 2001, Kunnath-Velayudhan *et al.*, 2010). These antibodies target all immunologically relevant pathogen specific antigens (Kunnath-Velayudhan *et al.*, 2012) and may shed some light on the *M. tuberculosis* protein expression profiles during infection. However, the antibody response differs widely among TB patients (Khan *et al.*, 2011) and this may be due to the disease progression stage during sampling, the relative specific antigen dominancy and the *M. tuberculosis* strain antigen composition (Kunnath-Velayudhan *et al.*, 2011, Senoputra *et al.*, 2015). Nevertheless, these antibodies can be used in proteomic studies to identify their cognate *M. tuberculosis* antigens or epitopes (Wallis *et al.*, 2009). This could lead to the discovery of disease specific protein biomarkers that can be used in the design of new drugs, vaccines and diagnostic or prognostic tools. In turn, antibodies produced against these epitopes can be used in order to detect them in clinical

specimens. Consequently, identification of LTBI specific biomarkers may result in the development of improved and accurate point-of-care diagnostic tools that can be used to monitor disease progression especially in HIV/AIDS patients.

Phage display is a powerful *in vitro* selection technology that can be used to specifically extract proteins with novel and desired properties from large protein libraries (Konthur *et al.*, 2003). In the present study, a *M. tuberculosis* F15/LAM4/KZN phage secretome library (Chiliza *et al.*, 2017) was interrogated by immunoscreening against patients' sera to identify immunogenic LTBI and ATB specific proteins recognized by their antibodies. Sera from LTBI and ATB patients were used for isolation by affinity binding of phage particles displaying TB disease state-specific peptides. Therefore, the current study was aimed at the identification of novel, easily accessible disease state-specific *M. tuberculosis* secreted biomarkers that can be used to distinguish ATB from LTBI.

3.3 MATERIALS AND METHODS

3.3.1 Patient Recruitment and Specimen Collection

The study was approved by the University of KwaZulu-Natal Biomedical Research Ethics Committee (Ref. BE236/13). Participants were considered for inclusion if they were ≥18 years of age and willing to undergo an HIV test. Fifty-five study participants were recruited with informed consent from the Centre for the AIDS Programme of Research in South Africa (CAPRISA) eThekwini Clinic, KwaZulu-Natal. All participants included in the study were HIV negative and all individuals testing HIV positive were excluded. The GeneXpert/*MTB*/RIF assay was used as a TB confirmatory test and GeneXpert positive participants were included in our active-TB (ATB) group (n=20). GeneXpert negative participants were subjected to tuberculin skin test (TST) in order to detect LTBI. Subjects with an induration of ≥10mm

(TST+) were classified as latent-TB infected (*n*=15), and TST negative (*n*=20) as healthy non-TB subjects. Peripheral blood (5 mL) was collected using the closed blood collection method into BD Vacutainer serum tubes (red top, 5mL tube). The specimen was immediately mixed by inverting 6 to 10 times and stored at 4°C before processing. After centrifugation at 1500xg for 10 min, serum was stored at -80°C until use.

3.3.2 Panning of *M. tuberculosis* F15/LAM4/KZN Phage Secretome Library

A phage library displaying *M. tuberculosis* F15/LAM4/KZN secretome was constructed previously (Chiliza *et al.*, 2017) using the pDJ01 phagemid vector (Jankovic *et al.*, 2007). Pooled serum samples containing patients' polyclonal antibodies were diluted 1:100 in PBS (pH 7.2) and used to coat the ELISA plate wells at 4ºC overnight. The secretome phage library was exposed to the immobilized serum samples. The concentration of phage particles at the beginning of each of 3 rounds for both latent-TB and active-TB panning was 1×10^{11} pfu. In order to eliminate non-specific immunogenic peptides, the library was pre-incubated with immobilized sera from TB-free participants. Selective targeting of disease state specific phage displayed protein peptides was achieved by exposing the TB-free pre-absorbed library to latent-TB sera before the selection on active-TB sera to identify active-TB specific proteins. Latent-TB proteins were selected by exposing the TB-free pre-absorbed library first to active-TB sera and then to latent-TB sera. The binding phages were eluted with 0.1M TEA (triethanolamine) for 10 minutes at room temperature. The eluate was neutralized with 1M Tris buffer (pH 7), and amplified by infecting exponentially growing *E. coli* TG1 cells. Amplified phages were subjected to two further rounds of selective screening to enrich for clones that are specifically recognized by the patients' sera. After three successive rounds of interrogation against TB patients' sera (Fig. 3.1), affinity binding and immunogenic latent-TB and active-TB specific

M. tuberculosis phage displayed protein peptides were enriched. For latent-TB selection, the amount of recovered phage particles for the three successive rounds was 1.8×10^3, 3.2×10^4 and 2.4×10^5 pfu respectively. The active-TB recovered phage particles for three successive rounds was 2.2×10^3, 5.2×10^4 and 3.9×10^5 pfu respectively. Enrichment was calculated as the ratio of number of eluted phages in the third round of panning relative to the first round of panning.

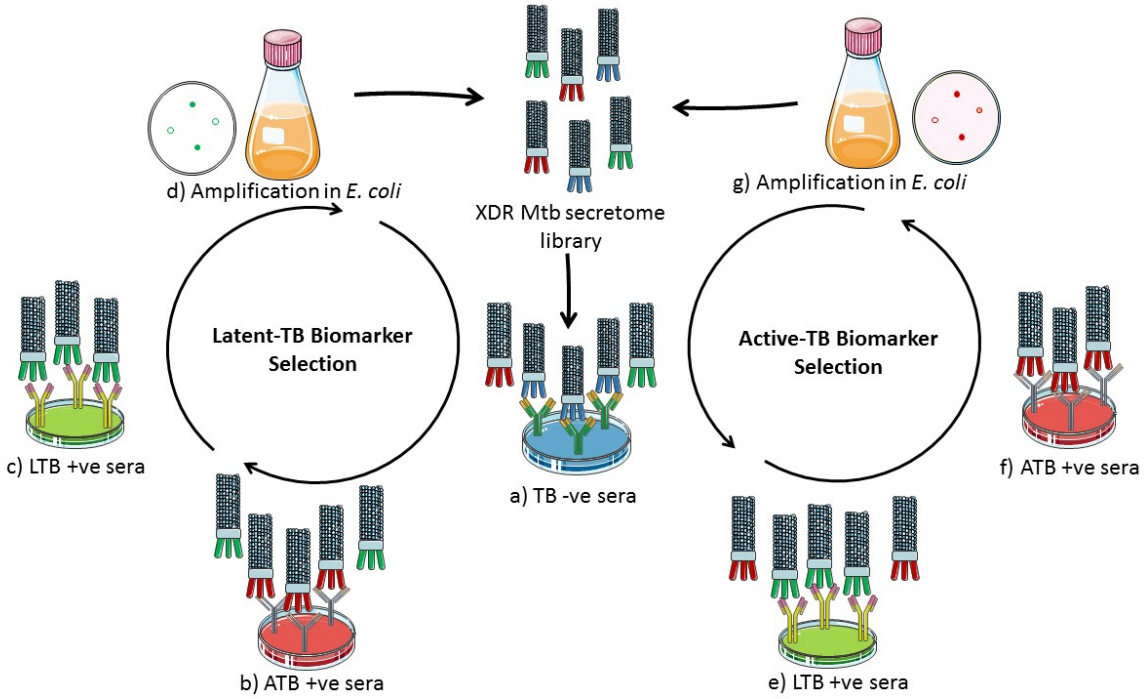

Fig. 3. 1: Overview of steps involved in *Mtb* phage secretome library immunoscreening against clinical serum samples. For selection of latent-TB specific biomarkers, the library was initially exposed to: a) control TB –ve sera from healthy participants to pre-absorb and remove phage particles displaying "ubiquitous" protein peptides. b) The unbound phages were transferred to a well coated with ATB +ve sera from active-TB participants to remove protein peptides recognised by patients' antibodies. c) Final selection was performed against LTB +ve sera from latent-TB infected participants. Selection of active-TB specific biomarkers began with a) TB –ve sera, followed by screening against e) LTB +ve sera from latent-TB infected patients and final screening f) against ATB +ve sera from active-TB patients.

3.3.3 DNA Sequence Analysis

The inserts of randomly selected clones were sequenced using the primer pDJ01R02 (5'-CCGGAAACGTCACCAATGAA-3') and pDJF03 (5'-ATGTTGCTGTTGATTCTTCA-3'). The DNA sequences were analysed using the CLCBio Workbench 2.0. BLAST analysis of the nucleotide sequences against the *M. tuberculosis* database was performed in order to retrieve encoded open reading frames (ORFs) and complete protein sequences.

3.3.4 Functional Categories and Gene Ontologies

The ORFs functional category information for annotated genes was retrieved from the TubercuList database (http://tuberculist.epfl.ch/) (Lew *et al.*, 2011). Protein sequences of unknown genes that are not documented on TubercuList database were categorized as unknown functional category. For further assignment of possible function, gene ontology (GO) features of identified latent-TB and active-TB proteins were obtained using Universal Protein Resource Knowledgebase (UniProtKB) database (http://www.uniprot.org/) (The_UniProt_Consortium, 2017). The Web Gene Ontology Annotation Plotting (WEGO) (http://wego.genomics.org.cn/cgi-bin/wego/index.pl) was used to analyze the GO categories (Ye *et al.*, 2006).

3.3.5 Metabolic Pathways and Specialty Proteins Identification

The Pathosystems Resource Integration Center (PATRIC) [https://www.patricbrc.org] and KEGG (Kyoto Encyclopedia of Genes and Genomes) [http://www.genome.ad.jp/kegg] databases were used to identify metabolic pathway proteins (Kanehisa *et al.*, 2017). PATRIC specialty protein list was used to identify proteins essential for *M. tuberculosis* growth, virulence and survival during host infection (Wattam *et al.*, 2017).

3.4 RESULTS AND DISCUSSION

3.4.1 Identification of Disease State Specific Proteins

After three successive rounds of immunoscreening against sera from LTBI (n=15) and ATB patients (n=20), affinity binding and immunogenic latent-TB and active-TB specific *M. tuberculosis* phage displayed protein peptides were enriched by 133-fold and 177-fold, respectively. At the last round of panning, clones were selected randomly, and 157 and 127 phage DNA were sequenced for LTBI and ATB, respectively. DNA sequence analysis revealed 125 LTBI and 100 ATB distinct sequences. The BLAST search against *M. tuberculosis* databases identified 118 LTBI and 96 ATB encoded ORFs (Appendix A, Table S3.1 and S3.2). The ORFs were allocated functional categories according to the TubercuList database annotation (Lew *et al.*, 2011). The LTBI and ATB selected proteins were distributed to ten different TubercuList functional categories (Fig. 3.2).

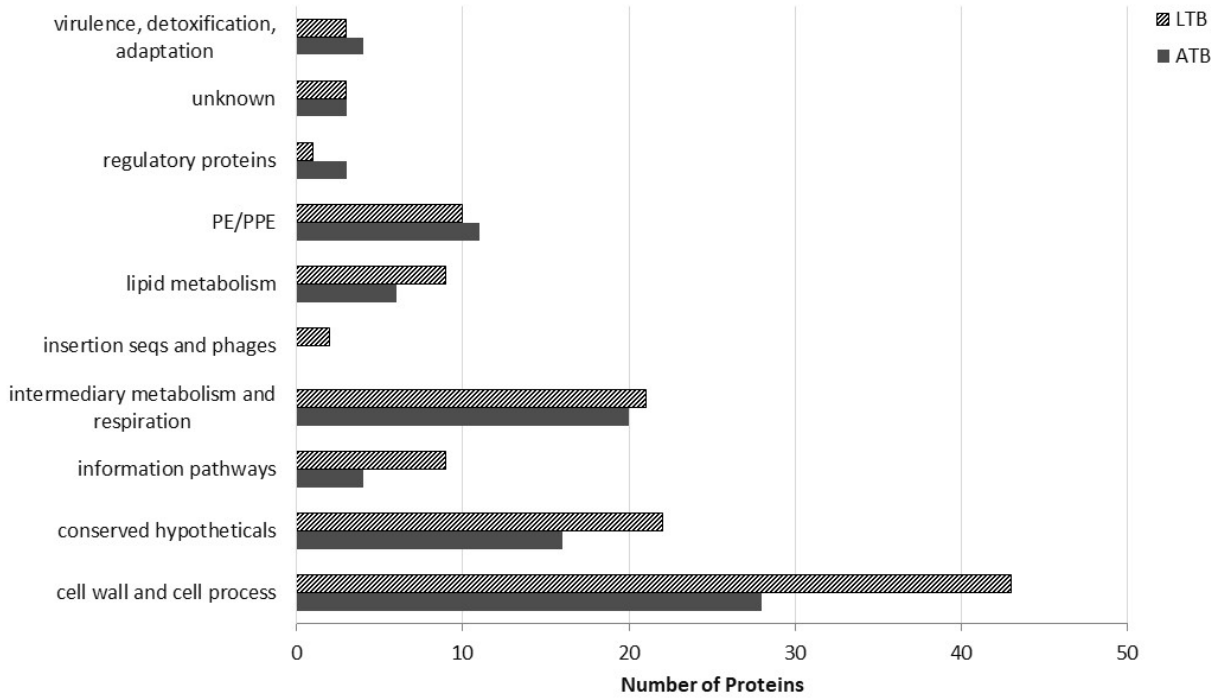

Fig. 3. 2: TubercuList functional categories of proteins identified during selection for active-TB and latent-TB biomarkers.

Table 3. 1: List and description of 23 proteins common to ATB and LTBI.

Rv	Name	Description
MT1330.1	MT1330.1	hypothetical protein
Rv0203	-	secreted protein with unknown function
Rv0327	Cyp135A1	Cytochromes P450, a heme-thiolate monooxygenase
Rv0630c	RecB	exonuclease V (beta chain) RecB
Rv1009	RpfB	resuscitation-promoting factor RpfB.
Rv1029	KdpA	potassium-transporting ATPase a chain KdpA
Rv1133c	MetE	Cobalamin-independent methionine synthase
Rv1157c	-	Conserved secreted ala-, pro-rich protein
Rv1268	-	hypothetical outer membrane protein
Rv1269c	-	Conserved probable secreted protein
Rv1271c	-	Conserved hypothetical secreted protein
Rv1357c	-	Conserved hypothetical protein
Rv1703	-	Probable catechol-O-methyltransferase.
Rv1806	PE20	PE family protein PE20
Rv1886c	FbpB	Secreted antigen 85-B FbpB (85B)
Rv1926c	Mpt63	Immunogenic 16 kDa protein Mpt63
Rv1980c	Mpt64	Immunogenic protein Mpt64
Rv2152c	-	UDP-N-acetylmuramate-alanine ligase MurC
Rv2414c	-	Conserved hypothetical protein
Rv3157	NuoM	NADH dehydrogenase I (chain M) NUOK
Rv3218	-	Conserved protein. Function unknown
Rv3318	SdhA	succinate dehydrogenase flavoprotein subunit sdhA
Rv3508	PE_PGRS54	PE-PGRS family protein PE_PGRS54

The majority of LTBI (43) and ATB (28) proteins were associated with the cell wall and cell processes. These were mainly surface membrane and transmembrane proteins involved in transportation of molecules across the cell wall and facilitate attachment to host target cells. The second most prevalent group of proteins were the conserved hypothetical proteins of unknown functions of which 22 were unique to the LTBI and 16 were unique to the ATB.

Twenty-three proteins were observed to be common between the LTBI and ATB lists (Table 3.1). These included hypothetical proteins and immunogenic proteins such as FbpB (Rv1886c), Mpt63 (Rv1926c), Mpt64 (Rv1980c) and two PE/PPE family proteins, PE20 (Rv1806) and PE_PGRS54 (Rv3508).

3.4.2 Gene Ontology Analysis

Gene Ontology functional categories of 60 ATB and 80 LTBI proteins GO annotation were retrieved from UniProtKB database (Appendix A, Table S3.3 and S3.4). WEGO analysis assigned the proteins to three functional categories; Cellular Component (CC), Molecular Function (MF), and Biological Process (BP), with a total of 31 GO terms (Fig. 3.3). Proteins involved in multiple activities or processes are assigned by WEGO to one or more associated GO terms and categories. Notably, 15 of 95 LTBI and 12 of 72 ATB proteins with either unknown function/s or unrecognized by Gene Ontology enrichment were not assigned to any functional categories.

Amongst the ATB proteins, the main CC terms comprised cell (42 proteins), cell part (41 proteins) and membrane (29 proteins). In the BP ontology, top terms included cellular process (27 proteins), metabolic process (26 proteins) and response to stimulus (11 proteins). The majority of MF category terms involved catalytic activity (31 proteins) and binding (24 proteins). The enzymes involved in catalytic activity were classified as oxidoreductase, transferase, hydrolase, lyase, ligase and isomerase.

For the LTBI specific proteins, the main CC terms included cell (60 proteins), cell part (59 proteins) and membrane (49 proteins). The main BP ontology terms comprised metabolic process (37 proteins), cellular process (31 proteins) and growth (13 proteins). The main MF

terms included catalytic activity (43 proteins) and binding (26 proteins). The enzymes involved in catalytic activity were classified as oxidoreductase, transferase, hydrolase, lyase and ligase.

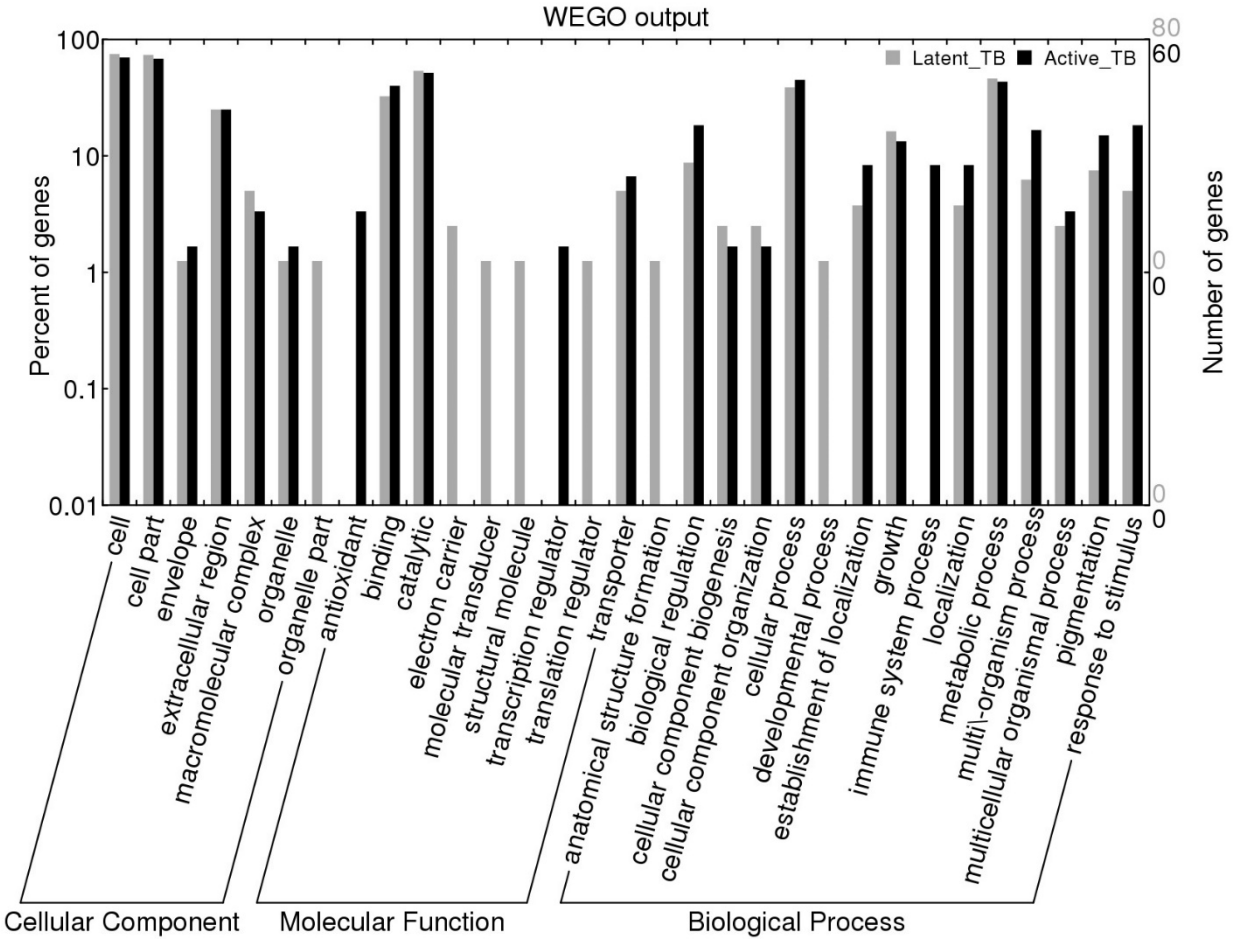

Fig. 3. 3: Histogram of gene ontology classifications of ATB and LTBI specific *M. tuberculosis* proteins recognised by sera of TB patients. The WEGO plot shows the three main GO categories: cellular component, molecular function and biological process. The right y-axis indicates the number of genes in a category. The left y-axis indicates the percentage of a specific category of genes in that main category. The WEGO plotting parameters were transformed into a log10 scale to represent both highly and lowly-enriched GO functions.

3.4.3 Functional Categories Unique to ATB and LTBI

3.4.3.1 ATB functional categories

The 3 GO terms unique to ATB included 2 MF terms (antioxidant and transcriptional regulator) and 1 BP term (immune system process). The antioxidant associated proteins were Rv2878c (MPT53) and Rv2429 (AhpD). MPT53 is a soluble secreted antigen reportedly recognized by TB patients' humoral response (Malen *et al.*, 2008) and AhpD is a reductase that is involved in oxidative stress response (Koshkin *et al.*, 2003). The transcription associated protein, Rv3058c is involved in transcriptional regulatory processes (TubercuList). The immune system process proteins comprised Rv2941 (fadD28), Rv3343c (PPE54), Rv1860 (Apa) and Rv1818c (PE_PGRS33) which are involved in negative regulation or modulation of host immune response. FadD28 is involved in fatty acid biosynthesis and together with PPE54, is involved in host phagosome maturation arrest (Brodin *et al.*, 2010). Alanine and proline rich secreted protein, Apa facilitates the bacterial attachment to host cells (Pitarque *et al.*, 2005). The immunogenic PE-PGRS family protein, PE_PGRS33 facilitates cell surface interactions among mycobacteria and interactions of bacteria with macrophages (Cohen *et al.*, 2014, Palucci *et al.*, 2016).

3.4.3.2 LTBI functional categories

The 7 GO terms unique to LTBI included 1 CC term (organelle part), 4 MF terms (electron carrier, molecular transducer, structural molecule and translation regulator), and 2 BP terms (anatomical structure formation and developmental process). The organelle part protein was identified as Rv3459c (RpsK) that is essential for selection of protein biosynthesis (UniProtKB). The electron carrier activity proteins were Rv1161 (NarG) and Rv0688 encoding the putative ferredoxin reductase. NarG catalyses the consumption of nitrate in *M. tuberculosis*

(Iona *et al.*, 2016) while the putative ferredoxin reductase is involved in the transfer of electrons in various metabolic reactions (TubercuList). The molecular transducer protein (Rv3365c) is a conserved hypothetical protein proposed to be involved in cell-associated regulatory functions (Mazandu *et al.*, 2012). The translation regulator activity protein Rv2839c encoding InfB is an essential component for the initiation of protein synthesis (TubercuList). The developmental process protein (Rv0050) also known as PonA1 is involved in cell wall formation by synthesis of cross-linked peptidoglycan from the lipid intermediates (Kieser *et al.*, 2015).

3.4.4 Differentially Expressed Functional Proteins

3.4.4.1 ATB differentially expressed proteins

The ATB differentially expressed proteins included 3 isomerases (Alr, TreY and TopA), the extracellular protein EspR (Rv3849) and the organelle proteins Rv1526c and PE_PGRS47 (Rv2741). Isomerase Alr is involved in peptidoglycan cross-linking through conversion of L-alanine to D-alanine which is a necessary precursor for peptidoglycan biosynthesis (Strych *et al.*, 2001). TreY is involved in starch and sucrose metabolism, specifically, trehalose biosynthesis. TopA is involved in DNA transcription and reportedly required for adaptation and survival of *M. tuberculosis* under stressful conditions. EspR is associated with cell wall functions and pathogenesis through regulation of multiple genes, including the *espACD* operon, which is a key ESX-1 component (Bitter *et al.*, 2009). Rv1526c is a glycosyltransferase and is thought to be involved in cellular metabolism (TubercuList). The PE-PGRS family protein, PE_PGRS47 is required for growth and survival of *M. tuberculosis* during chronic TB and is also involved in evasion of innate and adaptive immunity (Saini *et al.*, 2016).

3.4.4.2 LTBI differentially expressed proteins

The LTBI differentially expressed proteins comprised SahH (Rv3779) associated with growth, the host intracellular part protein, SapM (Rv3310) and three cellular component's organization or biogenesis proteins DnaB (Rv0058), EmbC (Rv3793), and FadD17 (Rv3506). The S-adenosyl-L-homocysteine hydrolase (SahH) is involved in regulation of the intracellular concentration of adenosylhomocysteine and is essential for bacterial growth (Cole *et al.*, 1998, Sassetti *et al.*, 2003a). The acid phosphatase (SapM) plays an important role in blocking phagosome-lysosome fusion, thus participating in the intracellular survival of the pathogen (Puri *et al.*, 2013). The Arabinosyltransferase C (EmbC) is an integral membrane protein involved in the biosynthesis of the mycobacterial cell wall arabinan required for resistance to anti-TB drug ethambutol, hence, essential for *M. tuberculosis* growth (Sassetti *et al.*, 2003a, Kieser *et al.*, 2014) and survival. EmbC is a potential drug target and can be considered for use in TB diagnosis. FadD17 is required for salvaging cholesterol. The *M. tuberculosis* genome encodes 35 *fadD* genes which are suspected to be associated with fatty acid biosynthesis (TubercuList). When cholesterol is used by *M. tuberculosis* as a carbon source, up-regulated genes include *fadD17, fadD18, fadD19* and *fadD3* (Wipperman *et al.*, 2014).

3.4.5 Metabolic Pathway Proteins

During LTBI and ATB, the expression of certain metabolic pathway genes is upregulated or downregulated (Gopinath *et al.*, 2015). The metabolic pathway gene encoded proteins (MPPs) are involved in catabolism and biosynthesis of molecules essential for growth of *M. tuberculosis.* Therefore, these are attractive drug targets since inhibition of their functions may kill the pathogen. The *M. tuberculosis* PATRIC and KEGG metabolic pathway databases have 840 and 605 MPPs, respectively. The individual databases have some overlapping MPPs

between databases and also database specific MPPs (Fig. 3.4). The databases were used to identify 14 ATB and 21 LTBI specific MPPs involved in *M. tuberculosis* metabolic processes (Table 3.2).

Analysis of the generated protein lists revealed that different *M. tuberculosis* metabolic pathways were enriched. The pathways unique to ATB were amino acid biosynthesis, carbohydrate metabolism and iron metabolism, while cofactor and coenzyme biosynthesis, nitrogen metabolism and carbonic metabolism were unique LTBI enriched pathways. Among the complete list of metabolic pathways (Table 3.2), cell wall biosynthesis and lipid metabolism pathways were common to both ATB and LTBI.

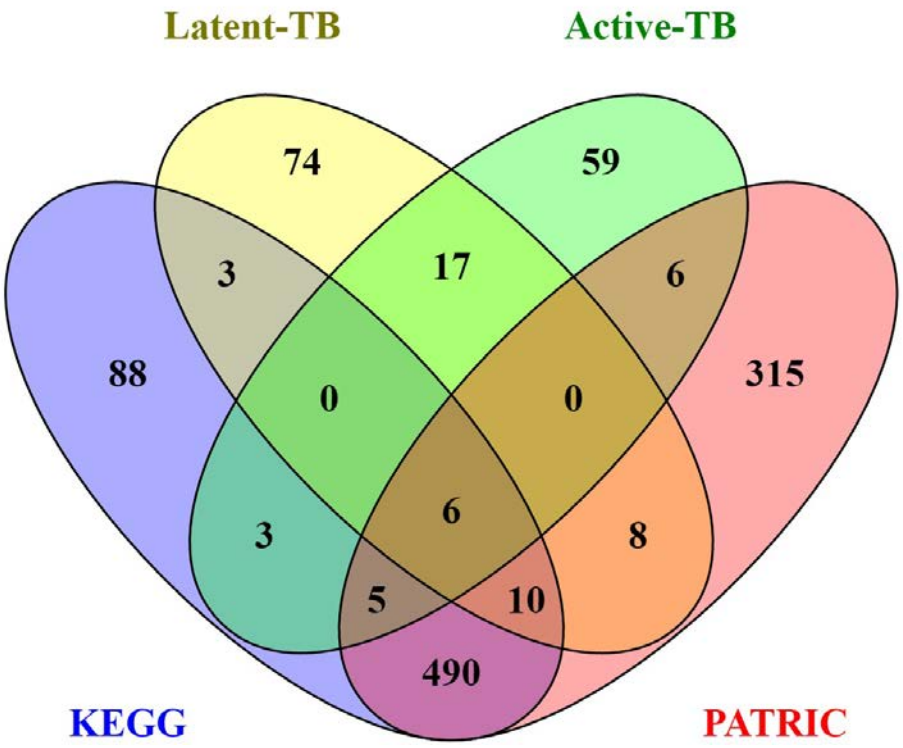

Fig. 3. 4: Venn diagram showing metabolic pathway proteins from KEGG and PATRIC databases that overlap with the selected ATB and LTBI proteins. There were 14 ATB and 21 LTBI specific metabolic pathway proteins that were identified.

Table 3. 2: List of metabolic pathway proteins that were uniquely enriched during selection in identified active- and latent -TB immunoscreening.

Pathway	Rv	Name	Product	Function
Active-TB specific MPPs				
Amino acid biosynthesis	Rv0013	*trpG*	Anthranilate synthase, amidotransferase component	Essential for the establishment, maintenance of infection and survival of *M. tuberculosis* (Zhang *et al.*, 2013a), no human homolog.
	Rv2537c	*aroD*	3-dehydroquinate dehydratase II	Required for protein synthesis
Carbohydrates Metabolism	Rv1563c	*treY*	maltooligosyl trehalose synthase	Key component of a variety of glycolipids required for growth and virulence of *M. tuberculosis* (De Smet *et al.*, 2000, Murphy *et al.*, 2005, Kalscheuer *et al.*, 2014), no human homolog.
	Rv3264c	*manB*	D-alpha-D-mannose-1-phosphate guanylyltransferase	
Lipid Metabolism	Rv1130	*metE/ prpD*	5-methyltetrahydropteroyltriglutamate--homocysteine methyltransferase	Expressed during infection (Savvi *et al.*, 2008) and required for intracellular survival (Mattow *et al.*, 2006).
	Rv2252		diacylglycerol kinase	
Cell wall biosynthesis	Rv3330	dacB1	D-alanyl-D-alanine carboxypeptidase penicillin-binding protein 5/6	Control of cell elongation and septum development via the regulation of the peptidoglycan cross-linking (Sauvage *et al.*, 2008, Prigozhin *et al.*, 2014).
	Rv3423c	alr	Alanine racemase	No human homolog, necessary precursor for peptidoglycan biosynthesis (Strych *et al.*, 2001).
	Rv3793	embC	Integral membrane indolylacetylinositol arabinosyltransferase EmbC	Required for anti-TB drug ethambutol resistance, essential for growth and intracellular survival (Sassetti *et al.*, 2003b, Kieser *et al.*, 2014).
other metabolic associated	Rv1876	bfrA	Bacterioferritin	Required for efficient utilization of stored iron under low iron conditions (Reddy *et al.*, 2012, Khare *et al.*, 2017), no human homolog.

Pathway	Rv	Name	Product	Function
	Rv2070c	CobK		Involved in cobalamin (vitamin B12) biosynthesis
	Rv3834c	*serS*	Seryl-tRNA synthetase	Essential gene involved in translation mechanism (Sassetti *et al.*, 2003a, Griffin *et al.*, 2011).
	Rv0509	HemA		
	Rv1393c	-	monoxygenase	
			Latent-TB specific MPPs	
Alternative energy metabolism	Rv1161	*narG*	Respiratory nitrate reductase alpha chain	Associated with virulence and supports anaerobic growth on glycerol (Weber *et al.*, 2000, Huang *et al.*, 2015) and up-regulated in the present of nitrate during dormant stage (Iona *et al.*, 2016).
	Rv3273	MtCA3	Na(+)-dependent bicarbonate transporter BicA / Carbonic anhydrase, beta class	Involved in utilization of alternative carbon sources such as carbon dioxide or carbonic acid in the absence of primary carbon sources (Nishimoria *et al.*, 2010).
cofactor and coenzyme biosynthesis	Rv3119	MoaE1		Essential for the catalytic activity of key enzymes involved in metabolism of carbon, nitrogen and sulphur (Srivastava *et al.*, 2016).
	Rv3324c	MoaC3		
	Rv3601c	PanD		Essential for virulence (Sambandamurthy *et al.*, 2002).
Lipid Metabolism	Rv0222	*echA1*	Enoyl-CoA hydratase	
	Rv0769		3-oxoacyl-[acyl-carrier protein] reductase	
	Rv1141c	*echA11*	Enoyl-CoA hydratase	
	Rv1142c	*echA10*	Enoyl-CoA hydratase	
	Rv2249c	*glpD1*	Glycerol-3-phosphate dehydrogenase	
	Rv2277c		Glycerophosphoryl diester phosphodiesterase	Required for in vivo growth (Sassetti *et al.*, 2003b).
	Rv3229c	*desA3*	linoleoyl-CoA desaturase; stearoyl-CoA 9-desaturase	Required for in vivo growth (Sassetti *et al.*, 2003b).

Pathway	Rv	Name	Product	Function
	Rv3280	*accD5*	Propionyl-CoA carboxylase beta chain	Key enzyme in the catabolic pathway of odd-chain fatty acids and essential for growth (Griffin *et al.*, 2011).
	Rv3506	*fadD17*	Long-chain fatty-acid-CoA ligase, Mycobacterial subgroup FadD17	
Cell wall biosynthesis	Rv0050	*ponA1*	Pyrroline-5-carboxylate reductase	Expressed during intracellular growth (Talaat *et al.*, 2004) and are involved in stationary-phase survival under non-replicating conditions (Rengarajan *et al.*, 2005).
	Rv3682	*ponA2*	Multimodular transpeptidase-transglycosylase	
	Rv1884c	*rpfC*	Resuscitation-promoting factor RpfC	Promote the resuscitation and growth of dormant, non-growing cell (Iona *et al.*, 2016).
Other metabolism associated	Rv1547	*dnaE*	DNA polymerase III alpha subunit	
	Rv3248c	*sahH*	Adenosylhomocysteinase	Involved in regulation of the intracellular concentration of adenosylhomocysteine and it is essential for bacterial growth (Cole *et al.*, 1998, Sassetti *et al.*, 2003a).
	Rv3309c	*upp*	Uracil phosphoribosyltransferase	
	Rv3310	*sapM*	Acid phosphatase	Block phagosome-lysosome fusion, required intracellular survival (Puri *et al.*, 2013).

3.4.6 Pathways Enriched in ATB

3.4.6.1 Amino acid biosynthesis

Two pathways, the amino acid biosynthesis and carbohydrate metabolism, were unique to ATB. Two amino acid biosynthesis enzymes identified included TrpG (Rv0013) and AroD (Rv2537c). Anthranilate synthase (TrpG) catalyses the biosynthesis of tryptophan which is essential for the establishment, maintenance of infection and survival of *M. tuberculosis* (Zhang *et al.,* 2013a). The tryptophan biosynthesis pathway is conserved in mycobacterial species, while absent in humans (Zhang *et al.*, 2013a). Thus, TrpG is an ideal anti-TB drug target (Bashiri *et al.*, 2015) and being a secreted protein, makes it an ideal TB diagnostic biomarker. The 3-dehydroquinase (AroD) catalyses the biosynthesis of aromatic amino acids such as tyrosine, tryptophan and phenylalanine (Parish *et al.*, 2002). Consequently, the co-selection of TrpG and AroD in the present study is indicative of actively growing *M. tuberculosis* that synthesizes precursors such as tryptophan and other aromatic amino acids that are required for protein synthesis. Interestingly, in the selected LTBI gene list no amino acid biosynthesis pathway associated proteins were identified. This is consistent with the report that the synthesis of certain *M. tuberculosis* proteins is reduced by more than 90% in order to shutdown most of the metabolic activities during dormancy (Hu *et al.*, 1998). This was supported by the down-regulation of tryptophan and methionine biosynthesis pathways during dormancy (Gopinath *et al.*, 2015).

3.4.6.2 Carbohydrate metabolism

Two carbohydrate metabolism enzymes, ManB (Rv3264c) and TreY (Rv1563c) essential for growth (Sassetti *et al.*, 2003a) were identified from the ATB list. Of these, TreY (malto-oligosyltrehalose synthase) is involved in starch and sucrose metabolism, specifically,

trehalose biosynthesis. Trehalose, a disaccharide, is a key component of a variety of glycolipids required for growth and virulence of *M. tuberculosis* (De Smet *et al.*, 2000, Murphy *et al.*, 2005, Kalscheuer *et al.*, 2014). Interestingly, trehalose is absent in mammals but produced by bacteria, plants, fungi and insects (De Smet *et al.*, 2000). The trehalose biosynthesis pathway enzymes, products and by-products are abundantly available during the chronic TB disease stage (Korte *et al.*, 2016), therefore, can serve as ATB diagnostic biomarkers.

3.4.6.3 Iron metabolism

Bacterioferritin A (BfrA; Rv1876), another ATB associated metabolic enzyme, is one of two iron storage proteins (BfrA and BfrB) encoded by the *M. tuberculosis* genome and is unique to bacteria (Reddy *et al.*, 2012, Khare *et al.*, 2017). BfrA is required for efficient utilization of stored iron under low iron conditions while BfrB (Rv3841) is required for storage of iron under iron excessive conditions and has been associated with the dormant phase (LTB) (Khare *et al.*, 2017). Since BfrA has no human homolog it may be a good ATB biomarker candidate.

3.4.7 Pathways Enriched in LTBI

3.4.7.1 Alternative energy metabolism

The NarG (Rv1161) and MtCA3 (Rv3273) enzymes involved in alternative energy metabolism were identified in the LTBI list. Nitrate reductase (NarG), is a nitrogen metabolism enzyme, that catalyses the consumption of nitrate in *M. tuberculosis*. This enzyme belongs to the *narGHJI* operon that is associated with virulence and supports anaerobic growth on glycerol (Weber *et al.*, 2000, Huang *et al.*, 2015). NarG is dependent on molybdopterin cofactor to perform its function, is conserved in mycobacteria (Williams *et al.*, 2011) and up-regulated in the presence of nitrate during the dormant stage (Iona *et al.*, 2016). Therefore, since NarG is

associated with LTBI, further investigations are required to determine potential application as drug target or diagnostic biomarker.

The *M. tuberculosis* genome contains at least three beta-class carbonic anhydrases, MtCA1 (Rv1284), MtCA2 (Rv3588c) and MtCA3 (Rv3273) (Ceruso *et al.*, 2014). The carbonic anhydrases are involved in utilization of alternative carbon sources such as carbon dioxide or carbonic acid in the absence of primary carbon sources (Nishimoria *et al.*, 2010). In this study, we identified the transmembrane MtCA3 reportedly associated with pathogenesis and is a TB drug target (Cau *et al.*, 2016).

3.4.7.2 Cofactor and coenzyme biosynthesis

The LTBI cofactor biosynthesis enzymes involved in the molybdopterin biosynthesis pathway include the molybdenum cofactor biosynthesis protein C (MoaC3) and molybdenum cofactor biosynthesis protein E (MoaE1), as well as aspartate alpha-decarboxylase (PanD) involved in the biosynthesis of the coenzyme, pantothenate. Molybdopterin cofactor is essential for the catalytic activity of key enzymes involved in metabolism of carbon, nitrogen and sulphur (Srivastava *et al.*, 2016).

3.4.8 Comparable Cell Wall Biosynthesis Pathway

Peptidoglycan, the key component of *M. tuberculosis* cell wall, is made up of glycan chains and determines the bacterium cell shape. Glycan chains consist of two different sugars that are cross-linked via short peptide side chains. Peptidoglycan polymerization is mediated by enzymes such as bifunctional penicillin-binding proteins that can both polymerize glycan strands and cross-link peptides (Kieser *et al.*, 2014).

Amongst the ATB list, two enzymes associated with peptidoglycan synthesis, the penicillin-binding protein DacB1 (Rv3330) and the alanine racemase, Alr (Rv3423c), were identified. The eleven penicillin-binding proteins (PBPs) encoded by the *M. tuberculosis* genome can be divided into class A, B and C (Sauvage *et al.*, 2008). The membrane associated DacB1 (D-alanyl-D-alanine carboxypeptidase) is a member of class C penicillin-binding proteins that is required for the control of cell elongation and septum development via the regulation of the peptidoglycan cross-linking (Sauvage *et al.*, 2008, Prigozhin *et al.*, 2014). Alr is involved in peptidoglycan cross-linking through conversion of L-alanine to D-alanine which is a necessary precursor for peptidoglycan biosynthesis (Strych *et al.*, 2001). Interestingly, DacB1 is known to be expressed by actively growing bacteria (Simpson *et al.*, 1994), while there is no known Alr homolog in humans making both proteins the potential ATB biomarkers for design of new drugs, vaccines or diagnostic assays.

The LTBI cell wall biosynthesis protein list included EmbC (Rv3793) and 2 bifunctional penicillin-binding protein enzymes, PonA1 (Rv0050) and PonA2 (Rv3682). Both PonA1 and PonA2 belong to class A penicillin-binding proteins (Sauvage *et al.*, 2008) and are involved in cell wall formation by synthesis of cross-linked peptidoglycan from the lipid intermediates (Kieser *et al.*, 2015). The *ponA1* and *ponA2* genes are reportedly expressed *in vivo* (Talaat *et al.*, 2004) and are involved in stationary-phase survival under non-replicating conditions (Rengarajan *et al.*, 2005). Both PonA1 and PonA2 may well be involved in inhibition of resuscitation protein factors (Hett *et al.*, 2010). This inhibition is achieved by the reported increased expression of PonA1 during the dormant stage (Saxena *et al.*, 2008).

In the present study, two resuscitation protein factors (Rpf), RpfB (found in both LTBI and ATB) and RpfC (in LTBI only) were identified. The *M. tuberculosis* genome encodes five Rpfs (A, B, C, D and E) and their function is to promote the resuscitation and growth of dormant cells (Iona *et al.*, 2016). The selection of RpfB and RpfC suggests a possible attempt to

stimulate growth of dormant *M. tuberculosis* cells, possibly triggered by host environmental factors. The selection of RpfB by both LTBI and ATB sera could be due to sero-conversion from LTBI to ATB state. Therefore, RpfB and RpfC are potential biomarkers for monitoring progress from LTBI to ATB in high risk groups such as HIV/AIDS patients. This will possibly permit initiation of TB treatment before the onset of clinical symptoms.

3.4.9 Virulence, Essential and Drug Target Proteins

Using PATRIC's specialty protein dataset that includes virulence factors, drug targets, antibiotic resistance and essential proteins, 17 and 18 specific *M. tuberculosis* proteins unique to ATB and LTBI lists respectively, were identified. Five proteins (Rv1009, Rv1980c, Rv2152c, Rv1357c and Rv1886c) were common to "Active-TB", "Latent-TB" and "Specialty" lists. Two of five are immunogenic proteins Mpt64 (Rv1980c) and FbpB (Rv1886c) (Malen *et al.*, 2008). The essential protein MurC (Rv2152c) is involved in the peptidoglycan biosynthesis pathway (Sassetti *et al.*, 2003a, Griffin *et al.*, 2011). The other 2 proteins are the resuscitation-promoting factor RpfB (Rv1009) and the conserved hypothetical protein (Rv1357c).

3.4.9.1 Virulence proteins

Amongst the ATB speciality proteins were 11 virulence factors including BfrA and the cell surface protein PirG (Rv3810), both required for intracellular survival and replication (Klepp *et al.*, 2009, Reddy *et al.*, 2012); the amino acid permease (Rv1979c) that modulates the host immune response (Kozak *et al.*, 2011); and the transcriptional regulator EspR (Rv3849), a key component of *M. tuberculosis* type VII secretion system (TSS7), ESX-1 system. EspR regulates the transcription of *espACD-rv3613c-rv3612c* operon required for ESX-1 system responsible for the secretion of 6-kDa early secreted antigen target (ESAT6) and 10-kDA

culture filtrate protein (CFP10) (Raghavan *et al.*, 2008, Hunt *et al.*, 2012). There were two PE/PPE family proteins, PE_PGRS33 (Rv1818c) that plays a role in adhesion to host cells (Delogu *et al.*, 2004) and PPE54 (Rv3343c), the member of MPTR (major polymorphic tandem repeats) subfamily which affects the phagosome (Brodin *et al.*, 2010). Lastly, the signal recognition particle receptor FtsY (Rv2921c) is required for bacterial cell division (Ligon *et al.*, 2012).

The LTBI specialty proteins included 12 virulence factors including NarG, RpfC, DesA3, SapM, PanD and PonA2 discussed in the previous section. Other virulence proteins comprised PPE5 (Rv0304c), HspR (Rv0353), Pks6 (Rv0405), MmpL10 (Rv1183), Rv1184c and Rv2277c. Polyketide synthase (Pks6), PPE5 and Rv2277c are reportedly expressed during initial lung infection (Kruh *et al.*, 2010), and therefore, involved in establishment or invasion of the host by *M. tuberculosis* during infection. The transcriptional repressor (HspR) is required to minimize the host immune-surveillance that may result in efficient killing of the pathogen by suppressing the expression of heat-shock genes that encode the immunodominant heat-shock proteins such as Hsp70 (Stewart *et al.*, 2001).

DesA3 (Rv3229c) is a membrane-bound stearoyl coenzyme A (CoA) desaturase that is involved in biosynthesis of oleic acid, an essential component of mycobacterial membrane phospholipids and triglycerides (Chang *et al.*, 2006), therefore, it is required for *in vivo* growth (Sassetti *et al.*, 2003a).

3.4.9.2 Essential proteins

Essential proteins of *M. tuberculosis* include those required for growth and survival during infection. Three ATB associated essential proteins, the TrpG, TopA (Rv3646c) and SerS were identified. TrpG and TopA are essential for the establishment, maintenance and survival of *M. tuberculosis* during infection (Zhang *et al.*, 2013b, Tan *et al.*, 2016). SerS is involved in the

translation mechanism and is, therefore, essential for growth (Sassetti *et al.*, 2003a). The 2 LTB identified specific essential proteins included the ribosomal protein S11p RpsK (Rv3459c) and ribonuclease P protein component RnpA (Rv3923c). Interestingly, MmpL7 (Rv2942), an efflux pump conferring antibiotic resistance, was the only antibiotic resistance transmembrane transport protein (Forrellad *et al.*, 2013) identified in this study.

3.4.9.3 Drug targets

Two potential drug targets identified within the ATB list included the AroD (Rv2537c) and Alr (Rv3423c). Drugs targeting these proteins will inhibit protein synthesis (Strych *et al.*, 2001, Parish *et al.*, 2002) resulting in bacterial death. The 3 LTBI specific prospective drug targets SahH (Rv3248c) and EmbC (Rv3793); and the peptide deformylase Def (Rv0429c) were reported to be essential for *M. tuberculosis* growth (Sassetti *et al.*, 2003a, Griffin *et al.*, 2011).

3.5 CONCLUSION

Phage display is one of the most powerful techniques used to identify proteins that bind to targets of interest. In this study, we identified protein biomarkers recognized by active- and latent-TB patients' antibodies. These included 4 ATB specific biomarkers (TrpG, TreY, Alr and BfrA) that have no human homologs and are essential for pathogenesis, and 5 LTBI specific biomarkers (PonA1, PonA2, NarG, MoaE1 and HspR) associated with dormancy. The identification of disease state-specific biomarkers will contribute towards the development of diagnostic tools that can differentiate active- from latent-TB infection. Some of the biomarkers are potentially suitable for monitoring of disease progression from latent- to active-TB in high risk groups such as HIV/AIDS patients. Future studies should investigate the potential

application of the identified biomarkers in the development of new vaccines, drugs and diagnostic tools.

3.6 REFERENCES

Bashiri, G., Johnston, J.M., Evans, G.L., Bulloch, E.M., Goldstone, D.C., Jirgis, E.N., *et al.* (2015). Structure and inhibition of subunit I of the anthranilate synthase complex of *Mycobacterium tuberculosis* and expression of the active complex. *Acta Crystallogr D Biol Crystallogr* **71,** 2297-2308.

Bitter, W., Houben, E.N.G., Luirink, J. and Appelmelk, B.J. (2009). Type VII secretion in mycobacteria: classification in line with cell envelope structure. *Trends in Microbiology* **17,** 337-338.

Brodin, P., Poquet, Y., Levillain, F., Peguillet, I., Larrouy-Maumus, G., Gilleron, M., *et al.* (2010). High content phenotypic cell-based visual screen identifies *Mycobacterium tuberculosis* acyltrehalose-containing glycolipids involved in phagosome remodeling. *PLoS Pathog* **6,** e1001100.

Cau, Y., Mori, M., Supuran, C.T. and Botta, M. (2016). Mycobacterial carbonic anhydrase inhibition with phenolic acids and esters: kinetic and computational investigations. *Organic & biomolecular chemistry* **14,** 8322-8330.

Ceruso, M., Vullo, D., Scozzafava, A. and Supuran, C.T. (2014). Sulfonamides incorporating fluorine and 1, 3, 5-triazine moieties are effective inhibitors of three β-class carbonic anhydrases from *Mycobacterium tuberculosis*. *Journal of enzyme inhibition and medicinal chemistry* **29,** 686-689.

Chang, Y. and Fox, B.G. (2006). Identification of Rv3230c as the NADPH oxidoreductase of a two-protein DesA3 acyl-CoA desaturase in *Mycobacterium tuberculosis* H37Rv. *Biochemistry* **45,** 13476-13486.

Chiliza, T.E., Pillay, M. and Pillay, B. (2017). Identification of unique essential proteins from a *Mycobacterium tuberculosis* F15/LAM4/KZN phage secretome library. *Pathogens and Disease* **75,** ftx001-ftx001.

Cohen, I., Parada, C., Acosta-Gio, E. and Espitia, C. (2014). The PGRS Domain from PE_PGRS33 of *Mycobacterium tuberculosis* is Target of Humoral Immune Response in Mice and Humans. *Front Immunol* **5,** 236.

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

Corbett, E.L., Watt, C.J., Walker, N., Maher, D., Williams, B.G., Raviglione, M.C. and Dye, C. (2003). The growing burden of tuberculosis: global trends and interactions with the HIV epidemic. *Archives of internal medicine* **163,** 1009-1021.

De Smet, K.A., Weston, A., Brown, I.N., Young, D.B. and Robertson, B.D. (2000). Three pathways for trehalose biosynthesis in mycobacteria. *Microbiology* **146,** 199-208.

Delogu, G., Pusceddu, C., Bua, A., Fadda, G., Brennan, M.J. and Zanetti, S. (2004). Rv1818c-encoded PE_PGRS protein of *Mycobacterium tuberculosis* is surface exposed and influences bacterial cell structure. *Mol Microbiol* **52,** 725-733.

Forrellad, M.A., Klepp, L.I., Gioffré, A., y García, J.S., Morbidoni, H.R., de la Paz Santangelo, M., *et al.* (2013). Virulence factors of the *Mycobacterium tuberculosis* complex. *Virulence* **4,** 3-66.

Gopinath, V., Raghunandanan, S., Gomez, R.L., Jose, L., Surendran, A., Ramachandran, R., *et al.* (2015). Profiling the proteome of *Mycobacterium tuberculosis* during dormancy and reactivation. *Molecular & Cellular Proteomics* **14,** 2160-2176.

Griffin, J.E., Gawronski, J.D., DeJesus, M.A., Ioerger, T.R., Akerley, B.J. and Sassetti, C.M. (2011). High-resolution phenotypic profiling defines genes essential for mycobacterial growth and cholesterol catabolism. *PLoS Pathog* **7,** e1002251.

Hett, E.C., Chao, M.C. and Rubin, E.J. (2010). Interaction and modulation of two antagonistic cell wall enzymes of mycobacteria. *PLoS Pathog* **6,** e1001020.

Hu, Y.M., Butcher, P.D., Sole, K., Mitchison, D.A. and Coates, A.R. (1998). Protein synthesis is shutdown in dormant *Mycobacterium tuberculosis* and is reversed by oxygen or heat shock. *FEMS Microbiol Lett* **158,** 139-145.

Huang, Q., Abdalla, A.E. and Xie, J. (2015). Phylogenomics of Mycobacterium Nitrate Reductase Operon. *Current microbiology* **71,** 121-128.

Hunt, D.M., Sweeney, N.P., Mori, L., Whalan, R.H., Comas, I., Norman, L., *et al.* (2012). Long-range transcriptional control of an operon necessary for virulence-critical ESX-1 secretion in *Mycobacterium tuberculosis*. *J Bacteriol* **194,** 2307-2320.

Iona, E., Pardini, M., Mustazzolu, A., Piccaro, G., Nisini, R., Fattorini, L. and Giannoni, F. (2016). *Mycobacterium tuberculosis* gene expression at different stages of hypoxia-induced dormancy and upon resuscitation. *Journal of Microbiology* **54,** 565-572.

Jankovic, D., Collett, M.A., Lubbers, M.W. and Rakonjac, J. (2007). Direct selection and phage display of a Gram-positive secretome. *Genome biology* **8,** R266.

Kalscheuer, R. and Koliwer-Brandl, H. (2014). Genetics of mycobacterial trehalose metabolism. *Microbiology spectrum* **2**.

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. and Morishima, K. (2017). KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res* **45,** D353-d361.

Khan, I.H., Ravindran, R., Krishnan, V.V., Awan, I.N., Rizvi, S.K., Saqib, M.A., *et al.* (2011). Plasma Antibody Profiles as Diagnostic Biomarkers for Tuberculosis. *Clinical and Vaccine Immunology* **18,** 2148-2153.

Khare, G., Nangpal, P. and Tyagi, A.K. (2017). Differential Roles of Iron Storage Proteins in Maintaining the Iron Homeostasis in *Mycobacterium tuberculosis*. *PloS one* **12,** e0169545.

Kieser, K.J., Baranowski, C., Chao, M.C., Long, J.E., Sassetti, C.M., Waldor, M.K., *et al.* (2015). Peptidoglycan synthesis in *Mycobacterium tuberculosis* is organized into networks with varying drug susceptibility. *Proceedings of the National Academy of Sciences* **112,** 13087-13092.

Kieser, K.J. and Rubin, E.J. (2014). How sisters grow apart: mycobacterial growth and division. *Nature Reviews Microbiology* **12,** 550-562.

Klepp, L.I., Soria, M., Blanco, F.C., Bianco, M.V., Santangelo, M.P., Cataldi, A.A. and Bigi, F. (2009). Identification of two proteins that interact with the Erp virulence factor from *Mycobacterium tuberculosis* by using the bacterial two-hybrid system. *BMC molecular biology* **10,** 3.

Konthur, Z. and Crameri, R. (2003). High-throughput applications of phage display in proteomic analyses. *Targets* **2,** 261-270.

Korte, J., Alber, M., Trujillo, C.M., Syson, K., Koliwer-Brandl, H., Deenen, R., *et al.* (2016). Trehalose-6-Phosphate-Mediated Toxicity Determines Essentiality of OtsB2 in *Mycobacterium tuberculosis* In Vitro and in Mice. *PLoS Pathogens* **12,** e1006043.

Koshkin, A., Nunn, C.M., Djordjevic, S. and de Montellano, P.R.O. (2003). The mechanism of *Mycobacterium tuberculosis* alkylhydroperoxidase AhpD as defined by mutagenesis, crystallography, and kinetics. *Journal of Biological Chemistry* **278,** 29502-29508.

Kozak, R.A., Alexander, D.C., Liao, R., Sherman, D.R. and Behr, M.A. (2011). Region of difference 2 contributes to virulence of *Mycobacterium tuberculosis*. *Infection and immunity* **79,** 59-66.

Kruh, N.A., Troudt, J., Izzo, A., Prenni, J. and Dobos, K.M. (2010). Portrait of a pathogen: the *Mycobacterium tuberculosis* proteome in vivo. *PLoS One* **5,** e13938.

Kunnath-Velayudhan, S., Davidow, A.L., Wang, H.Y., Molina, D.M., Huynh, V.T., Salamon, H., *et al.* (2012). Proteome-Scale Antibody Responses and Outcome of *Mycobacterium tuberculosis* Infection in Nonhuman Primates and in Tuberculosis Patients. *J. Infect. Dis.* **206,** 697-705.

Kunnath-Velayudhan, S. and Gennaro, M.L. (2011). Immunodiagnosis of Tuberculosis: a Dynamic View of Biomarker Discovery. *Clinical Microbiology Reviews* **24,** 792-805.

Kunnath-Velayudhan, S., Salamon, H., Wang, H.-Y., Davidow, A.L., Molina, D.M., Huynh, V.T., *et al.* (2010). Dynamic antibody responses to the *Mycobacterium tuberculosis* proteome. *Proceedings of the National Academy of Sciences* **107,** 14703-14708.

Laal, S., Samanich, K.M., Sonnenberg, M.G., Zolla-Pazner, S., Phadtare, J.M. and Belisle, J.T. (1997). Human humoral responses to antigens of *Mycobacterium tuberculosis*: immunodominance of high-molecular-mass antigens. *Clinical and Diagnostic Laboratory Immunology* **4,** 49-56.

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

Ligon, L.S., Hayden, J.D. and Braunstein, M. (2012). The ins and outs of *Mycobacterium tuberculosis* protein export. *Tuberculosis* **92,** 121-132.

Lyashchenko, K., Colangeli, R., Houde, M., Al Jahdali, H., Menzies, D. and Gennaro, M.L. (1998). Heterogeneous antibody responses in tuberculosis. *Infection and Immunity* **66,** 3936-3940.

Malen, H., Softeland, T. and Wiker, H.G. (2008). Antigen analysis of *Mycobacterium tuberculosis* H37Rv culture filtrate proteins. *Scand J Immunol* **67,** 245-252.

Mattow, J., Siejak, F., Hagens, K., Becher, D., Albrecht, D., Krah, A., *et al.* (2006). Proteins unique to intraphagosomally grown *Mycobacterium tuberculosis*. *Proteomics* **6,** 2485-2494.

Mazandu, G.K. and Mulder, N.J. (2012). Function prediction and analysis of *Mycobacterium tuberculosis* hypothetical proteins. *International journal of molecular sciences* **13,** 7283-7302.

McShane, H. (2005). Co-infection with HIV and TB: double trouble. *International journal of STD & AIDS* **16,** 95-101.

Murphy, H.N., Stewart, G.R., Mischenko, V.V., Apt, A.S., Harris, R., McAlister, M.S., *et al.* (2005). The OtsAB pathway is essential for trehalose biosynthesis in *Mycobacterium tuberculosis*. *Journal of Biological Chemistry* **280,** 14524-14529.

Nishimoria, I., Minakuchia, T., Marescab, A., Cartab, F., Scozzafava, A. and T Supuran, C. (2010). The β-carbonic anhydrases from *Mycobacterium tuberculosis* as drug targets. *Current pharmaceutical design* **16,** 3300-3309.

Palucci, I., Camassa, S., Cascioferro, A., Sali, M., Anoosheh, S., Zumbo, A., *et al.* (2016). PE_PGRS33 Contributes to *Mycobacterium tuberculosis* Entry in Macrophages through Interaction with TLR2. *PLoS One* **11,** e0150800.

Parish, T. and Stoker, N.G. (2002). The common aromatic amino acid biosynthesis pathway is essential in *Mycobacterium tuberculosis*. *Microbiology* **148,** 3069-3077.

Pitarque, S., Herrmann, J.L., Duteyrat, J.L., Jackson, M., Stewart, G.R., Lecointe, F., *et al.* (2005). Deciphering the molecular bases of *Mycobacterium tuberculosis* binding to the lectin DC-SIGN reveals an underestimated complexity. *Biochem J* **392,** 615-624.

Prigozhin, D.M., Krieger, I.V., Huizar, J.P., Mavrici, D., Waldo, G.S., Hung, L.-W., *et al.* (2014). Subfamily-specific adaptations in the structures of two penicillin-binding proteins from *Mycobacterium tuberculosis*. *PloS one* **9,** e116249.

Puri, R.V., Reddy, P.V. and Tyagi, A.K. (2013). Secreted acid phosphatase (SapM) of *Mycobacterium tuberculosis* is indispensable for arresting phagosomal maturation and growth of the pathogen in guinea pig tissues. *PloS one* **8,** e70514.

Raghavan, S., Manzanillo, P., Chan, K., Dovey, C. and Cox, J.S. (2008). Secreted transcription factor controls *Mycobacterium tuberculosis* virulence. *Nature* **454,** 717-U749.

Reddy, P.V., Puri, R.V., Khera, A. and Tyagi, A.K. (2012). Iron storage proteins are essential for the survival and pathogenesis of *Mycobacterium tuberculosis* in THP-1 macrophages and the guinea pig model of infection. *Journal of bacteriology* **194,** 567-575.

Rengarajan, J., Bloom, B.R. and Rubin, E.J. (2005). Genome-wide requirements for *Mycobacterium tuberculosi*s adaptation and survival in macrophages. *Proc Natl Acad Sci U S A* **102,** 8327-8332.

Rustad, T.R., Sherrid, A.M., Minch, K.J. and Sherman, D.R. (2009). Hypoxia: a window into *Mycobacterium tuberculosis* latency. *Cellular Microbiology* **11,** 1151-1159.

Saini, N.K., Baena, A., Ng, T.W., Venkataswamy, M.M., Kennedy, S.C., Kunnath-Velayudhan, S., *et al.* (2016). Suppression of autophagy and antigen presentation by *Mycobacterium tuberculosis* PE_PGRS47. *Nat Microbiol* **1,** 16133.

Samanich, K., Belisle, J. and Laal, S. (2001). Homogeneity of antibody responses in tuberculosis patients. *Infection and immunity* **69,** 4600-4609.

Sambandamurthy, V.K., Wang, X., Chen, B., Russell, R.G., Derrick, S., Collins, F.M., *et al.* (2002). A pantothenate auxotroph of *Mycobacterium tuberculosis* is highly attenuated and protects mice against tuberculosis. *Nat Med* **8,** 1171-1174.

Sassetti, C.M., Boyd, D.H. and Rubin, E.J. (2003a). Genes required for mycobacterial growth defined by high density mutagenesis. *Mol. Microbiol.* **48,** 77-84.

Sassetti, C.M. and Rubin, E.J. (2003b). Genetic requirements for mycobacterial survival during infection. *Proceedings of the National Academy of Sciences* **100,** 12989-12994.

Sauvage, E., Kerff, F., Terrak, M., Ayala, J.A. and Charlier, P. (2008). The penicillin-binding proteins: structure and role in peptidoglycan biosynthesis. *FEMS microbiology reviews* **32,** 234-258.

Savvi, S., Warner, D.F., Kana, B.D., McKinney, J.D., Mizrahi, V. and Dawes, S.S. (2008). Functional characterization of a vitamin B12-dependent methylmalonyl pathway in *Mycobacterium tuberculosis*: implications for propionate metabolism during growth on fatty acids. *Journal of bacteriology* **190,** 3886-3895.

Saxena, A., Srivastava, V., Srivastava, R. and Srivastava, B.S. (2008). Identification of genes of *Mycobacterium tuberculosis* upregulated during anaerobic persistence by fluorescence and kanamycin resistance selection. *Tuberculosis* **88,** 518-525.

Senoputra, M.A., Shiratori, B., Hasibuan, F.M., Koesoemadinata, R.C., Apriani, L., Ashino, Y., *et al.* (2015). Diagnostic value of antibody responses to multiple antigens from *Mycobacterium tuberculosis* in active and latent tuberculosis. *Diagnostic microbiology and infectious disease* **83,** 278-285.

Simpson, E.B., Hancock, T.W. and Buchanan, C.E. (1994). Transcriptional control of dacB, which encodes a major sporulation-specific penicillin-binding protein. *Journal of bacteriology* **176,** 7767-7769.

Srivastava, S., Pathak, M., Pandey, H., Tripathi, S., Garg, R., Misra-Bhattacharya, S. and Arora, A. (2016). Molecular characterization of novel immunodominant molybdenum cofactor biosynthesis protein C1 (Rv3111) from *Mycobacterium tuberculosis* H37Rv. *Biochimica et Biophysica Acta (BBA)-General Subjects* **1860,** 694-707.

Stewart, G.R., Snewin, V.A., Walzl, G., Hussell, T., Tormay, P., O'Gaora, P., *et al.* (2001). Overexpression of heat-shock proteins reduces survival of *Mycobacterium tuberculosis* in the chronic phase of infection. *Nat Med* **7,** 732-737.

Strych, U., Penland, R.L., Jimenez, M., Krause, K.L. and Benedik, M.J. (2001). Characterization of the alanine racemases from two mycobacteria. *FEMS Microbiol Lett* **196,** 93-98.

Talaat, A.M., Lyons, R., Howard, S.T. and Johnston, S.A. (2004). The temporal expression profile of *Mycobacterium tuberculosis* infection in mice. *Proceedings of the National Academy of Sciences of the United States of America* **101,** 4602-4607.

Tan, K., Cao, N., Cheng, B., Joachimiak, A. and Tse-Dinh, Y.-C. (2016). Insights from the Structure of *Mycobacterium tuberculosis* Topoisomerase I with a novel protein fold. *Journal of molecular biology* **428,** 182-193.

The_UniProt_Consortium (2017). UniProt: the universal protein knowledgebase. *Nucleic Acids Research* **45,** D158-D169.

Wallis, R.S., Doherty, T.M., Onyebujoh, P., Vahedi, M., Laang, H., Olesen, O., *et al.* (2009). Biomarkers for tuberculosis disease activity, cure, and relapse. *The Lancet Infectious Diseases* **9,** 162-172.

Wattam, A.R., Davis, J.J., Assaf, R., Boisvert, S., Brettin, T., Bun, C., *et al.* (2017). Improvements to PATRIC, the all-bacterial Bioinformatics Database and Analysis Resource Center. *Nucleic Acids Research* **45,** D535-D542.

Weber, I., Fritz, C., Ruttkowski, S., Kreft, A. and Bange, F.-C. (2000). Anaerobic nitrate reductase (narGHJI) activity of *Mycobacterium bovis* BCG in vitro and its contribution to virulence in immunodeficient mice. *Mol. Microbiol.* **35,** 1017-1025.

WHO (2017). Global Tuberculosis Report 2017 *World Health Organization*.

Williams, M.J., Kana, B.D. and Mizrahi, V. (2011). Functional analysis of molybdopterin biosynthesis in mycobacteria identifies a fused molybdopterin synthase in *Mycobacterium tuberculosis*. *Journal of bacteriology* **193,** 98-106.

Wipperman, M.F., Sampson, N.S. and Thomas, S.T. (2014). Pathogen roid rage: Cholesterol utilization by *Mycobacterium tuberculosis*. *Critical reviews in biochemistry and molecular biology* **49,** 269-293.

Ye, J., Fang, L., Zheng, H., Zhang, Y., Chen, J., Zhang, Z., *et al.* (2006). WEGO: a web tool for plotting GO annotations. *Nucleic acids research* **34,** W293-W297.

Zhang, Y.J., Reddy, M.C., Ioerger, T.R., Rothchild, A.C., Dartois, V., Schuster, B.M., *et al.* (2013a). Tryptophan biosynthesis protects mycobacteria from CD4 T-cell-mediated killing. *Cell* **155,** 1296-1308.

Zhang, Y.J. and Rubin, E.J. (2013b). Feast or famine: the host-pathogen battle over amino acids. *Cell Microbiol* **15,** 1079-1087.

CHAPTER 4: B-cell epitope derived T-cell epitopes as vaccine candidates to confer antibody and cellular mediated immunity against *M. tuberculosis* infection

Thamsanqa E. Chiliza[1], Manormoney Pillay[2] and Balakrishna Pillay[1*]

[1]Discipline of Microbiology, School of Life Sciences, College of Agriculture, Engineering and Science, University of KwaZulu-Natal, Westville Campus, South Africa;

[2]Medical Microbiology, School of Laboratory Medicine and Medical Sciences, College of Health Sciences, University of KwaZulu-Natal, Durban, South Africa

Manuscript in preparation

4.1 ABSTRACT

Mycobacterium bovis bacille Calmette–Guérin (BCG) vaccine is widely used to provide immunity against *Mycobacterium tuberculosis* (*Mtb*) infection in new-borns and infants. However, BCG efficacy decreases in adolescents and also does not provide consistent immunity to the development of active TB upon exposure to *Mtb* in adults who are mostly affected by nearly all new *Mtb* infections. The present study aimed to identify and confirm *Mtb*-specific B-cell and T-cell epitopes with potential use in the design of a new multi-epitope peptide TB vaccine. Previously, we selected and identified 191 *Mtb* proteins recognized by TB patients' humoral response. Using reverse vaccinology, 40 proteins were selected and ranked according to predicted subcellular localization, transmembrane domains, adhesive properties and antigenicity. B-cell epitopes were identified and major histocompatibility complex (MHC) class I and II epitopes were predicted. Five proteins (Rv1424c, Rv1884c, Rv1926c, Rv2376 and Rv3036c) and two proteins (Rv1271c and Rv3803c) possessed epitopes with high binding affinity to MHC class I ($CD8^+$ specific) and II ($CD4^+$ specific) molecules, respectively. Two epitopes RMPTGMPPK (Rv0361) and QLPPTDPRY (Rv3682) with binding affinity for non-classical MHC class I HLA-E*01:01 and HLA-E*01:03 alleles were identified, respectively. The HLA-E epitopes demonstrated the highest population coverage (99.88%) among the world's population, and therefore, are potential TB vaccine candidates since HLA-E alleles are not down-regulated by HIV co-infection. In conclusion, the B- and T-cells epitopes identified in this study represent potential candidates that may invoke both humoral and cellular immune responses. These findings will potentially accelerate and expedite the formulation of effective and cost-efficient multi-epitope peptide vaccines and diagnostic tests against *Mtb* infection.

KEYWORDS: BCG vaccine, reverse vaccinology, T-cell epitope, multi-epitope vaccine

4.2 INTRODUCTION

Mycobacterium tuberculosis (*Mtb*) is the etiological agent of tuberculosis (TB), the leading cause of death from infectious diseases. In 2016, the World Health Organization (WHO) recorded approximately 1.7 million deaths and 6.3 million new cases of TB (WHO, 2017). The available treatment regimens have been rendered impotent by the emergence of deadly multidrug-resistant (MDR), extensively drug resistant (XDR), and totally drug resistant (TDR) strains of *Mtb* (Pillay *et al.*, 2007, Klopper *et al.*, 2013). The treatment of drug resistant TB pathogens requires a prolonged course of multiple antimicrobial agents, some of which have serious toxic side effects, leading to non-adherence in patients (WHO, 2017). Exposure of such cases to household contacts and health care personnel during this period results in increasing the pool of drug resistant strains. Thus, new and efficacious vaccines are urgently needed to prevent such infection.

The globally used *Mycobacterium bovis* bacille Calmette–Guérin (BCG) vaccine, first introduced in 1921, is the only TB vaccine currently available (WHO, 2017). The BCG vaccine comprises an attenuated strain of *Mycobacterium bovis*, the causative agent of TB in cattle (Sakula, 1983). It is administered to over 90% of new-borns annually, to provide protection from *Mtb* infection in infants and young children, globally. However, BCG vaccine efficacy is reportedly between 50% and 80%, providing infants and children with protection for up to 10 years, and is unsuitable for HIV-exposed infants (Colditz *et al.*, 1995, Andersen *et al.*, 2005, Kaufmann *et al.*, 2017). Thus, there is a need for an improved vaccine for new-borns. The BCG efficacy decreases in adolescents and also does not provide consistent immunity to the development of active TB upon exposure to *Mtb* in adults (Orr *et al.*, 2014). The development of an effective vaccine to prevent TB transmission in these groups will have enormous impact on curbing the epidemic. Since an ideal vaccine should be able to prevent the establishment or reactivation of *Mtb* infection (Boggiano *et al.*, 2017), the development of novel TB vaccine

candidate biomarkers to fulfil this requirement is an imperative in the national and global context. In previous studies, the immunodominant antigens such as early secretory antigen target 6 (ESAT-6), culture filtrate protein 10 (CFP-10), and antigen 85 complex were investigated as vaccine candidate (Li *et al.*, 2014). However, the immunodominant antigens are most likely used to elicit the host immune response that is beneficial to *Mtb* pathogen (Orr *et al.*, 2014). Similar observation regarding use of these antigens as diagnostic biomarkers have yielded no results to date, thus the ongoing search for new biomarkers for vaccine and diagnostic application. Therefore, TB vaccine development investigators should not only focus on immunodominant proteins but must include the subdominant subsets which could illicit a protective immune response against *Mtb* (Orr et al., 2014).

The identification of T-cell epitopes capable of producing both a humoral immune response and T-cell mediated immunity could lead to the design of an effective subunit polypeptide TB vaccine. The subunit vaccine can be composed of one or more immunogenic T-cell epitopes to improve vaccine efficacy and provide wider coverage of binding affinity to MHC class I and II molecules of individuals of different ethnicity globally. Therefore, the identification of potential epitopes with higher binding affinity for MHC's is necessary to design peptide vaccines (Rashid *et al.*, 2017).

Currently, there are twelve candidate vaccines undergoing clinical trials (WHO, 2017). These can be classified into three categories based on their composition: (i) whole-cell or lysates of mycobacteria, (ii) viral vector vaccines, and (iii) adjuvanted recombinant protein vaccines. The whole-cell or lysates may consist of several thousand *Mtb*-specific antigens. The viral vector based vaccines and recombinant protein vaccines comprise up to four *Mtb*-specific antigens (Kaufmann *et al.*, 2017).

Since *Mtb* secretory proteins are targets of host humoral and cellular mediated immune responses, they represent potential vaccine candidates and immunodiagnostic targets

(Lyashchenko *et al.*, 1998, Kunnath-Velayudhan *et al.*, 2012). In our previous study, 191 antigenic proteins recognized by latent-TB and active-TB patients' humoral response were identified by immunoscreening of the *Mtb* phage secretome library against the serum samples (Chiliza et al., under review in Tuberculosis journal). In the present study, the 191 antigenic proteins (*Mtb* fragments) were evaluated for the design of a recombinant subunit polypeptide TB vaccine. This was done using an *in silico* strategy combined with the verification of B-cell epitopes of 191 phage displayed *Mtb* proteins and by identifying B-cell epitope-derived T-cell epitopes binding with the MHC class I and II molecules.

4.3 MATERIALS AND METHODS

4.3.1 Antigenic *Mtb* Proteins

In a previous study, blood samples were collected with informed consent from active-TB (*n*=20), latent-TB (*n*=15) patients and healthy tuberculin negative participants (*n*=20). The serum from active- and latent-TB patients was used to identify phage-displayed *Mtb* proteins recognized by patients' antibodies (Chapter 3, Chiliza et al., submitted in Jan 2018). NCBI Gene IDs for 191 *Mtb* proteins were retrieved in UniProtKB for analysis using the *Mtb* H37Rv genome from the Vaxign program list as the reference. This study was approved by the University of KwaZulu-Natal Biomedical Research Ethics Committee (Ref. BE236/13).

4.3.2 Extracellular Protein Preselection

Vaxign (http://www.violinet.org/vaxign/index.php), the web-based pipeline, dedicated to vaccine design, was used to shortlist potential vaccine candidates based on their cellular localization, probability of possessing adhesin-like characteristics, and the number of

transmembrane helices. The Vaxign integrated bioinformatics tools included PSORTb2.0, SPAAN and the HMMTOP used for subcellular localization; and the hidden Markov model for the prediction of adhesin characteristics and transmembrane helix topology, respectively (Gardy *et al.*, 2004, Sachdeva *et al.*, 2004, Käll *et al.*, 2007). Extracellular proteins, outer membrane proteins, and proteins of unknown localization with an adhesin probability score >0.50 and either 1 or no transmembrane helices were pre-selected for further analysis.

4.3.3 Protein Antigenicity

VaxiJen v2.0, a server for the prediction of protective antigens, was used to predict the antigenicity of each protein (http://www.ddg-pharmfac.net/vaxijen/VaxiJen/VaxiJen.html). The antigenic score was determined based on the physicochemical properties of proteins and those with a value ≥ 0.4 were selected (Doytchinova *et al.*, 2007).

4.3.4 Subcellular Localization Prediction

Subcellular localization of selected proteins was predicted using SignalP 4.1 for Sec-dependent signal peptides whilst Tat-dependent signal peptides were predicted with TatP 1.0 (Bendtsen *et al.*, 2005, Petersen *et al.*, 2011). SecretomeP 2.0 was used for the prediction of non-classical secreted proteins (Bendtsen *et al.*, 2005). All three programs were freely available from the Centre for Biological Sequence Analysis at the Technical University of Denmark (http://www.cbs.dtu.dk/services). The subcellular localization of all remaining proteins was predicted with TBPred (http://www.imtech.res.in/raghava/tbpred/) (Rashid *et al.*, 2007).

4.3.5 Verification of B-cell Epitopes

The previously identified *Mtb* protein fragments of pre-selected vaccine candidates were analysed for B-cell epitopes. BCPreds software was used to analyse amino acid sequences (http://ailab.ist.psu.edu/bcpred/), using 2 different algorithms: the amino acid pair (AAP) antigenicity method (Chen *et al.*, 2007) and string kernels method (EL-Manzalawy *et al.*, 2008). Antigenic linear non-overlapping 20-mer epitopes were predicted from the whole protein sequence, and B-cell epitopes of each preselected protein with a score >0.8 were further characterized.

4.3.6 T-Cell Epitope Prediction

T-cell epitopes were predicted from the selected B-cell epitopes. T-cell Epitope Prediction Tools (TepiTool) from Immune Epitope Database and Analysis Resource (IEDB) (http://tools.iedb.org/tepitool/) were used for the prediction of affinity for MHC class I and MHC class II alleles (Paul *et al.*, 2016).

T-cell epitopes binding to MHC class I alleles were predicted using the SMM (Peters *et al.*, 2005) and NetMHCpan (Nielsen *et al.*, 2007, Hoof *et al.*, 2009) methods. The binding affinity of 9-mer and 10-mer epitope peptides to the 27 most common human MHC class I HLA-A and HLA-B alleles was determined. Peptides binding to most alleles with percentile ranks below 1% and/or half-maximal inhibitory concentration (IC_{50}) of ≤250 were selected.

The MHC II binding was predicted using the Consensus method which employs SMM_align, NN_align, Combinatorial library, Sturniolo methods (Wang *et al.*, 2008, Wang *et al.*, 2010) and NetMHCIIpan (Nielsen *et al.*, 2008, Karosiene *et al.*, 2013). IEDB recommended settings were applied and 15mer peptides with 10 amino acid residue overlaps were selected if their percentile score was below 20% and binding of two-third (8/15) of HLA-DR allele species was demonstrated.

4.3.7 Population Coverage Prediction

The percentage of individuals within the global population responding to predicted T-cell epitopes was estimated using the IEDB population coverage calculation tool (http://tools.iedb.org/tools/population/iedb_input) (Bui *et al.*, 2006). HLA-DRB3*02:02, HLA-DRB3*01:01, HLA-DRB4*01:01, HLA-DRB5*01:01 alleles were excluded by the IEDB population coverage server during calculation. The combined score for MHC classes I and II was assessed for the analysis of the population coverage.

4.4 RESULTS AND DISCUSSION

4.4.1 Protein Preselection

M. tuberculosis surface membrane, and membrane anchored and secreted proteins (secretome), are known to be involved in virulence and participate in host-pathogen interaction (Barh *et al.*, 2010). TB patients produce antigen-specific antibodies in response to the antigenic *Mtb* secretome. These antibodies specifically recognise the B-cell epitopes on the antigens. In our previous study, an *in vitro* based approach was used to select immunogenic phage displayed *Mtb* F15/LAM4/KZN secretome fragments recognized by latent- and active-TB patients' sera antibodies. The identified fragments encoded 191 open reading frames which were investigated in the present study for their potential application in TB vaccine design.

Forty-four of the 191 proteins were shortlisted as potential vaccine antigens according to the following criteria: localization (includes proteins with unknown localization), adhesion features, and number of transmembrane helices (Table 4.1). The 22 proteins with an adhesion score of ≥0.5, included 12 extracellular proteins and 10 proteins with unknown localization. The other 22 proteins were of unknown localization.

All 44 proteins were further analysed to determine their cellular localization. SignalP 4.1 server identified 27 secreted proteins with signal peptides for secretion. An additional 6 proteins that

are possibly secreted via the alternative non-classical pathway were confirmed as extracellular proteins by SecretomeP. TBPred was used to confirm the localization of 11 proteins; 6 are membrane anchored, 3 are integral membrane proteins, and 2 are secreted proteins (Table 4.1).

4.4.2 Protein Selection According To Antigenicity

The antigenicity scores predicted by VaxiJen v2.0 server (Doytchinova *et al.*, 2007) for the 44 proteins ranged from 0.3289 to 2.4465 with 40 candidates displaying a score ≥ 0.4. Only 4 proteins (Rv0680c, Rv2429, Rv3054c and Rv3265c) with an antigenicity score below 0.4, were excluded (Table 4.1). The data on 40 identified candidate vaccines listed in Table 4.1 makes a compelling case for their consideration in the design of new protective TB vaccines.

Highly ranked antigenic proteins included the polymorphic GC-rich-repetitive sequence (PGRS) subfamily proteins of the PE (Proline and Glutamic acid rich) family, PE_PGRS17 (Rv0978c), PE_PGRS25 (Rv1396c), PE_PGRS27 (Rv1450c) and PE_PGRS50 (Rv3345c) with antigenicity scores of 2.4465, 2.3904, 2.2197 and 0.8921, respectively. None of these antigenic proteins had been previously identified among the 16 PE_PGRS proteins reported as potential candidate vaccines with antigenicity scores ranging from 0.6034 to 3.4881 (Monterrubio-Lopez *et al.*, 2015). PE_PGRS17 is known to elicit a strong B-cell humoral response during different clinical stages of both adult and paediatric TB patients (Narayana *et al.*, 2007). This protein also triggers the innate immune response via activation of dendritic cells, resulting in $CD4^+$ T-cells stimulation and hence, regulation of the progression of TB disease (Bansal *et al.*, 2010). The other highly antigenic proteins included Apa (Rv1860), MPT53 (Rv2878c), MPT63 (Rv1926c), MPT64 (Rv1980c), FbpB (Rv1886c) and FbpD (Rv3803c). MPT32, an alanine and proline-rich protein (Apa) binds to host macrophages during infection (Romain *et al.*, 1999) and elicits a B-cell humoral response and stimulates

both $CD4^+$ and $CD8^+$ T-cell protective immunity against TB (Kumar *et al.*, 2003). MPT63 and MPT64 are extracellular proteins that reportedly provide protective immunity (Horwitz *et al.*, 1995, Zvi *et al.*, 2008). The FbpB and FbpD that also function as adhesins are responsible for the attachment of *Mtb* to host cells including macrophages (Puech *et al.*, 2002, Wilson *et al.*, 2004). The resuscitation-promoting factors [RpfB (Rv1009) and RpfC (Rv1884c)] are highly immunogenic, triggering a B-cell response and elicits T-cell protective immunity against TB (Yeremeev *et al.*, 2003, Romano *et al.*, 2012). The protein with the lowest antigenicity of 0.4045 was identified as the conserved immunogenic secretory protein TB22.2 (Rv3036c).

Table 4. 1: List of proteins meeting selection criteria of potential candidate vaccines. Proteins were ranked and selected based on antigenicity score of ≥0.4.

Locus Tag	Gene Symbol	Protein Note	Localization	Probability	Adhesin Probability	Trans-membrane helices	Antigenicity
Rv1450c	PE_PGRS27	PE-PGRS family protein	Extracellular	0.797	0.731	0	2.4465
Rv3345c	PE_PGRS50	PE-PGRS family protein	Extracellular	0.797	0.718	0	2.3904
Rv1396c	PE_PGRS25	PE-PGRS family protein	Unknown	0.646	0.711	0	2.2197
Rv0978c	PE_PGRS17	PE-PGRS family protein	Unknown	0.646	0.612	0	0.8921
Rv0361	Rv0361	Probable conserved membrane protein	Unknown	0.25	0.224	1	0.7237
Rv3333c	Rv3333c	Hypothetical Proline rich protein	Unknown	0.333	0.536	0	0.6905
Rv1813c	Rv1813c	Hypothetical protein Rv1813c	Unknown	0.333	0.287	1	0.6637
Rv1435c	Rv1435c	Probable conserved Proline, Glycine, Valine-rich secreted protein	Unknown	0.333	0.241	0	0.6481
Rv1910c	Rv1910c	Probable exported protein	Unknown	0.25	0.624	1	0.6481
Rv1926c	mpt63	Immunogenic protein MPT63	Extracellular	1	0.506	1	0.6419
Rv1271c	Rv1271c	Hypothetical conserved secreted protein	Unknown	0.333	0.714	1	0.6324
Rv3212	Rv3212	Hypothetical conserved Alanine, Valine rich protein	Unknown	0.333	0.319	1	0.6303
Rv1886c	fbpB	Secreted antigen 85-B FbpB (Ag85B)	Extracellular	1	0.618	1	0.5926
Rv1980c	mpt64	Immunogenic protein MPT64	Extracellular	1	0.635	1	0.5856
Rv2878c	mpt53	Soluble secreted antigen MPT53 precursor	Extracellular	1	0.604	1	0.5816

Locus Tag	Gene Symbol	Protein Note	Localization	Probability	Adhesin Probability	Trans-membrane helices	Antigenicity
Rv1804c	Rv1804c	Hypothetical protein Rv1804c	Unknown	0.25	0.338	0	0.5750
Rv3682	ponA2	Bifunctional membrane-associated penicillin-binding protein PonA2	Extracellular	0.972	0.495	1	0.5648
Rv2499c	Rv2499c	Posible oxidase regulatory-related protein	Unknown	0.25	0.321	0	0.5586
Rv1269c	Rv1269c	Conserved probable secreted protein	Extracellular	0.913	0.614	0	0.5539
Rv3256c	Rv3256c	Hypothetical protein Rv3256c	Unknown	0.25	0.147	0	0.5514
Rv1424c	Rv1424c	Possible membrane protein	Unknown	0.25	0.234	0	0.5438
Rv3218	Rv3218	Hypothetical protein Rv3218	Unknown	0.25	0.264	0	0.5404
Rv1860	apa	Alanine and Proline rich secreted protein Apa	Extracellular	1	0.564	1	0.5244
Rv1291c	Rv1291c	Hypothetical conserved secreted protein	Unknown	0.333	0.719	1	0.518
Rv1009	rpfB	Resuscitation-promoting factor RpfB	Unknown	0.25	0.253	1	0.5117
Rv2376c	cfp2	Low molecular weight antigen Cfp2	Extracellular	1	0.619	1	0.5109
Rv1268c	Rv1268c	Hypothetical protein Rv1268c	Unknown	0.333	0.396	1	0.5014
Rv0179c	lprO	Possible lipoprotein LprO	Unknown	0.25	0.547	1	0.4921
Rv3896c	Rv3896c	Hypothetical protein Rv3896c	Unknown	0.25	0.298	0	0.4798
Rv2944	Rv2944	IS1533 transposase	Unknown	0.25	0.245	0	0.4687
Rv1884c	rpfC	Resuscitation-promoting factor RpfC	Unknown	0.25	0.388	0	0.4657
Rv3572	Rv3572	Hypothetical protein Rv3572	Unknown	0.333	0.646	0	0.4650

Locus Tag	Gene Symbol	Protein Note	Localization	Probability	Adhesin Probability	Trans-membrane helices	Antigenicity
Rv1810	Rv1810	Hypothetical protein Rv1810	Unknown	0.333	0.442	1	0.4634
Rv3922c	Rv3922c	Hypothetical protein Rv3922c	Unknown	0.25	0.199	0	0.4505
Rv3803c	fbpD	Secreted MPT51/MPB51 antigen protein FbpD (Ag85C)	Extracellular	1	0.560	1	0.4504
Rv0559c	Rv0559c	Possible conserved secreted protein	Unknown	0.333	0.599	0	0.4490
Rv1984c	cfp21	Probable cutinanse precursor Cfp21	Unknown	0.333	0.626	0	0.4470
Rv3395A	Rv3395A	Probable membrane protein	Unknown	0.25	0.379	2	0.4469
Rv1329c	dinG	Probable ATP-dependent helicase DinG	Unknown	0.25	0.170	0	0.4427
Rv3036c	TB22.2	Probable conserved secreted protein TB22.2	Extracellular	0.973	0.665	0	0.4045
Rv3265c	wbbL1	Probable dTDP-RHA:A-D-GlcNAc-diphosphoryl polyprenol	Unknown	0.25	0.166	0	0.3626
Rv3054c	Rv3054c	Hypothetical protein Rv3054c	Unknown	0.25	0.222	0	0.3501
Rv2429	ahpD	Alkyl hydroperoxide reductase D protein AHPD	Unknown	0.25	0.441	0	0.3289
Rv0680c	Rv0680c	Probable conserved transmembrane protein	Unknown	0.333	0.386	2	0.3421

4.4.3 Validation of B-cell Epitopes

The first step in vaccine design has been reported to be the identification of B-cell epitopes (Barh *et al.*, 2010). Therefore, the complete sequences of 40 proteins were assessed in terms of B-cell epitopes using BCPreds and AAP algorithms. An alternative, ABCPred program (http://www.imtech.res.in/raghava/abcpred/) was used for B-cell epitope prediction of 8 proteins (Rv1271c, Rv1291c, Rv1329c, Rv1396c, Rv1435c, Rv1813c, Rv2499c and Rv3345c). Of the original phage displayed *Mtb* secretome peptides recognized by the B-cell humoral response of TB patients, a total of 97 B-cell epitopes (Appendix A, Table S4.1) was confirmed. This validated the *in vitro* approach that we previously used to identify *Mtb* antigens recognized by TB patients' antibodies. The length of the epitopes was 15 and 20 amino acids for ABCPred and BCPreds program, respectively.

Generally, the antibody response during TB has been directed against extracellular and membrane-associated antigens (Kunnath-Velayudhan *et al.*, 2010). Antibodies targeting membrane associated antigens have been suggested to enhance the protective immunity against TB (Achkar *et al.*, 2014). Since antibodies can bind *Mtb* antigens at the site of infection, they may inhibit cell adhesion if directed against B-cell epitopes of surface-exposed antigens. This will prevent entry of mycobacteria into host cells and limit the dissemination of *Mtb* during reinfection of host cells (Jacobs *et al.*, 2016). Surface-binding antibodies are able to trigger killing by opsono-phagocytosis, therefore, the identified B-cell epitopes may induce antibodies that are able to opsonize mycobacteria and initiate complement-mediated lysis and uptake into neutrophils for destruction (Plotkin *et al.*, 2008).

4.4.4 B-cell Epitope Derived T-cell Epitopes

4.4.4.1 Classical MHC class I and II epitopes

Previous studies have shown that vaccines that induce both antibody and cellular responses are able to elicit protection that is superior to live BCG (Prados-Rosales *et al.*, 2014). Therefore, a good vaccine candidate should produce both the B-cell and T-cell mediated immunity for immune protection against *Mtb* infection (Kaufmann *et al.*, 2017). In order to identify T-cell epitopes, each B-cell epitope was analysed for the identification of T-cell epitopes within the B-cell epitope sequence. The MHC molecules that recognize T-cell epitopes are termed human leukocyte antigen (HLA) alleles and are the most polymorphic molecules in humans (Bui *et al.*, 2006). They are classified into MHC class I (specific to $CD8^+$ T-cells) and MHC class II (specific to $CD4^+$ T-cells) (Guermonprez *et al.*, 2002). MHC class I molecules are divided into classical and non-classical alleles. The classical MHC class I molecules consists of HLA-A, HLA-B, and HLA-C which are highly polymorphic and their function is to present allele-specific host or pathogen peptides to Cytotoxic T-cells (CD8+) (Kraemer *et al.*, 2015). The non-classical MHC class I molecules consists of HLA-E, HLA-H, HLA-G and HLA-F (Carlini *et al.*, 2016). The specificity and diversity of HLA alleles is essential to stimulate effective cellular immune responses (Germain, 1994). Therefore, identification of peptides (T-cell epitopes) that bind to more than one HLA allele on both classes could lead to the discovery of promiscuous T-cell epitopes that can provide protective immunity to a large proportion of human population.

In the current study, IEDB analysis predicted both classical MHC class I and MHC class II restricted alleles on the basis of the percentile score and IC_{50} value, respectively. For the first level screening, common alleles across global populations (default setting of 27 MHC class I and 26 MHC class II) were used to identify T-cell epitopes recognized by most alleles. Of these, 8 classical MHC class I and 18 MHC class II T-cell epitopes that interacted with at least

8 and 10 alleles, respectively (Table 4.2) were selected. Two B-cell epitope peptides QQEMAAAVAHAFETGE (Rv1329c) and IVTSNKAFGRWGEVFGGDDV (Rv2944) were predicted to possess the highest binding affinity for 13 and 12 class I alleles, respectively. Figure 4.1 summarizes the approach used to identify and select T-cell epitopes from 191 antigenic *Mtb* proteins.

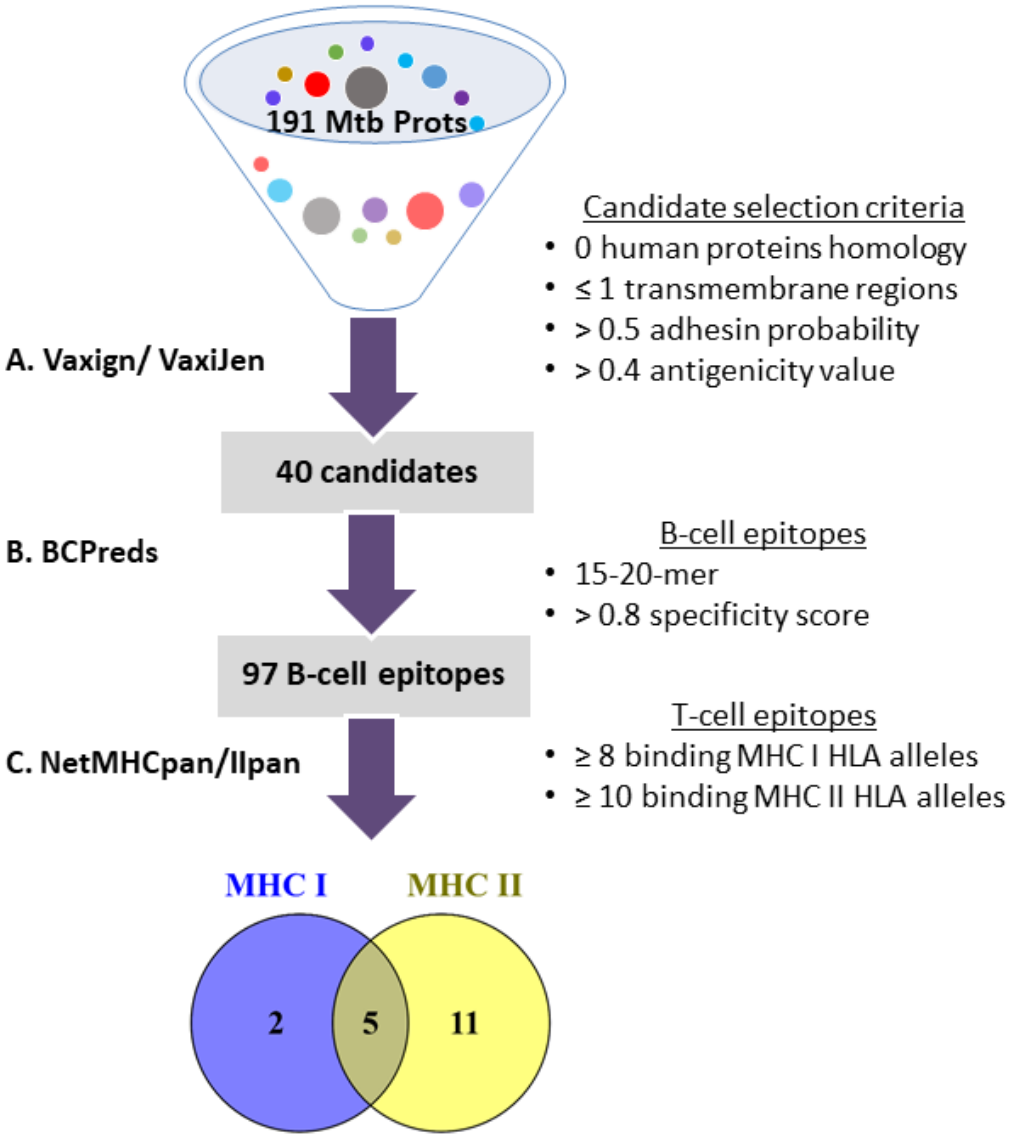

Fig. 4. 1: Vaccinomics workflow approach applied to the 191 pre-selected *M. tuberculosis* antigenic protein fragments to select novel vaccine candidates. The process starts with Vaxign/VaxiJen selecting 40 vaccine candidate proteins. These candidates were analyzed for B-cell (BCPreds) and T-cell (NetMHC) epitopes. At the end, MHC I and II epitopes recognized by most HLA alleles were chosen as potential vaccine antigens. (All bioinformatics tools used are detailed in methods section).

Table 4. 2: List of MHC Class I and II epitopes with high scoring restricted alleles.

Rv	Description	Epitope	Allele
		MHC Class I	
Rv1329c	Probable ATP-dependent helicase DinG	QQEMAAAVAHAFETGE	HLA-A*02:03, HLA-A*23:01, HLA-A*32:01, HLA-A*68:01, HLA-A*68:02, HLA-B*07:02, HLA-B*15:01, HLA-B*35:01, HLA-B*40:01, HLA-B*44:03, HLA-B*53:01, HLA-B*57:01, HLA-B*58:01
Rv1424c	Possible membrane protein	RPPAEKLVFPVLLGILTLLL	HLA-A*02:01, HLA-A*02:03, HLA-A*02:06, HLA-A*32:01, HLA-A*68:02, HLA-B*07:02, HLA-B*35:01, HLA-B*40:01, HLA-B*44:02, HLA-B*58:01
Rv1884c	Resuscitation-promoting factor (RpfC)	GNASATSGDMSSMTRIAKPL	HLA-A*03:01, HLA-A*11:01, HLA-A*26:01, HLA-A*30:01, HLA-A*31:01, HLA-A*68:01, HLA-B*08:01, HLA-B*15:01, HLA-B*35:01
Rv1926c	Immunogenic protein MPT63	ATFAAPVALAAYPITGKLGS	HLA-A*02:03, HLA-A*02:06, HLA-A*03:01, HLA-A*11:01, HLA-A*30:02, HLA-A*32:01, HLA-A*68:01, HLA-A*68:02, HLA-B*15:01, HLA-B*35:01
		TMTDTVGQVVLGWKVSDLKS	HLA-A*01:01, HLA-A*02:01, HLA-A*02:03, HLA-A*02:06, HLA-A*68:02, HLA-B*53:01, HLA-B*57:01, HLA-B*58:01
Rv2376c	Low molecular weight antigen CFP2	AGGPVVYQMQPVVFGAPLPL	HLA-A*02:01, HLA-A*02:03, HLA-A*02:06, HLA-A*23:01, HLA-A*24:02, HLA-A*30:02, HLA-A*32:01, HLA-B*07:02, HLA-B*15:01, HLA-B*35:01
Rv2944	IS1533 transposase	IVTSNKAFGRWGEVFGGDDV	HLA-A*02:03, HLA-A*02:06, HLA-A*11:01, HLA-A*23:01, HLA-A*24:02, HLA-A*30:01, HLA-A*31:01, HLA-A*33:01, HLA-A*68:01, HLA-B*40:01, HLA-B*57:01, HLA-B*58:01
Rv3036c	Conserved secreted protein TB22.2	HVHASGPKYMLDMTFPVDYP	HLA-A*02:01, HLA-A*02:03, HLA-A*02:06, HLA-A*23:01, HLA-A*26:01, HLA-A*30:02, HLA-A*31:01, HLA-A*32:01, HLA-A*68:02, HLA-B*15:01, HLA-B*35:01

MHC Class II

Rv	Epitope	Allele

Rv	Description	Epitope	Allele
Rv0978c	PE-PGRS family protein PE_PGRS17	RIGSAINTANTAAAA	HLA-DRB1*01:01, HLA-DRB1*03:01, HLA-DRB1*04:01, HLA-DRB1*04:05, HLA-DRB1*07:01, HLA-DRB1*08:02, HLA-DRB1*09:01, HLA-DRB1*11:01, HLA-DRB1*13:02, HLA-DRB3*01:01
Rv1271c	Conserved hypothetical secreted protein	KDEAFIAQMESIGVT	HLA-DRB1*01:01, HLA-DRB1*03:01, HLA-DRB1*04:01, HLA-DRB1*04:05, HLA-DRB1*07:01, HLA-DRB1*08:02, HLA-DRB1*09:01, HLA-DRB1*11:01, HLA-DRB1*12:01, HLA-DRB1*13:02, HLA-DRB1*15:01, HLA-DRB3*02:02, HLA-DRB4*01:01, HLA-DRB5*01:01
Rv1291c	Conserved hypothetical secreted protein	FTRRFAASMVGTTLT	HLA-DRB1*01:01, HLA-DRB1*03:01, HLA-DRB1*04:01, HLA-DRB1*04:05, HLA-DRB1*07:01, HLA-DRB1*08:02, HLA-DRB1*09:01, HLA-DRB1*11:01, HLA-DRB1*15:01, HLA-DRB5*01:01
Rv1329c	Probable ATP-dependent helicase DinG	ESVSMSVPELLAIAV	HLA-DRB1*01:01, HLA-DRB1*03:01, HLA-DRB1*04:01, HLA-DRB1*04:05, HLA-DRB1*09:01, HLA-DRB1*11:01, HLA-DRB1*13:02, HLA-DRB1*15:01, HLA-DRB3*02:02, HLA-DRB4*01:01, HLA-DRB5*01:01
Rv1396c	PE-PGRS family protein PE_PGRS25	IPTGFRGTVMSFLFA	HLA-DRB1*03:01, HLA-DRB1*04:01, HLA-DRB1*04:05, HLA-DRB1*07:01, HLA-DRB1*08:02, HLA-DRB1*09:01, HLA-DRB1*11:01, HLA-DRB1*15:01, HLA-DRB3*01:01, HLA-DRB5*01:01
		SFLFAQPEMLGAAAT	HLA-DRB1*01:01, HLA-DRB1*03:01, HLA-DRB1*04:01, HLA-DRB1*04:05, HLA-DRB1*07:01, HLA-DRB1*08:02, HLA-DRB1*09:01, HLA-DRB1*11:01, HLA-DRB1*12:01, HLA-DRB1*13:02, HLA-DRB3*02:02, HLA-DRB4*01:01, HLA-DRB5*01:01
Rv1424c	Possible membrane protein	KLVFPVLLGILTLLL	HLA-DRB1*01:01, HLA-DRB1*03:01, HLA-DRB1*04:01, HLA-DRB1*04:05, HLA-DRB1*08:02, HLA-DRB1*11:01, HLA-DRB1*12:01, HLA-DRB1*15:01, HLA-DRB4*01:01, HLA-DRB5*01:01
Rv1804c	Conserved protein	LMIGLAVPAHAGPSG	HLA-DRB1*01:01, HLA-DRB1*03:01, HLA-DRB1*04:01, HLA-DRB1*04:05, HLA-DRB1*08:02, HLA-DRB1*09:01, HLA-DRB1*11:01, HLA-DRB1*12:01, HLA-DRB1*15:01, HLA-DRB5*01:01
Rv1813c	Conserved hypothetical protein	GLGILLVPTVDAHLA	HLA-DRB1*01:01, HLA-DRB1*03:01, HLA-DRB1*04:01, HLA-DRB1*04:05, HLA-DRB1*07:01, HLA-DRB1*08:02, HLA-DRB1*09:01, HLA-DRB1*11:01, HLA-DRB1*12:01, HLA-DRB1*13:02, HLA-DRB1*15:01, HLA-DRB4*01:01, HLA-DRB5*01:01

Rv	Description	Epitope	Allele
Rv1926c	Immunogenic protein MPT63	ATFAAPVALAAYPIT	HLA-DRB1*01:01, HLA-DRB1*04:01, HLA-DRB1*04:05, HLA-DRB1*07:01, HLA-DRB1*08:02, HLA-DRB1*09:01, HLA-DRB1*11:01, HLA-DRB1*12:01, HLA-DRB1*15:01, HLA-DRB3*01:01, HLA-DRB5*01:01
Rv1984c	Probable cutinase precursor CFP21	IVGVVVATTLALVSA	HLA-DRB1*01:01, HLA-DRB1*03:01, HLA-DRB1*04:01, HLA-DRB1*04:05, HLA-DRB1*07:01, HLA-DRB1*08:02, HLA-DRB1*09:01, HLA-DRB1*11:01, HLA-DRB1*12:01, HLA-DRB1*13:02, HLA-DRB1*15:01, HLA-DRB4*01:01, HLA-DRB5*01:01
Rv2376c	Low molecular weight antigen CFP2	AGGPVVYQMQPVVFG	HLA-DRB1*01:01, HLA-DRB1*03:01, HLA-DRB1*04:01, HLA-DRB1*04:05, HLA-DRB1*07:01, HLA-DRB1*08:02, HLA-DRB1*09:01, HLA-DRB1*11:01, HLA-DRB1*12:01, HLA-DRB1*13:02, HLA-DRB1*15:01, HLA-DRB3*01:01, HLA-DRB3*02:02, HLA-DRB4*01:01, HLA-DRB5*01:01
Rv2878c	Secreted antigen MPT53 precursor	FCNAEAPSLSQVAAA	HLA-DRB1*01:01, HLA-DRB1*03:01, HLA-DRB1*04:01, HLA-DRB1*04:05, HLA-DRB1*07:01, HLA-DRB1*08:02, HLA-DRB1*09:01, HLA-DRB1*11:01, HLA-DRB1*13:02, HLA-DRB4*01:01, HLA-DRB5*01:01
Rv3036c	Conserved secreted protein TB22.2	GPKYMLDMTFPVDYP	HLA-DRB1*03:01, HLA-DRB1*04:01, HLA-DRB1*04:05, HLA-DRB1*07:01, HLA-DRB1*09:01, HLA-DRB1*11:01, HLA-DRB1*12:01, HLA-DRB1*13:02, HLA-DRB1*15:01, HLA-DRB3*01:01, HLA-DRB3*02:02, HLA-DRB5*01:01
Rv3212	Conserved alanine valine rich protein	WTSDARATISRPAAV	HLA-DRB1*01:01, HLA-DRB1*03:01, HLA-DRB1*04:01, HLA-DRB1*07:01, HLA-DRB1*08:02, HLA-DRB1*09:01, HLA-DRB1*11:01, HLA-DRB1*12:01, HLA-DRB1*13:02, HLA-DRB3*01:01
Rv3218	Conserved protein	VLIVNPTATATTPAG	HLA-DRB1*01:01, HLA-DRB1*03:01, HLA-DRB1*04:01, HLA-DRB1*08:02, HLA-DRB1*09:01, HLA-DRB1*11:01, HLA-DRB1*12:01, HLA-DRB1*13:02, HLA-DRB3*02:02, HLA-DRB5*01:01
		GTTPVRPVPAVAVVP	HLA-DRB1*01:01, HLA-DRB1*04:01, HLA-DRB1*07:01, HLA-DRB1*08:02, HLA-DRB1*09:01, HLA-DRB1*11:01, HLA-DRB1*13:02, HLA-DRB1*15:01, HLA-DRB3*01:01, HLA-DRB5*01:01
Rv3803c	Secreted MPT51/MPB51 antigen	APYENLMVPSPSMGR	HLA-DRB1*01:01, HLA-DRB1*03:01, HLA-DRB1*04:01, HLA-DRB1*04:05, HLA-DRB1*08:02, HLA-DRB1*11:01, HLA-DRB1*12:01, HLA-DRB1*15:01, HLA-DRB3*02:02, HLA-DRB5*01:01

Five 15-mer peptides predicted to display high binding affinity for MHC class II alleles were identified as follows: AGGPVVYQMQPVVFG (15 alleles), KDEAFIAQMESIGVT (14 alleles), SFLFAQPEMLGAAAT (13 alleles), GLGILLVPTVDAHLA (13 alleles) and IVGVVVATTLALVSA (13 alleles). Five T-cell epitope regions from 5 proteins, Rv1329c, Rv1424c, Rv1926c, Rv2376c and Rv3036c, were observed to overlap between MHC classes I and II alleles (Fig. 4.1 and Table 4.2).

4.4.4.2 Non-classical MHC class 1 HLA-E epitopes

The non-classical HLA-E play a key role in the modulation of both the innate and adaptive immune system. HLA-E is known to present only a small pool of peptides to Cytotoxic T-cells as antigen presentation is not its primary role (Felício *et al.*, 2014). Of the 13 HLA-E alleles, only two, HLA-E*01:01 and HLA-E*01:03 contribute to the immune function, and are frequently found in worldwide populations (Felício *et al.*, 2014).

The present study investigated whether any of the 97 B-cell epitopes were recognized by HLA-E*01:01 and/or HLA-E*01:03 for presentation to Cytotoxic T-cell. The rationale for this approach was that down-regulation of HLA-A and HLA-B molecules expression during HIV infection decreased the antigen presentation capabilities of classical MHC class I (Cohen *et al.*, 1999). In addition, the decreased $CD4^+$ cell counts result in impaired functioning of *Mtb*-specific $CD4^+$ T-cells (MHC class II) which are capable of containing bacterial replication (Riou *et al.*, 2016). Interestingly, in contrast, HLA-E is resistant to HIV-mediated down-regulation. Therefore, HLA-E dependent antigen presentation is not affected or down-regulated by HIV-infection (Cohen *et al.*, 1999). Thus, targeting *Mtb*-specific HLA-E restricted immunity by vaccination may be a novel and advantageous approach, especially, in high HIV endemic countries like South Africa.

Table 4. 3: List of identified HLA-E*01:01 and *01:03 binding 9-mer peptide epitopes.

Allele	Rv	Description	Peptide	Percentile rank	Antigenicity
HLA-E*01:01	Rv0361	Conserved membrane protein	TTPPRMPTG	0.15	-0.1113
HLA-E*01:01	Rv1291c	Conserved hypothetical secreted protein	TLTAATLGL	0.15	0.6772
HLA-E*01:01	Rv0361	Conserved membrane protein	RMPTGMPPK	0.2	0.6896
HLA-E*01:01	Rv1435c	Conserved proline, glycine, valine-rich secreted protein	NMSPAAPGR	0.2	0.4507
HLA-E*01:01	Rv1984c	Probable cutinase precursor CFP21	VVVATTLAL	0.2	0.4859
HLA-E*01:01	**Rv3682**	penicillin-binding protein 1A/1B PonA2	**QLPPTDPRY**	**0.2**	0.4994
HLA-E*01:01	Rv1291c	Conserved hypothetical secreted protein	ITPPSAARA	0.3	0.1337
HLA-E*01:01	Rv1424c	Possible membrane protein	RPPAEKLVF	0.4	-0.6094
HLA-E*01:03	**Rv3682**	penicillin-binding protein 1A/1B PonA2	**QLPPTDPRY**	**0.6**	0.4994
HLA-E*01:01	Rv0361	Conserved membrane protein	NQAPTTPPR	0.9	0.0878
HLA-E*01:01	Rv0978c	PE-PGRS family, PE_PGRS17	AAAATTQVL	1.0	0.3291

In this study, 10 epitopes binding HLA-E*01:01 alleles with percentile score ranging from 0.15 to 1.0 were identified (Table 4.3). One epitope [QLPPTDPRY (Rv3682)] was found to bind to HLA-E*01:03 and HLA-E*01:01 allele. This may be due to the lack of allelic variation in the peptide binding capability of HLA-E, as a similar peptide is presented by both HLA-E variants (Strong *et al.*, 2003). The peptide QLPPTDPRY that binds both HLA-E alleles, is encoded by Rv3682, a bifunctional membrane-associated penicillin-binding protein (PonA2) (Sauvage *et al.*, 2008). PonA2 is involved in the final stages of the synthesis of cross-linked peptidoglycan, the major component of the bacterial cell wall (Kieser *et al.*, 2015). The cell wall biosynthesis is essential for growth, cell division and cellular structure maintenance (Dover *et al.*, 2007). Therefore, inhibition of penicillin-binding protein such as PonA2 could lead to irregularities in cell wall structure and eventual cell death and lysis. The *ponA2* gene is expressed during *in vivo* growth and is involved in stationary-phase survival under non-replicating conditions (Talaat *et al.*, 2004, Rengarajan *et al.*, 2005). The highly ranked epitopes were obtained from the

conserved membrane Rv0361 [TTPPRMPTG] and conserved secreted Rv1291c [TLTAATLGL] proteins. The epitope encoded by PE_PGRS17 [AAAATTQVL] displayed the lowest score of 1.0.

4.4.5 Population Coverage Prediction

Determining the world population coverage for HLA alleles is a prerequisite for the design of an effective epitope-based vaccine to target all major ethnic groups. In the current study, the IEDB world population coverage analysis for individual epitopes identified the following 5 classical MHC class I epitopes: RPPAEKLVFPVLLGILTLLL (Rv1424c), AGGPVVYQMQPVVFGAPLPL (Rv2376c) and HVHASGPKYMLDMTFPVDYP (Rv3036c) each with a coverage of 39.08%, and TMTDTVGQVVLGWKVSDLKS (Rv1926c) and GNASATSGDMSSMTRIAKPL (Rv1884c) with a coverage of 17.34% and 16.81%, respectively. When all five epitopes were pooled together, a world population coverage of 63.79% was achieved. The other three MHC class I epitopes each displayed an insignificant population coverage of 0.97%. Amongst the 18 MHC class II epitopes, 16 had a population coverage of 11.53% when assessed individually and two epitopes, IPTGFRGTVMSFLFA (Rv1396c) and GPKYMLDMTFPVDYP (Rv3036c) each had a population coverage of 17.84%.

Further investigation of the population coverage of the identified non-classical HLA-E epitopes revealed that each of the HLA-E*01:01 epitopes had a population coverage of 74.79%. This was similar to that of the HLA-E*01:03 epitope QLPPTDPRY (Rv3682) at 71.7%. The combination of HLA-E*01:03 epitope with any one of the HLA-E*01:01 epitopes resulted in a combined population coverage of 99.88%.

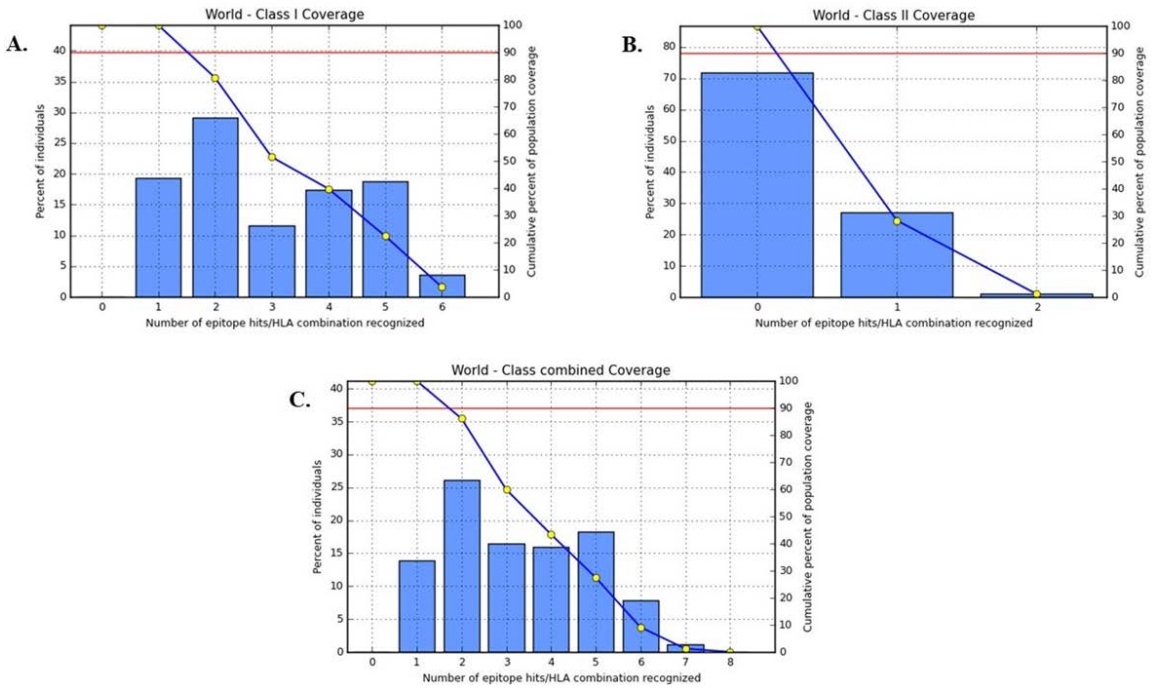

Fig. 4. 2: Population coverage analysis for the top predicted epitopes based on the HLA interaction. The world populations were assessed for the proposed coverage by (A) MHC class I, (B) MHC class II and (C) the combined prediction for both of the MHC, the number 1 bar for all the analyses represents out-predicted epitope. Notes: in the graphs, the line (-o-) represents the cumulative percentage of population coverage of the epitopes, whilst the bars represent the population coverage for each epitope.

We identified 9 epitopes with significant global population coverage. These included the 5 classical MHC class I epitopes encoded by Rv1424c, Rv1884c, Rv1926c, Rv2376 and Rv3036c, and 2 non-classical HLA-E epitopes encoded by Rv0361 and Rv3682. The 2 MHC class II epitopes selected were encoded by Rv1271c and Rv3803c. The potential world population coverage of the combined epitopes if used in a poly-peptide vaccine format was calculated to be 99.97% (Fig. 4.2).

4.5 CONCLUSION

The main limitation of the present study is that T-cell epitopes were identified *in silico* using bioinformatics tools. However, in mitigation of this limitation, B-cell epitopes were experimentally selected and confirmed with *in silico* B-cell epitope prediction tools. Therefore, we expect T-cell epitopes to be immunogenic and induce cell-mediated immunity *in vitro* and *in vivo*.

Future studies will include the design of a recombinant multi-epitope TB vaccine candidates by joining different DNA fragments encoding selected epitopes into one gene fragment, cloning and expression as recombinant polypeptide proteins in *E. coli*. The different recombinant polypeptide proteins will be used to immunize mice and the humoral and cellular immune responses will be analysed in order to identify the most promising candidate vaccine with potential to provide immunity to *Mtb* infection. The promising TB vaccine will be tested by infecting the immunized mice (vaccinated) with virulent *Mtb* strain to determine it efficacy. The proposed approach is expected to lead to the development of new TB vaccine with added advantage of not only provoke the T-cell immune response but also the B-cell mediated antibody response that will trigger the killer T-cells to eliminate *Mtb* and infected cells.

There is an urgent need for a new vaccine that will provide protection against TB in adolescents and adults. The present study identified B- and T-cell epitopes with potential to elicit antibody and cellular mediated immunity against TB in global wide populations. Therefore, future studies should investigate the *in vivo* capacity of the HLA-E binding epitope QLPPTDPRY to induce $CD8^+$ T-cell response, especially in HIV/AIDS patients. The efficiency of the identified epitopes as a multivalent *Mtb* poly-peptide vaccine should also be evaluated.

4.6 REFERENCES

Achkar, J.M., Chan, J. and Casadevall, A. (2014). Role of B-cells and Antibodies in Acquired Immunity against *Mycobacterium tuberculosis*. *Cold Spring Harbor Perspectives in Medicine***,** a018432.

Andersen, P. and Doherty, T.M. (2005). Opinion: The success and failure of BCG--implications for a novel tuberculosis vaccine. *Nature reviews. Microbiology* **3,** 656.

Bansal, K., Elluru, S.R., Narayana, Y., Chaturvedi, R., Patil, S.A., Kaveri, S.V., *et al.* (2010). PE_PGRS Antigens of *Mycobacterium tuberculosis* Induce Maturation and Activation of Human Dendritic Cells. *The Journal of Immunology* **184,** 3495-3504.

Barh, D., Misra, A.N., Kumar, A. and Vasco, A. (2010). A novel strategy of epitope design in *Neisseria gonorrhoeae*. *Bioinformation* **5,** 77.

Boggiano, C., Eichelberg, K., Ramachandra, L., Shea, J., Ramakrishnan, L., Behar, S., *et al.* (2017). "The Impact of *Mycobacterium tuberculosis* Immune Evasion on Protective Immunity: Implications for TB Vaccine Design" – Meeting report. *Vaccine* **35,** 3433-3440.

Bui, H.H., Sidney, J., Dinh, K., Southwood, S., Newman, M.J. and Sette, A. (2006). Predicting population coverage of T-cell epitope-based diagnostics and vaccines. *BMC Bioinformatics* **7,** 153.

Carlini, F., Ferreira, V., Buhler, S., Tous, A., Eliaou, J.-F., René, C., *et al.* (2016). Association of HLA-A and Non-Classical HLA Class I Alleles. *PLOS ONE* **11,** e0163570.

Chen, J., Liu, H., Yang, J. and Chou, K.-C. (2007). Prediction of linear B-cell epitopes using amino acid pair antigenicity scale. *Amino acids* **33,** 423-428.

Cohen, G.B., Gandhi, R.T., Davis, D.M., Mandelboim, O., Chen, B.K., Strominger, J.L. and Baltimore, D. (1999). The selective downregulation of class I major histocompatibility

complex proteins by HIV-1 protects HIV-infected cells from NK cells. *Immunity* **10,** 661-671.

Colditz, G.A., Berkey, C.S., Mosteller, F., Brewer, T.F., Wilson, M.E., Burdick, E. and Fineberg, H.V. (1995). The efficacy of bacillus Calmette-Guerin vaccination of newborns and infants in the prevention of tuberculosis: meta-analyses of the published literature. *Pediatrics* **96,** 29-35.

Dover, L.G., Alderwick, L.J., Brown, A.K., Futterer, K. and Besra, G.S. (2007). Regulation of cell wall synthesis and growth. *Current molecular medicine* **7,** 247-276.

Doytchinova, I.A. and Flower, D.R. (2007). VaxiJen: a server for prediction of protective antigens, tumour antigens and subunit vaccines. *BMC Bioinformatics* **8,** 4.

EL-Manzalawy, Y., Dobbs, D. and Honavar, V. (2008). Predicting linear B-cell epitopes using string kernels. *Journal of molecular recognition* **21,** 243-255.

Felício, L.P., Porto, I.O.P., Mendes-Junior, C.T., Veiga-Castelli, L.C., Santos, K.E., Vianello-Brondani, R.P., *et al.* (2014). Worldwide HLA-E nucleotide and haplotype variability reveals a conserved gene for coding and 3′ untranslated regions. *Tissue Antigens* **83,** 82-93.

Gardy, J.L., Laird, M.R., Chen, F., Rey, S., Walsh, C., Ester, M. and Brinkman, F.S. (2004). PSORTb v. 2.0: expanded prediction of bacterial protein subcellular localization and insights gained from comparative proteome analysis. *Bioinformatics* **21,** 617-623.

Germain, R.N. (1994). MHC-dependent antigen processing and peptide presentation: providing ligands for T lymphocyte activation. *Cell* **76,** 287-299.

Guermonprez, P., Valladeau, J., Zitvogel, L., Théry, C. and Amigorena, S. (2002). Antigen Presentation and T-cell Stimulation by Dendritic Cells. *Annual Review of Immunology* **20,** 621-667.

Hoof, I., Peters, B., Sidney, J., Pedersen, L.E., Sette, A., Lund, O., *et al.* (2009). NetMHCpan, a method for MHC class I binding prediction beyond humans. *Immunogenetics* **61,** 1.

Horwitz, M.A., Lee, B.W., Dillon, B.J. and Harth, G. (1995). Protective immunity against tuberculosis induced by vaccination with major extracellular proteins of *Mycobacterium tuberculosis*. *Proceedings of the National Academy of Sciences* **92,** 1530-1534.

Jacobs, A.J., Mongkolsapaya, J., Screaton, G.R., McShane, H. and Wilkinson, R.J. (2016). Antibodies and tuberculosis. *Tuberculosis* **101,** 102-113.

Käll, L., Krogh, A. and Sonnhammer, E.L. (2007). Advantages of combined transmembrane topology and signal peptide prediction—the Phobius web server. *Nucleic acids research* **35,** W429-W432.

Karosiene, E., Rasmussen, M., Blicher, T., Lund, O., Buus, S. and Nielsen, M. (2013). NetMHCIIpan-3.0, a common pan-specific MHC class II prediction method including all three human MHC class II isotypes, HLA-DR, HLA-DP and HLA-DQ. *Immunogenetics* **65,** 711-724.

Kaufmann, S.H.E., Weiner, J. and von Reyn, C.F. (2017). Novel approaches to tuberculosis vaccine development. *International Journal of Infectious Diseases* **56,** 263-267.

Kieser, K.J., Baranowski, C., Chao, M.C., Long, J.E., Sassetti, C.M., Waldor, M.K., *et al.* (2015). Peptidoglycan synthesis in *Mycobacterium tuberculosis* is organized into networks with varying drug susceptibility. *Proceedings of the National Academy of Sciences* **112,** 13087-13092.

Klopper, M., Warren, R.M., Hayes, C., van Pittius, N.C.G., Streicher, E.M., Müller, B., *et al.* (2013). Emergence and spread of extensively and totally drug-resistant tuberculosis, South Africa. *Emerging infectious diseases* **19,** 449.

Kraemer, T., Celik, A.A., Huyton, T., Kunze-Schumacher, H., Blasczyk, R., Bade-D, *et al.* (2015). HLA-E: Presentation of a Broader Peptide Repertoire Impacts the Cellular

Immune Response - Implications on HSCT Outcome. *Stem Cells International* **2015,** 12.

Kumar, P., Amara, R.R., Challu, V.K., Chadda, V.K. and Satchidanandam, V. (2003). The Apa protein of *Mycobacterium tuberculosis* stimulates gamma interferon-secreting CD4+ and CD8+ T-cells from purified protein derivative-positive individuals and affords protection in a guinea pig model. *Infection and immunity* **71,** 1929-1937.

Kunnath-Velayudhan, S., Davidow, A.L., Wang, H.Y., Molina, D.M., Huynh, V.T., Salamon, H., *et al.* (2012). Proteome-Scale Antibody Responses and Outcome of *Mycobacterium tuberculosis* Infection in Nonhuman Primates and in Tuberculosis Patients. *J. Infect. Dis.* **206,** 697-705.

Kunnath-Velayudhan, S., Salamon, H., Wang, H.-Y., Davidow, A.L., Molina, D.M., Huynh, V.T., *et al.* (2010). Dynamic antibody responses to the *Mycobacterium tuberculosis* proteome. *Proceedings of the National Academy of Sciences* **107,** 14703-14708.

Li, W., Deng, G., Li, M., Zeng, J., Zhao, L., Liu, X., Wanga, Y. (2014). A recombinant adenovirus expressing CFP10, ESAT6, Ag85A and Ag85B of *Mycobacterium tuberculosis* elicits strong antigen-specific immune responses in mice. *Molecular Immunology* **62**, 86-95

Lyashchenko, K., Colangeli, R., Houde, M., Al Jahdali, H., Menzies, D. and Gennaro, M.L. (1998). Heterogeneous antibody responses in tuberculosis. *Infection and Immunity* **66,** 3936-3940.

Monterrubio-Lopez, G.P., Gonzalez, Y.M.J.A. and Ribas-Aparicio, R.M. (2015). Identification of Novel Potential Vaccine Candidates against Tuberculosis Based on Reverse Vaccinology. *BioMed research international* **2015,** 16.

Narayana, Y., Joshi, B., Katoch, V.M., Mishra, K.C. and Balaji, K.N. (2007). Differential B-Cell Responses Are Induced by *Mycobacterium tuberculosis* PE Antigens Rv1169c, Rv0978c, and Rv1818c. *Clinical and Vaccine Immunology* **14,** 1334-1341.

Nielsen, M., Lundegaard, C., Blicher, T., Lamberth, K., Harndahl, M., Justesen, S., *et al.* (2007). NetMHCpan, a method for quantitative predictions of peptide binding to any HLA-A and-B locus protein of known sequence. *PloS one* **2,** e796.

Nielsen, M., Lundegaard, C., Blicher, T., Peters, B., Sette, A., Justesen, S., *et al.* (2008). Quantitative predictions of peptide binding to any HLA-DR molecule of known sequence: NetMHCIIpan. *PLoS computational biology* **4,** e1000107.

Orr, M.T., Ireton, G.C., Beebe, E.A., Huang, P.D., Reese, V.A., Argilla, D., *et al.*, (2014). Immune subdominant antigens as vaccine candidates against *Mycobacterium tuberculosis*. *Journal of Immunology* **193,** 2911-2918.

Paul, S., Sidney, J., Sette, A. and Peters, B. (2016). TepiTool: a pipeline for computational prediction of T-cell epitope candidates. *Current Protocols in Immunology***,** 18.19. 11-18.19. 24.

Peters, B. and Sette, A. (2005). Generating quantitative models describing the sequence specificity of biological processes with the stabilized matrix method. *BMC bioinformatics* **6,** 132.

Pillay, M. and Sturm, A.W. (2007). Evolution of the extensively drug-resistant F15/LAM4/KZN strain of *Mycobacterium tuberculosis* in KwaZulu-Natal, South Africa. *Clinical infectious diseases* **45,** 1409-1414.

Plotkin, S.A. and Plotkin, S.A. (2008). Correlates of Vaccine-Induced Immunity. *Clinical Infectious Diseases* **47,** 401-409.

Prados-Rosales, R., Carreño, L.J., Batista-Gonzalez, A., Baena, A., Venkataswamy, M.M., Xu, J., *et al.* (2014). Mycobacterial membrane vesicles administered systemically in mice

induce a protective immune response to surface compartments of *Mycobacterium tuberculosis*. *MBio* **5,** e01921-01914.

Puech, V., Guilhot, C., Perez, E., Tropis, M., Armitige, L.Y., Gicquel, B. and Daffé, M. (2002). Evidence for a partial redundancy of the fibronectin-binding proteins for the transfer of mycoloyl residues onto the cell wall arabinogalactan termini of *Mycobacterium tuberculosis*. *Mol. Microbiol.* **44,** 1109-1122.

Rashid, M.I., Naz, A., Ali, A. and Andleeb, S. (2017). Prediction of vaccine candidates against *Pseudomonas aeruginosa*: An integrated genomics and proteomics approach. *Genomics* **109,** 274-283.

Rengarajan, J., Bloom, B.R. and Rubin, E.J. (2005). Genome-wide requirements for *Mycobacterium tuberculosis* adaptation and survival in macrophages. *Proc Natl Acad Sci U S A* **102,** 8327-8332.

Riou, C., Strickland, N., Soares, A.P., Corleis, B., Kwon, D.S., Wherry, E.J., *et al.* (2016). HIV Skews the Lineage-Defining Transcriptional Profile of *Mycobacterium tuberculosis*–Specific CD4+ T-cells. *The Journal of Immunology* **196,** 3006-3018.

Romain, F., Horn, C., Pescher, P., Namane, A., Riviere, M., Puzo, G., *et al.* (1999). Deglycosylation of the 45/47-kilodalton antigen complex of *Mycobacterium tuberculosis* decreases its capacity to elicit in vivo or in vitro cellular immune responses. *Infection and immunity* **67,** 5567-5572.

Romano, M., Aryan, E., Korf, H., Bruffaerts, N., Franken, C.L., Ottenhoff, T.H. and Huygen, K. (2012). Potential of *Mycobacterium tuberculosis* resuscitation-promoting factors as antigens in novel tuberculosis sub-unit vaccines. *Microbes Infect* **14,** 86-95.

Sachdeva, G., Kumar, K., Jain, P. and Ramachandran, S. (2004). SPAAN: a software program for prediction of adhesins and adhesin-like proteins using neural networks. *Bioinformatics* **21,** 483-491.

Sakula, A. (1983). BCG: who were Calmette and Guerin? *Thorax* **38,** 806.

Sauvage, E., Kerff, F., Terrak, M., Ayala, J.A. and Charlier, P. (2008). The penicillin-binding proteins: structure and role in peptidoglycan biosynthesis. *FEMS microbiology reviews* **32,** 234-258.

Strong, R.K., Holmes, M.A., Li, P., Braun, L., Lee, N. and Geraghty, D.E. (2003). HLA-E allelic variants. Correlating differential expression, peptide affinities, crystal structures, and thermal stabilities. *J Biol Chem* **278,** 5082-5090.

Talaat, A.M., Lyons, R., Howard, S.T. and Johnston, S.A. (2004). The temporal expression profile of *Mycobacterium tuberculosis* infection in mice. *Proceedings of the National Academy of Sciences of the United States of America* **101,** 4602-4607.

Wang, P., Sidney, J., Dow, C., Mothe, B., Sette, A. and Peters, B. (2008). A systematic assessment of MHC class II peptide binding predictions and evaluation of a consensus approach. *PLoS computational biology* **4,** e1000048.

Wang, P., Sidney, J., Kim, Y., Sette, A., Lund, O., Nielsen, M. and Peters, B. (2010). Peptide binding predictions for HLA DR, DP and DQ molecules. *BMC bioinformatics* **11,** 568.

WHO (2017). Global Tuberculosis Report 2017 *World Health Organization.*

Wilson, R.A., Maughan, W.N., Kremer, L., Besra, G.S. and Futterer, K. (2004). The structure of *Mycobacterium tuberculosis* MPT51 (FbpC1) defines a new family of non-catalytic alpha/beta hydrolases. *J Mol Biol* **335,** 519-530.

Yeremeev, V.V., Kondratieva, T.K., Rubakova, E.I., Petrovskaya, S.N., Kazarian, K.A., Telkov, M.V., *et al.* (2003). Proteins of the Rpf family: immune cell reactivity and vaccination efficacy against tuberculosis in mice. *Infect Immun* **71,** 4789-4794.

Zvi, A., Ariel, N., Fulkerson, J., Sadoff, J.C. and Shafferman, A. (2008). Whole genome identification of *Mycobacterium tuberculosis* vaccine candidates by comprehensive data mining and bioinformatic analyses. *BMC Med Genomics* **1,** 18.

CHAPTER 5: General Discussion, Recommendations and Conclusion

5.1 GENERAL DISCUSSION AND RECOMMENDATIONS

The serological diagnosis of infectious diseases can be achieved by the detection of either antibodies or antigens in patient blood samples (Khurshid *et al.*, 2017). However, in 2011 the WHO suspended the use of serological tests based on antibody detection for TB diagnosis due to the diverse antibody response by patients, as well as significant inconsistencies in their performance due to the lack of a single robust antigen (Bekmurzayeva *et al.*, 2013). Furthermore, the emergence of MDR/XDR-TB strains has created a public health crisis and a global health security risk (WHO, 2014). Therefore, there is an urgent need for innovative strategies and interventions to tackle the growing epidemic of TB. Consequently, the current study used the phage display method to interrogate the *M. tuberculosis* (*Mtb*) secretome in an effort to identify new pathogen-derived antigenic biomarkers for the potential development of rapid diagnostics, effective drugs and vaccines.

The *Mtb* secretome possesses key characteristics that are essential when selecting candidate vaccines, targets for therapeutic development, and biomarkers for diagnosing TB. The bacterial secretome comprises surface-membrane proteins, transmembrane proteins and proteins secreted into the extracellular environment. The *Mtb* secretome is responsible for the regulation of many physiological processes and its composition changes in response to environmental stimuli (Caccia *et al.*, 2013). Thus, the ability of *Mtb* to secrete and export certain proteins outside of the bacterial cell is directly linked to its pathogenicity (Majlessi *et al.*, 2015). Hence, there is an increasing interest in the *Mtb* secretome as it plays an important role in the host-pathogen interaction and the proteins involved are therefore, potential biomarkers and therapeutic targets against TB. Therefore, the current study focused on profiling the

mycobacterial secretome in terms of localization, disease state-specific expression, and immunogenicity properties of potential TB biomarkers.

Several approaches have been previously used to identify potential antigens and epitopes for the development of new diagnostic tools, vaccines and drugs. The bioinformatics approach used with whole genome sequencing identifies bacterial secretome proteins specific to the secretion pathway (Song *et al.*, 2008, Vizcaino *et al.*, 2010, Xu *et al.*, 2014). This approach generated a list of secreted protein candidates which were later validated using targeted secretion assays based on reporter fusion systems (McDonough *et al.*, 2005, McCann *et al.*, 2011). The basic principle behind the study of the entire *Mtb* genome (proteome) is to first eliminate all general "housekeeping" genes (proteins) and then to target the conserved, pathogen-specific proteins involved in host-pathogen interaction (Perkowski *et al.*, 2017).

An alternative culture filtrate based proteomics approach or phage display, does not require the secretion pathway information (Maffei *et al.*, 2017). Therefore, a phage display system designed to investigate the bacterial secretome (Jankovic *et al.*, 2007, Liu *et al.*, 2011) was utilized in the present study. Chapter 2 of this thesis described the construction of a *Mtb* secretome library. The analysis and characterization of randomly selected clones from the library revealed the presence of complete and partial signal peptide sequences in the majority of *Mtb* protein peptides. Ng *et al.*, found that a cut-off of between 3 and 5 amino acids before the cleavage site is sufficient for peptidase binding and substrate protein processing during secretion (Ng *et al.*, 2009). This partly explained the identification of peptides with a partial signal peptide sequence and translocation through their respective secretion pathways. Therefore, the signal peptide sequence was demonstrated to be crucial for proper processing of the majority of the analysed *Mtb* protein fragments.

Integral membrane proteins have been reported to be particularly difficult to identify by commonly used mass spectrometry methods due to interference by detergents (Molloy *et al.*,

2000). This difficulty was addressed by the use of the phage display system in the present study that facilitated the selective incorporation of integral membrane and transmembrane protein fragments into the phage surface during packaging and assembly. It is assumed that the identified integral membrane or transmembrane protein fragments were likely to have been delivered to the membrane site via insertase which is responsible for insertion of transmembrane proteins into the cell membrane (Schneewind *et al.*, 2014). An extensive literature search of library clones with signal peptide sequences, as well as integral and transmembrane proteins, facilitated the identification of unique extracellular proteins with known and unknown function(s), and those that are essential for *in vivo* growth and survival of *Mtb* during infection. These proteins, viz., Nrp, PssA, MmpL5, SirA, GatB, EspA, TopA, EccCa1, Rv1634 and Rv3103c could represent potential candidates for the development of diagnostic tools, new drugs and vaccines, since they are directly in contact with the infected host and will be easily accessible if they are targeted for such purposes. Thus, the sequence analysis of randomly selected library clones validated the role of phage display in the study of the *Mtb* secretome and enabled the successful identification of integral and transmembrane proteins.

Phage display is considered to be one of the most powerful techniques used to identify proteins that bind to targets of interest (Mullen *et al.*, 2006). It is an appropriate tool for the investigation of the immunological spectrum during *Mtb* infection and identification of possible TB antigens that elicit humoral and cellular immune responses. TB patients produce antibodies against surface membrane and extracellular proteins secreted by *Mtb* during different stages of infection (Kunnath-Velayudhan *et al.*, 2013). Chapter 3 described the identification of *Mtb* peptides recognized by these antibodies via screening of the *Mtb* phage secretome library against sera from active- and latent-TB patients. The phage DNA was sequenced for the identification of novel disease state-specific biomarkers. These consisted mainly of secreted and cell wall-

associated proteins, suggesting that the immunodominant antigens of *Mtb* are proteins that are secreted by the pathogen during infection. As expected, the analysis of selected latent-TB specific proteins suggested that the humoral immune response targets a different antigenic repertoire compared to immune responses in active-TB infected individuals. This difference may be due to higher bacterial burdens in active-TB than in latently infected individuals (Kunnath-Velayudhan *et al.*, 2013). Therefore, the approach used in the present study allowed for the profiling of the *Mtb* secretome during latent infection and TB disease. The identified *Mtb* proteins may represent virulence factors expressed to advance the disease agenda during infection, and therefore, are potential TB candidate biomarkers.

The significance of the discovery of disease state-specific biomarkers lies in their potential use to develop tools that can differentiate active- from latent-TB infection and possibly, monitor TB disease progression from latent- to active-TB, especially in high risk groups such as HIV infected patients. Interestingly, 23 proteins were commonly induced by both latent-TB and active-TB patients, including those involved in resuscitation of dormant *Mtb* cells to actively growing bacteria. Therefore, some of these proteins may be involved in the progression from latent-TB to active-TB state. Future studies should investigate the role of the identified proteins in TB pathogenesis for consideration in the design of new drugs that can be used to treat MDR/XDR-TB. The role of the transition proteins, especially the hypothetical proteins and those of unknown function(s), also needs to be established. Interestingly, Rv3310 was identified in Chapter 2 by random screening of library members and also in Chapter 3 during immunoscreening against latent-TB patients' sera. Rv3310 is the only acid phosphatase (SapM) produced by *Mtb*. It is involved in maintaining the pH levels high within the phagosome, blocking the phagosome-lysosome fusion, therefore participate in the intracellular survival, replication and persistence of the pathogen within the macrophage (Puri *et al.*, 2013). SapM is an ideal target for the design of new TB drug that will interfere with it normal function and

allow antimicrobial activities of the phagosome to eliminate *Mtb*. The newly identified biomarkers should also be tested for use in multiple antigen based serological assays to improve the sensitivity and specificity in TB diagnosis.

Phage display library screening against patients' sera can lead to the identification of potential candidate vaccines against specific diseases (Mullen *et al.*, 2006). In Chapter 4, the suitability of the identified active- and latent-TB proteins as candidate vaccines was investigated. Since the proteins were selected as phage displayed peptide fragments and not as complete length proteins, it was assumed that the fragments were immunogenic, and therefore, contained an epitope(s). An epitope is that part of an antigen that specifically binds an antibody (B-cell epitope) or a T-cell receptor (T-cell epitope) (Wang *et al.*, 2004). The *Mtb* phage secretome library generated in the current study is a significant source of TB candidate vaccines and biomarkers/epitopes, as it consisted of extracellular and surface exposed proteins, which is one of the requirements for a good vaccine epitope. The selected active-TB and latent-TB phage displayed *Mtb* protein fragments and the identified B-cell epitopes were therefore analysed using bioinformatics tools. The antibody-mediated response (B-cells) plays only a supportive role to the cell-mediated (T-cell) immunity which provides protection against *Mtb* infection (Kerns *et al.*, 2014). In order to achieve both antibody-mediated and cell-mediated immunity against TB, the presence of T-cell epitopes within the B-cell epitopes of highly antigenic protein biomarkers was assessed. T-cell epitopes that can provide more than 99% of global population coverage, when incorporated into a new multi-epitope polypeptide TB vaccine, were identified. This included a novel epitope QLPPTDPRY (Rv3682) with binding affinity to HLA-E alleles which are not down-regulated by HIV co-infection (Cohen *et al.*, 1999, McMurtrey *et al.*, 2017) and could potentially provide HIV patients with immunity to *Mtb* infection or *Mtb* reactivation. Our study had several limitations. A previous study showed that only 10% of the whole *Mtb* proteome was immunogenic (Kunnath-Velayudhan *et al.*, 2013). In the current study, the

complete *Mtb* phage secretome library could not be sequenced due to the scope of the study; thus resulting in an incomplete secretome repertoire analysis. Nevertheless, the analysis of 120 secretome library members predicted that more than 95% of these were extracellular proteins. Similarly, not all enriched active-TB and latent-TB phage displayed *Mtb* protein fragments could be sequenced. In addition, since antibody response varies from patient to patient, it would have been interesting to screen the library against individual serum samples to identify proteins (epitopes) commonly recognized by humoral immune response of most, if not all, study participants. Such epitopes can provide consistency, increased specificity, and sensitivity for TB diagnosis. Finally, the T-cell epitopes for the development of TB vaccines were identified only *in silico*, based on B-cell epitopes. Since phages are regarded as natural immunostimulators (adjuvant) and recognized for eliciting a powerful and effective immune response (Haq *et al.*, 2012), there is a need to investigate the use of phage particles displaying *Mtb* epitopes on their surface coat protein for immunization for protective immunity, thus acting as a vaccine delivery vehicle.

5.2 CONCLUSION

To the best of our knowledge, this is the first study to investigate an XDR *Mtb* clinical strain secretome using the phage display technique and immunoscreening of phage library against ATB and LTBI patients' sera to identify antigens that could be relevant for diagnosis of ATB and LTBI, respectively. The findings of this study confirm that the *Mtb* phage secretome library generated is an extraordinary source of clinically significant TB biomarker epitopes. The *Mtb* secretome comprise mainly immunogenic proteins expressed during infection that are the primary mycobacterial structures that interact with the host to facilitate invasion, virulence and survival inside host. It is therefore crucial to identify accurately the sequential expression of

these proteins as well as their association with the specific disease stage/state for potential exploitation in combating TB. The present study contributes new knowledge of the *Mtb* secretome profile during infection, providing some insights into cellular processes involved in TB pathogenesis. This knowledge will potentially accelerate research toward new treatments, diagnostic tools and vaccines for the benefit of the most vulnerable against TB.

5.3 REFERENCES

Bekmurzayeva, A., Sypabekova, M. and Kanayeva, D. (2013). Tuberculosis diagnosis using immunodominant, secreted antigens of *Mycobacterium tuberculosis*. *Tuberculosis* **93,** 381-288.

Caccia, D., Dugo, M., Callari, M. and Bongarzone, I. (2013). Bioinformatics tools for secretome analysis. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics* **1834,** 2442-2453.

Cohen, G.B., Gandhi, R.T., Davis, D.M., Mandelboim, O., Chen, B.K., Strominger, J.L. and Baltimore, D. (1999). The selective downregulation of class I major histocompatibility complex proteins by HIV-1 protects HIV-infected cells from NK cells. *Immunity* **10,** 661-671.

Haq, I.U., Chaudhry, W.N., Akhtar, M.N., Andleeb, S. and Qadri, I. (2012). Bacteriophages and their implications on future biotechnology: a review. *Virology journal* **9,** 9.

Jankovic, D., Collett, M.A., Lubbers, M.W. and Rakonjac, J. (2007). Direct selection and phage display of a Gram-positive secretome. *Genome biology* **8,** R266.

Kerns, P.W., Ackhart, D.F., Basaraba, R.J., Leid, J.G. and Shirtliff, M.E. (2014) *Mycobacterium tuberculosis* pellicles express unique proteins recognized by the host humoral response, pp. 347-358.

Khurshid, S., Afzal, M., Khalid, R., Akhtar, M.W. and Qazi, M.H. (2017). Potential of multi-component antigens for tuberculosis diagnosis. *Biologicals* **48,** 109-113.

Kunnath-Velayudhan, S. and Porcelli, S.A. (2013). Recent advances in defining the immunoproteome of *Mycobacterium tuberculosis*. *Frontiers in Immunology* **4**.

Liu, S., Han, W., Sun, C., Lei, L., Feng, X., Yan, S., *et al.* (2011). Subtractive screening with the *Mycobacterium tuberculosis* surface protein phage display library. *Tuberculosis* **91,** 579-586.

Maffei, B., Francetic, O. and Subtil, A. (2017). Tracking Proteins Secreted by Bacteria: What's in the Toolbox? *Frontiers in Cellular and Infection Microbiology* **7**.

Majlessi, L., Prados-Rosales, R., Casadevall, A. and Brosch, R. (2015). Release of mycobacterial antigens. *Immunological Reviews* **264,** 25-45.

McCann, J.R., McDonough, J.A., Sullivan, J.T., Feltcher, M.E. and Braunstein, M. (2011). Genome-Wide Identification of *Mycobacterium tuberculosis* Exported Proteins with Roles in Intracellular Growth. *Journal of Bacteriology* **193,** 854-861.

McDonough, J.A., Hacker, K.E., Flores, A.R., Pavelka, M.S. and Braunstein, M. (2005). The twin-arginine translocation pathway of *Mycobacterium smegmatis* is functional and required for the export of mycobacterial β-lactamases. *Journal of bacteriology* **187,** 7667-7679.

McMurtrey, C., Harriff, M.J., Swarbrick, G.M., Duncan, A., Cansler, M., Null, M., *et al.* (2017). T-cell recognition of *Mycobacterium tuberculosis* peptides presented by HLA-E derived from infected human cells. *PLOS ONE* **12,** e0188288.

Molloy, M.P., Herbert, B.R., Slade, M.B., Rabilloud, T., Nouwens, A.S., Williams, K.L. and Gooley, A.A. (2000). Proteomic analysis of the *Escherichia coli* outer membrane. *The FEBS Journal* **267,** 2871-2881.

Mullen, L.M., Nair, S.P., Ward, J.M., Rycroft, A.N. and Henderson, B. (2006). Phage display in the study of infectious diseases. *Trends in Microbiology* **14,** 141-147.

Ng, S.Y., VanDyke, D.J., Chaban, B., Wu, J., Nosaka, Y., Aizawa, S.-I. and Jarrell, K.F. (2009). Different minimal signal peptide lengths recognized by the archaeal prepilin-like peptidases FlaK and PibD. *Journal of bacteriology* **191,** 6732-6740.

Perkowski, E.F., Zulauf, K.E., Weerakoon, D., Hayden, J.D., Ioerger, T.R., Oreper, D., *et al.* (2017). The EXIT Strategy: an Approach for Identifying Bacterial Proteins Exported during Host Infection. *MBio* **8**.

Puri, R.V., Reddy, P.V. and Tyagi, A.K. (2013). Secreted acid phosphatase (SapM) of *Mycobacterium tuberculosis* is indispensable for arresting phagosomal maturation and growth of the pathogen in guinea pig tissues. *PloS one* **8,** e70514.

Schneewind, O. and Missiakas, D. (2014). Sec-secretion and sortase-mediated anchoring of proteins in Gram-positive bacteria. *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research* **1843,** 1687-1697.

Song, H., Sandie, R., Wang, Y., Andrade-Navarro, M.A. and Niederweis, M. (2008). Identification of outer membrane proteins of *Mycobacterium tuberculosis*. *Tuberculosis* **88,** 526-544.

Vizcaino, C., Restrepo-Montoya, D., Rodriguez, D., Nino, L.F., Ocampo, M., Vanegas, M., *et al.* (2010). Computational Prediction and Experimental Assessment of Secreted/Surface Proteins from *Mycobacterium tuberculosis* H37Rv. *Plos Computational Biology* **6**.

Wang, L.-F. and Yu, M. (2004). Epitope identification and discovery using phage display libraries: applications in vaccine development and diagnostics. *Current drug targets* **5,** 1-15.

WHO, U., FIND, Stop TB Partnership (2014). Progress in diagnosis: a key step in overcoming the MDR-TB crisis. *World Health Organization* http://www.who.int/campaigns/tb-day/2014/tb-brochure/en/.

Xu, G., Ni, Z., Shi, Y., Sun, X., Wang, H., Wei, C., *et al.* (2014). Screening essential genes of *Mycobacterium tuberculosis* with the pathway enrichment method. *Mol Biol Rep***,** 1-6.

APPENDICES

APPENDIX A: SUPPLEMENTARY INFORMATION TABLES

Table S1. 1: *M. tuberculosis* H37Rv metabolic classes and pathways according to PATRIC database.

Classes	Pathways
Amino Acid Metabolism	Glycine, serine and threonine metabolism
	Cysteine and methionine metabolism
	Valine, leucine and isoleucine degradation
	Lysine degradation
	Arginine and proline metabolism
	Histidine metabolism
	Tryptophan metabolism
	Valine, leucine and isoleucine biosynthesis
	Phenylalanine metabolism
	Tyrosine metabolism
	Alanine, aspartate and glutamate metabolism
	Phenylalanine, tyrosine and tryptophan biosynthesis
	Lysine biosynthesis
Biosynthesis of Polyketides and Nonribosomal Peptides	Biosynthesis of ansamycins
	Polyketide sugar unit biosynthesis
	Biosynthesis of type II polyketide backbone
	Biosynthesis of type II polyketide products
	Biosynthesis of 12-, 14- and 16-membered macrolides
	Biosynthesis of siderophore group nonribosomal peptides
	Biosynthesis of vancomycin group antibiotics
Biosynthesis of Secondary Metabolites	Streptomycin biosynthesis
	Limonene and pinene degradation
	Phenylpropanoid biosynthesis
	Tetracycline biosynthesis
	Isoquinoline alkaloid biosynthesis
	Insect hormone biosynthesis
	Terpenoid backbone biosynthesis
	Tropane, piperidine and pyridine alkaloid biosynthesis
	Flavonoid biosynthesis
	Carotenoid biosynthesis
	Zeatin biosynthesis
	Anthocyanin biosynthesis
	Flavone and flavonol biosynthesis
	Stilbenoid, diarylheptanoid and gingerol biosynthesis
	Betalain biosynthesis
	Puromycin biosynthesis
	Novobiocin biosynthesis
	Penicillin and cephalosporin biosynthesis
	Diterpenoid biosynthesis

Classes	Pathways
	Brassinosteroid biosynthesis
	Isoflavonoid biosynthesis
	Caffeine metabolism
	beta-Lactam resistance
Carbohydrate Metabolism	Inositol phosphate metabolism
	Starch and sucrose metabolism
	Citrate cycle (TCA cycle)
	Glyoxylate and dicarboxylate metabolism
	Pentose and glucuronate interconversions
	Galactose metabolism
	Glycolysis / Gluconeogenesis
	Ascorbate and aldarate metabolism
	Pyruvate metabolism
	Propanoate metabolism
	Butanoate metabolism
	Pentose phosphate pathway
	Fructose and mannose metabolism
	Amino sugar and nucleotide sugar metabolism
	C5-Branched dibasic acid metabolism
Energy Metabolism	Reductive carboxylate cycle (CO2 fixation)
	Methane metabolism
	Sulfur metabolism
	Nitrogen metabolism
	Carbon fixation in photosynthetic organisms
	Oxidative phosphorylation
	Photosynthesis
Glycan Biosynthesis and Metabolism	Peptidoglycan biosynthesis
	Glycosaminoglycan degradation
	O-Glycan biosynthesis
	High-mannose type N-glycan biosynthesis
	Lipopolysaccharide biosynthesis
	Glycosylphosphatidylinositol(GPI)-anchor biosynthesis
	Glycosphingolipid biosynthesis - lacto and neolacto series
	Glycosphingolipid biosynthesis - globo series
	Glycosphingolipid biosynthesis - ganglio series
Immune System	T-cell receptor signaling pathway
Lipid Metabolism	Fatty acid metabolism
	Glycerolipid metabolism
	Fatty acid elongation in mitochondria
	alpha-Linolenic acid metabolism
	Secondary bile acid biosynthesis
	Glycerophospholipid metabolism
	Primary bile acid biosynthesis
	Fatty acid biosynthesis
	C21-Steroid hormone metabolism

Classes	Pathways
	Sphingolipid metabolism
	Linoleic acid metabolism
	Synthesis and degradation of ketone bodies
	Biosynthesis of unsaturated fatty acids
	Ether lipid metabolism
	Arachidonic acid metabolism
	Steroid biosynthesis
Metabolism of Cofactors and Vitamins	One carbon pool by folate
	Nicotinate and nicotinamide metabolism
	Pantothenate and CoA biosynthesis
	Retinol metabolism
	Thiamine metabolism
	Porphyrin and chlorophyll metabolism
	Ubiquinone and other terpenoid-quinone biosynthesis
	Biotin metabolism
	Folate biosynthesis
	Vitamin B6 metabolism
	Riboflavin metabolism
	Lipoic acid metabolism
Metabolism of Other Amino Acids	Glutathione metabolism
	Cyanoamino acid metabolism
	Selenoamino acid metabolism
	beta-Alanine metabolism
	Taurine and hypotaurine metabolism
	D-Glutamine and D-glutamate metabolism
	D-Arginine and D-ornithine metabolism
	Phosphonate and phosphinate metabolism
	D-Alanine metabolism
Nucleotide Metabolism	Pyrimidine metabolism
	Purine metabolism
Signal Transduction	mTOR signaling pathway
	Phosphatidylinositol signaling system
Translation	Aminoacyl-tRNA biosynthesis
Xenobiotics Biodegradation and Metabolism	Tetrachloroethene degradation
	Geraniol degradation
	Benzoate degradation via hydroxylation
	1,4-Dichlorobenzene degradation
	Caprolactam degradation
	Atrazine degradation
	Trinitrotoluene degradation
	Naphthalene and anthracene degradation
	Metabolism of xenobiotics by cytochrome P450
	Drug metabolism - cytochrome P450
	Drug metabolism - other enzymes

Classes	Pathways
	2,4-Dichlorobenzoate degradation
	1,1,1-Trichloro-2,2-bis(4-chlorophenyl)ethane (DDT) degradation
	Biphenyl degradation
	Toluene and xylene degradation
	1- and 2-Methylnaphthalene degradation
	Ethylbenzene degradation
	Bisphenol A degradation
	gamma-Hexachlorocyclohexane degradation
	Styrene degradation
	Fluorobenzoate degradation

Table S2. 1: Insert DNA sequences, encoded peptide sequences of 98 distinct sequences and their corresponding protein name (Rv no.).

ID	Rv no.	DNA Insert Sequence	Peptide Fragment	SP	TM
S1	MT3042	CGCGGCCCCGAATTAGGACAATTTTCGCACCTAGCGCATCCAATATC GCTTTCGAAGAACGTTCACGCCAGTCCCACTGGGCCGGTGCGAATG GTGCAACGCGCCTTTCGTCGAAGGAAACGCCGTCCGCCACCGAGCC CGCGCTAGGCAAGTCGGTCCCAAGAACGTCGCAAGGATACGCCAA GCGGCCGCGGTCAATCTTGACTTGTCGGCCACCGCCGGCAAACCAA CATTCAGCCACAACGCGACAGAGAGGTACCCAATGTTCACTGCCCG TATCCGCGCCCTCGCCGGCATGTCTCTGCTAGCCTCGGCGATCGGAC TGGCGGCCTTCGGAGCCGCTACCGGCACCGCCAATGCCGCCCCGAC CCAC	AAPNDNFRTRIQYRFRRTFTPVPLGRCEWCNAPFVEGNAVRHRARARQVGPK NVARIRQAAAVNLDL**SATAGKPTFSHNATERYPMFTARIRALAGMSLLASAIG LAAFGAATGTANAAPTH**	**Y**	
S2	Rv0101	CGCTGCTCCGGAACGGACGGTGTCGTCGATCGATGCGCTGGATGG GACCGAGCGTGCCCGGTTGGATGAGTGGGGTAACCGCGCTGTGCT GACTGCGCCCGCGCCCACGCCGGTGTCGATCCCGCAGATGTTGGCC GCCCAGGTGGCACGTATCCCCGAAGCGGAGGCGGTGTGTTGCGGG GACGCGTCGATGACGTATCGGGAACTCGACGAGGCGTCCAACCGG TTAGCGCATCGGCTGGCAGGTTGTGGGGCCG	**AAPERTVSSIDALDGTERARLDEWGNRAVLTAPAPTPVSIPQMLAAQVARIP EAEAVCCGDASMTYRELDEASNRLAHRLAGCGA**	**N**	
S3	Rv0116c	ARAWTGTGGTGKGTYKATCSMTCCATCTCKKGATGGCGGACGCCCT GCGTCGAGTGGTTCGTTATCTATCCGTTGTGGTCGCGATCACGCTGA TGCTCACCGCGGAATCAGTCAGCATAGCGACCGCCGCGGTCCCGCC ACTCCAACCGATCCCAGGCGTTGCGTCGGTGTCGCCGGCTAATGGT GCCGTGGTGGGGGTGGCGCACCCGGTGGTG	VVSHLMADAL**RRVVRYLSVVVAITLMLTAESVSIATAAVPPLQPIPGVASVSP ANGAVVGVAHPVV**	**Y**	
S4	Rv0125	TAGAAAAATCCTGCCGCCCGGACCCTTAAGGCTGGGACAATTTCTG ATAGCTACCCCGACACAGGAGGTTACGGGATGAGCAATTCGCGCCG CCGCTCACTCAGGTGGTCATGGTTGCTGAGCGTGCTGGCTGCCGTC GGGCTGGGCCTGGCCACGGCGCCGGCCCAGGCGGCCCCGCCGGCC TTGTCGCAGGACCGGTTCGCCGACTTCCCCGCGCTGCCCCTCGACCC GTCCGCGATGGTCGCCCAAGTGGGGCCACAGGTGGTCAACATCAAC ACCAAACTGGGCTACAACAACGCCGTGGGCGCCGAGGTCGGCGAG GCTTTGGTCTCGATGCTCAAGGATCACGGTGTCGGCTTCCATCCTCG CAAGGCCCTAGCTCGCGTCGATGAGGCCGCAAGGACGATGCACTTC GGTGACGGCACGTCCGAACCGTTCGATCTGCTTGCCGTGGTCCCCC CGCACGTGCCCTCCGCCGCGGCGCGGTCAGCGGGTCTCAGCGAATC	KNPAARTLKAGTISDSYPDTGGYG**MSNSRRRSLRWSWLLSVLAAVGLGLATA PAQAAPPALSQDRFADFPALPLDPSAMVAQVGPQVVNINTKLGYNNAVGA** EVGEALVSMLKDHGVGFHPRKALARVDEAARTMHFGDGTSEPFDLLAVVPPH VPSAAARSAGLSESGWIPVDPRTLSTSAD VWAIGDATVLTLPNGKPLPKGSRD	**Y**	

ID	Rv no.	DNA Insert Sequence	Peptide Fragment	SP	TM
		CGGGTGGATACCCGTGGACCCGCGCACCCTGTCCACTAgCGCCGAC WACGTGTGGGCCATCGGCGATGCGACCGTGCTGACGCTGCCGAAT GGCAAACCGCTGCCCAAGGGGTCTAGAGA			
S5	Rv0125	CAATTTCTGATAGCTACCCCGACACAGGAGGTTACGGGATGAGCAA TTCGCGCCGCCGCTCACTCAGGTGGTCATGGTTGCTGAGCGTGCTG GCTGCCGTCGGGCTGGGCCTGGCCACGGCGCCGGCCCAGGCGGCC CCGCCGGCCTTGTCGCAGGACCGGTTCGCCGACTTCCCCGCGCTGC CCCTCGACCCGTCCGCGATGGTCGCCCAAGTGGGGCCACAG	ISDSYPDTGGYG**MSNSRRRSLRWSWLLSVLAAVGLGLATAPAQAAPPALSQD RFADFPALPLDPSAMVAQVGPQ**	**Y**	
S6	Rv0129c	CATGCCCAGACACTGCGGAAATGCCACCTTCAGGCCGTCGCGTCGG TCCCGAATTGGCCGTGAACGACCGCCGGATAAGGGTTTCGGCGGTG CGCTTGATGCGGGTGGACGCCCGAAGTTGTGGTTGACTACACGAGC ACTGCCGGGCCCAGCGCCTGCAGTCTGACCTAATTCAGGATGCGCC CAAACATGCATGGATGCGTTGAGATGAGGATGAGGGAAGCAAGAA TGCAGCTTGTTGACAGGGTTCGTGGCGCCGTCACGGGTATGTCGCG TCGACTCGTGGTCGGGGCCGTCGGCGCGGCCCTAGTGTCGGGTCTG GTCGGCGCCGTCGGTGGCACGGCGACCGCGGGGGCATTTTCC	HAQTLRKCHLQAVASVPNWP&TTAG#GFRRCA&CGWTPEVVVDYTSTAGPS ACSLT#FR**MRPNMHGCVEMRMREARMQLVDRVRGAVTGMSRRLVVGAV GAALVSGLVGAVGGTATAGAFS**	**Y**	
S7	Rv0179c	CTGTTGCTGGGTGGATTCATCACACACCTGTACGACCGCAGCAGCTT CCACCGAATTCCGGGAGCGGTCGAGATGGCGATCGCCGTGGAACA CCAAGCTCAATACCAAGCTCGCCAAAGCGCACGCGAAAGCTCAAGC GAACAACCGTAATTCGAACGCGGCCGAACCGTCACGCAGCTGTTAC TGAAGTTGACAGTGTGATCAGTGTGCCTAATGCTGGGATGGTTGCA TGTGGATTCGGGCCGAGAGGGTTGCTGTGCTGACACCGACTGCCAG CCTGCGCCGATTGACGGCTTGCTACGCCGCGTTGGCGGTGTGCGCC GCCCTCGCCTGCACCACCGGGCAGCCGGCCGCCCGCGCCGCCGACG GGCGCGAGATGCTCGCCCAAGCGATAGCCACCACCAGGGGCTCAT ACCTGGTGTACAACTTCGGCGGCGGTCATCCTATGCCGCTGCTCAAC GCAGGTGGTCACTGGTACGAGATGAACAACGGCGGCCATCTGATG ATCATCAAGAATGCCTCCCAACGGCTTTCACCACATCTACTGGTAGA CACCCACACTGGAGACCAGGCGCGCTGCGAA	CCWVDSSHTCTTAAASTEFRERSRWRSPWNTKLNTKLAKAHAKAQANNRNS NAAEPSRSCY&S&QCDQCA#CWDGC**MWIRAERVAVLTPTASLRRLTACYAA LAVCAALACTTGQPAARAADGREMLAQAIATTRGSYLVYNFGGGHPMPLLN AGGHWYEMNNGGHLMIIKNASQRLSPHLLVDTHTGDQARCE**	**Y**	

ID	Rv no.	DNA Insert Sequence	Peptide Fragment	SP	TM
S8	Rv0192	GTCTATTCGCCGCTCTGAACATAGCCGCAGTGGTTGCGGTGCTGATGCTGGGTGCTGGCGTTGCCGTGGCGGACCCGGTTTTC	**SIRRSEHSRSGCGADAGCWRCRGGPG**F	**N**	
S9	Rv0203	CGCAAACCCGTCGCAAACCCGTCGCAAACCGTAAGGAGTCATCCATGAAGACAGGCACCGCGACGACGCGGCGCAGGCTGTTGGCAGTACTGATCGCCCTCGCGTTGCCGGGGGCCGCCGTTGCGCTGCTGGCCGAACCATCAGCGACCGGCGCGTCGGACCCGTGCGCGGCCAGCGAAGTGGCGAGGACGGTCGGTTCGGTCGCCAAGTCGATGGGCGACTACCTGGATTCACACCCAGAGACCAACCAGGTGATGACCGCGGTCTTGCAGCAGCAGGTAGGGCCGGGGTCGGTCGCATCGCTGAAGGCCCATTTCGAGGCG	QTRRKPVANRKESS**MKTGTATTRRRLLAVLIALALPGAAVALLAEPSATGASDPCAASEVARTVGSVAKSMGDYLDSHPETNQVMTAVLQQQVGPGSVASLKAHFEA**	**Y**	
S10	Rv0236c	CTGCCGTCGCCCGGTCCGCCAGAACGCCAGCAGAGCTAGCAGGGGCAGCAGGGCCAGCCCTATCGCCAGGCTCGCCCGATACAGCGAGTTCGGTGCGAATGTCAGCGTGATGGTGCCGGATTCGCCTGCGTCG	DAGESGTITLTF**APNSLYRASLAIGLALLPLLALLAFWRTGRRQ**	**P**	**TM**
S11	Rv0255c	CGCTGGTCGCGGGGTTTGTGGTCAATAAGTTTCGGGGCGACTCCGACCCTGCTGGCGCCAGGTCTGCGCGACCTGGAACG	**LVAGFVVNKFRGDSDLLAPGLRDLER**	**N**	
S12	Rv0320	CCAGTTGATGAACAATGTGCCCCAAGCGCTGCAACAACTGGCCCAGCCCACGAAAAGCATCTGGCCGTTCGACCAACTGAGTGAACTCTGGAAAGCCATCTCGCCGCA**CGCTGGCTCAGCGGATGTAGAGTTCCTCGCCGGTGGGATACCCGCCGCCCACGGTCGTGATGTGGATGTGGTCGTAGTGGCCGGCCCCGGTTGTCCGCGCGCCGTTGGGCGTGTAGTAGGCGCCACGCCAAATCACATCTTGCATCCCAAATCGGGTCGCGTTCTTCAGTACGAAAGCGACGATCTCGTTGCCCAGCGCTATGCCCTCGGCGGTGCCGGGGTTGGGAACCATCACGTCGAGCGCCAAACCATTGGGATGCCATCTCAGCGCATCCGGCCGAACGCCGCCGATTTCGCGAATTTCGGGGAAAGCCGCACTGATACTGCGGGCTGTCAAGACGGTCTTGACCTGTAGACCCTGCTCGGGTGCCACCCCTACCGGCAAATACCGAGAYACGATGCGCCATCTCGATGCCGAAGCGAACTGGCTGGCAGCAAGGCCCGTCCCGATTTCACCGCCGGATATCTGTGAGGAGAAAGCCAGCGGTGCAACCGGCACGATCTGCACACAACAGGGCGCGTCGTCGCCGAGAACCGGGTTGGCATCGGCGCGGGCGGCAAGCGTACTGACATCTCCGCCCGTGGCGAAGAACACGGCGGCCGGGGCGAGGACCGCA	PRCSEHVGPDGGRRGHGWHRGSHSAVRRYPRTDPVAIADLGRS&RRSSRD#IFILPLSCRNLLSV**GRHELARDRRKSSAVLAAVLAPAAVFFATGGDVSTLAARADANPVLGDDAPCCVQIVPVAPLAFSSQISGGEIGTGLAASQFASASRWRIVSRYLPVGVAPEQGLQVKTVLTARSISAAFPEIREIGGVRPDALRWHPNGLALDVMVPNPGTAEGIALGNEIVAFVLKNATRFGMQDVIWRGAYYTPNGARTTGAGHYDHIHITTVGGGYPTGEELYIR**&ASVRRDGFPEFTQLVERPDAFRGLGQLLQRLGHIVHQL	**Y**	

ID	Rv no.	DNA Insert Sequence	Peptide Fragment	SP	TM
		GCCAGGACCGCTGACGACTTTCGCCGGTCCCTAGCTAATTCATGTC GGCCCACGCTGAGCAGATTACGACAAGATAACGGAAGAATAAATA TCTAATCTCTAGAAGACCGCCGTCAGCTGCGCCCTAGATCTGCAATC GCGACGGGATCGGTTCGCGGATATCTGCGGACTGCCGAATGGCTG CCTCGATGCCAGCCGTGACCCCGGCGACCGCCGTCTGGACCGACAT GCTCGGAACACCGAGG			
S13	Rv0425c	CACCCCGGGGTCGGGTCGGCCAGCCTGAACTACCCGTTGTCCCGTG TCGTCGTCGCTATCGACGATCCCGACACATCACTGCGCGAACTTTGC CGCATTGTCGATGACGCCGAAAAAGCCGAAAGGCACCGGCCACCC AGGGGGAACAGACCCCGACCGAAC	**HPGVGSASLNYPLSRVVVAIDDPDTSLRELCRIVDDAEKAERHRPPRGNRPRP N**	**P**	**TM**
S14	Rv0436c	ccatgacctggcagagctggacaaggtgggggcctacgacgtgatcaccgtgttcgacg cgattcacgaccagGCGCAGCCAGCGCTGGCTTGCCCAACACCCTGAGG TGTGGGACGACAAGCCCAAGCAACGGCGCGCGGTGCGGCGCGCGA GCCGCCGGGCGCATCCCTACCGGCCGTCGATGGCGCGGCTGGGCC TGCGCAAGCCGGGTCGACGGCTGTGAcccacc	MTWQSWTRWGPTT&SPCSTRFTTR**RSQRWLAQHPEVWDDKPKQRRAVRR ASRRAHPYRPSMARLGLRKPGRRL**&PT	**P**	**TM**
S15	Rv0455c	CGACTACAGCCAAGCCCCACCGGAACTGCGACCCGACCCCACGCAC GTCATCACTCCTGATGGGAGCATGTAGTGAGACAGACTATGAACCT TGTCGTCTGCCCGCCTCGAGGCGCCCCAAAAGTGCTACGGTAACCA CTATGTCTCGGCTGAGTTCCATCCTGCGTGCCGGCGCGGCATTTCTG GTTCTCGGCATCGCCGCTGCGACATTTCCACAAAGCGCGGCAGCCG ACTCCACGGAAGACTTTCCAATACCTCGCCGGATGATCGCAACCACC TGCGACGCCGAACAATATCTGGCGGCGGTGCGGGATACCAGTCCG GTGTACTACCAGCGGTACATGATCGACTTCAACAACCATGCAAACCT TCAGCAAGCGACGATCAACAAGGCGCACTGGTTCTTCTCGCTGTCA CCGGCGGAGCGCCGAGACTACTCCGAACACTTTTACAATGGCGATC CGCTGACGTTTGCCTGGGTCAATCACATGAAAATCTTC	TTAKPHRNCDPTPRTSSLLMGACSETDYEPCRL**PASRRPKSATVTTMSRLSSILR AGAAFLVLGIAAATFPQSAAADSTEDFPIPRRMIATTCDAEQYLAAVRDTSPV YYQRYMIDFNNHANLQQATINKAHWFFSLSPAERRDYSEHFYNGDPLTFAW VNHMKIF**	**Y**	
S16	Rv0559c	TGAAGGGAACAAAGCTGGCTGTTGTCGTCGGCATGACGGTGGCTG CCGTTAGTTTGGCAGCGCCGGCGCAGGCCGACGACTACGACGCCCC CTTCAACAACACGATCCATCGCTTCGGGATCTACGGCCCGCAGGACT ACAACGCTTGG	**KGTKLAVVVGMTVAAVSLAAPAQADDYDAPFNNTIHRFGIYGPQDYNAW**	**Y**	
S17	Rv0603	AGGAGAAAGCGATGAATCGCATCGTGCAGTTCGGAGTTTCCGCCGT GGCCGCGGCGGCGATCGGCATCGGAGCCGGGTCGGGGATCGCGG	EKA**MNRIVQFGVSAVAAAAIGIGAGSGIAAAFDGEDEVTGPDADRARAAAV Q**	**Y**	

ID	Rv no.	DNA Insert Sequence	Peptide Fragment	SP	TM
		CGGCGTTCGACGGCGAGGACGAGGTGACCGGCCCCGACGCCGACCGCGCGCGCGCCGCCGCGGTGCAG			
S18	Rv0675	ATCGGGCCGGGACCCGTCCGGTGCACAGAGTTGGCCTCCGGTGTGCCAAAGGCCTTCAAATCGGCTCCCGCACAAAAGGTTCCACCCGCACCCCAGAGTACGGCCACCGACGCGGCGTCGTCCCGGTCGAATTGCTCGAACGCCGCGCACAACGCCGCGGCGGTCGGGCCGTTGACCGCGTTGCGGGAGGCCGGCCGGTTCAGAATCACCGTGGTCACCCGACCTTTGCGCTCCACACGCACCAGATCACTCATGTCACCTCCAGGAGTTGAGTTGCATCGCGTCGTTTCGCCATTGCAGTGGCGAAGTCGTGGTACGCCGCCCGTAACCCGGCGCCCGGCCAGTCGGCGGGCAACAGTTCAGCGGGCAACATCGGATCGGTGAG	HRSDVAR&TVARRLAGRRVTGGVPRLRHCNGETTRCNSTPGGD**MSDLVRVERKGRVTTVILNRPASRNAVNGPTAAALCAAFEQFDRDDAASVAVLWGAGGTFCAGADLKAFGTPEANSVHRTGPGP**	**Y**	
S19	Rv0676c	CCAGGGCCACCGCACCGACCAGGATGGGGCCGGGCCAGCGGACGATGGCGGCCCCGACCTTGCGCCAGCCCCGCACCCGCGCCATCCGCTTGGGCTCGAGCAGCTTGCCGAACCGGCTCGTCACGGCGATTATCGCCGGGCCCAGGGTGAGTGCGGCGGCGACG	**VAAALTLGPAIIAVTSRFGKLLEPKRMARVRGWRKVGAAIVRWPGPILVGAVAL**	**N**	**TM**
S20	Rv0676c	TCGCCGACGTGATGATCGACCGGTAGACCAGCAGCAACATCACGATGATCACGGTGAACGTGACCGCCTCGATCACCTGCAGACTACGGTCGCCGGCCTGCTGCTGATCGGCGACCAGCGCGGCCGAACCGGTGACGTACACCTTGACACCGGGTGGCGGCGCAAGGCGCTC	**ERLAPPPGVKVYVTGSAALVADQQQAGDRSLQVIEAVTFTVIIVMLLLVYRSIITSA**	**N**	**TM**
S21	Rv0822c	CAATCATCATCGCCGTGCTGCTGGCCACCACAGTGTTTTTTTCGCCCGGCAACGAACAGGCTGCCGCCACCGTGGCCGCCGTGTTCGGCCAGTCAAAGATCGAGCGGGTGACCGGGATCGGCCAACTGGTCCAGGTGGTGCTGGGCCAAGACTTCAGCGCGGTGCGCGCTCCCCTGCCGAGTGGCTCCACCGTCAGCGTGCAGATAAGCCGCAACTCCTCCAGCCCACCGACCAAGCTGCCCGAGGACCTGACGGTCACCAACCCTGCTTGCGGCCGGCACTACTCGGGCCCCCTTCTCGACCTGGGCGGTGGTTGCTGTCGTGGTTGTTCCGGCGGTGGTGCTTTGTGTGGTTGG	IIIAV**LLATTVFFSPGNEQAAATVAAVFGQSKIERVTGIGQLVQVVLGQDFSAVRAPLPSGSTVSVQISRNSSSPPTKLPEDLTVTN**PACGRHYSGPLLDLGGGCCRGCSGGGALCGW	**N**	
S22	Rv0824c	CTCTCCGCGAAATAGTGGCCCTGGTGGTTTTGGCCTGGGCTGAAGCCCCGGTTGACTACCTCGAGGCGAAGTTTCTCCAACTCGACAGGGTCGACCGATCGGGTCACCACCAGGTAGTCGCGCAGCGCGATGCCGTGCCGATTCTCCTCGGCGGTCCAACGGTTGACCCACTGCCCCCACGCGCCGTCCATGCC	**GMDGAWGQWVNRWTAEENRHGIALRDYLVVTRSVDPVELEKLRLEVVNRGFSPGQNHQGHYFAE**	**N**	

ID	Rv no.	DNA Insert Sequence	Peptide Fragment	SP	TM
S23	Rv0842	CGGCGGCGGCCACCAGCATGGTGGATACCAGCGTCACCCGCGCCCCGACACGGTCCACCAACCACCCACACAGCAGCTGCCCCACCGCCCAACCCAGGCCGTAGGCCGCCACGACCGCGCCGACCGCTCCCGCACCATGTCCTCGTCCAGCCACGTGGTAGGCTAGGAACGGATAGCCGAATCCGGCCGAGCGCACCAGCAAGTTGCCGCCCAGCAG	**LLGGNLLVRSAGFGYPFLAYHVAGRGHGAGAVGAVVAAYGLGWAVGQLLCGWLVDRVGARVTLVSTMLVAAA**	**P24**	**TM**
S24	Rv0983	TACCTGGCGTGATTCCGACCATGACGCCCCCTCCTGGGATGGTTCGCCAACGCCCTCGTGCAGGCATGTTGGCCATCGGCGCGGTGACGATAGCGGTGGTGTCCGCCGGCATCGGCGGCGCGGCCGCATCCCTGGTCGGGTTCAACCGGGCACCCGCCGGCCCCAGCGGCGGCCCAGTGGCTGCCAGCGCGGCGCCAAGCATCCCCGCAGCAAACATGCCGCCGGGGTCGGTCGAACAGGTGGCGGCCAAGGTGGTGCCCAGTGTCGTCATGTTGGAAACCGATCTGGGCCGCCAGTCGGAGGAGGGCTCCGGCATCATTCTGTCTGCCGAGGGGCTGATCTTGACCAACAACCACGTGATCGCGGCGGCCGCCAAGCCTCCCCTGGGCAGTCCGCCG	**PGVIPTMTPPPGMVRQRPRAGMLAIGAVTIAVVSAGIGGAAASLVGFNRAPAGPSGGPVAASAAPSIPAANMPPGSVEQVAAKVVPSVVMLETDLGRQSEEGSGIILSAEGLILTNNHVIAAAAKPPLGSPP**	**N**	
S25	Rv1029	CCAGCCAGGTCGCGATCAAGCAGCTCGGCACCAACGGCGGCGGGTTCTTCAACGTGAACTCCGCGCATCCGTTCGAAAACTACACGCCGATAGGCAATTTCGTCGAAAACTGGGCGATCCTGATCATCCCGTTCGCGCTGTGCTTCGCCTTCGGCAAGATGGTGCACGACCGTCGTCAAGGCTGGGCGGTGCTGGCCATCATGGGCATCATTTGGATCGGAATGTCAGTCGCGGCAATGTCATTCGAGGCCAAGGGCAACCCGCGGCTGGATGCGCTGGGGGTGACACAGCAGACGACGGTCGACCAGTCCGGCGGCAACCTGGAGGGCAAGGAGGTGCGCTTTGGC	**SQVAIKQLGTNGGGFFNVNSAHPFENYTPIGNFVENWAILIIPFALCFAFGKMVHDRRQGWAVLAIMGIIWIGMSVAAMSFEAKGNPRLDALGVTQQTTVDQSGGNLEGKEVRFG**	**P**	**TM**
S26	Rv1029	TGCATGGTTTCATCGTCGCCAACACGCTGGAGGGCGCCCCCCAGCTCATTCCAGGCGGGCCGGTGGCCAGCCAGGTCGCGATCAAGCAGCTCGGCACCAACGGCGGCGGGTTCTTCAACGTGAACTCCGCGCATCCGTTCGAAAACTACACGCCGATAGGCAATTTCGTCGAAAACTGGGCGATCCTGATCATCCCGTTCGCGCTGTGCTTCGCCTTCGGCAAGATGGTGCACGACCGTCGTCAAGGCTGGGCGGTGCTGGCCATCATGGGCATCATTTGGATCGGAATGTCAGTCGCGGCAATGTCATTCGAGGCCAAGGGCAACCCGCGGCTGGATGCGCTGGGGGTGACACAGCAGACGACGGTCGACCAGTCCGGCGGCAACCTGGAGGGCAAGGAGGTGCGCTTTGGCGTCGGTGCGTCTGGGTTATGGGCGGCGTCGACGACCGGCACCTCC	**HGFIVANTLEGAPQLIPGGPVASQVAIKQLGTNGGGFFNVNSAHPFENYTPIGNFVENWAILIIPFALCFAFGKMVHDRRQGWAVLAIMGIIWIGMSVAAMSFEAKGNPRLDALGVTQQTTVDQSGGNLEGKEVRFGVGASGLWAASTTGTSNGSVNSMHDSYTPLGGM**	**P**	**TM**

ID	Rv no.	DNA Insert Sequence	Peptide Fragment	SP	TM
		AACGGCTCGGTCAACTCGATGCACGACAGCTACACACCACTGGGCG GCATG			
S27	Rv1118c	CACGCCAATCGCGACCAGGAGGACAAACTGCTGCGGGTCATCGCG CGGATGAACGGCACGCCTTTCCCAACCACCGCCCGGTTGACCGGCC GGTGGCTGCGCGGCCGGCTTCCSACCCTCAMCGATTGGCTGCK	**HANRDQEDKLLRVIARMNGTPFPTTARLTGRWLRGRLPTL DWL**	**N**	
S28	Rv1156	TCCGTCCCGACAGCTCTACACTGAGGACGTGCCAAATCTGCAGCTTG TCCAAGAGCCGGCAGCCGACGCGCTGCTGAACGCCAACCCATTCGC GTTGCTGGTGGGCATGTTGCTCGACCAGCAGGTGCCGATGGAGACC GCCTTCGGCCACTGGCATTGGTGCACTGCACCGTGCGCCATTCGTG GCGACAACTGCGAGCGGGAGCGGGACCAAGGATGATGGTCCCGGT CGCGACGGGCGCGATCCCGCTCCG	PSRQLYTEDV**PNLQLVQEPAADALLNANPFALLVGMLLDQQVPMETAF**GHW HWCTAPCAIRGDNCERERDQG&WSRSRRARSRS	**N**	
S29	Rv1157c	GGCGTCAAGATCCGGGACAAGGTCCAGGACAAAGTAATCGCCATC ACCGGCGGCGCCCGGGGGATCGGATTGGCCACAGCGGCCGCGCTG CCCTGACGCATCTGTGAGAAGGCTGACCAACACCGAACACCGGGA GAACACCACCGTGGCAAGCACTTGGAGTGTGTGCAAAGGTTTGGCC GCCGTCGTCATCACCTCGGCTGCCGCGTTCGCGCTGTGTCCGAACG CGGCAGCCGACCCGGCGACGCCGCAGCCCAACCCCACTCAACAGCT ACCGGGCTTGCCGGCGTTGGCCCAGCTGAGTCCGATAATCCAGCAA	ASRSGTRSRTK#SPSPAAPGGSDWPQRPRCPDASVRRLTNTEHRENTTV**ASTW SVCKGLAAVVITSAAAFALCPNAAADPATPQPNPTQQLPGLPALAQLSPIIQQ**	**Y**	
S30	Rv1161	GGTGTCGTGCTGCAGCGCGGTCAGCACCACATCGGTGCGAGTGCCC AGATGACGTTTCGCCAGCGCACTGAAAGCACGCGCGATGGCGGCG AATGCGTCAAAGTCCGAACGGGTTTCCCACGGCGGATCGATCGCCG GACTGAA	**FSPAIDPPWETRSDFDAFAAIARAFSALAKRHLGTRTDVVLTALQHDT**	**N**	
S31	Rv1200	GAACCAGCCCGACAGTCACGGTTGCCAGGCCCATGATCAACAGTGT GGCGACCAGGGTCTTCTTGCGGCCGAGGCGGTCTCCAAAGTATCCA AAGACGGCCGCGCCGAACGGCCGGGATAGGAACGCCACAGCAAAT GTCCCCATC	**MGTFAVAFLSRPFGAAVFGYFGDRLGRKKTLVATLLIMGLATVTVGLV**	**Y**	**M**

ID	Rv no.	DNA Insert Sequence	Peptide Fragment	SP	TM
S32	Rv1228	TAGCGCTCAACAAGATGCGTGCCAACTCACCCTGCCCCGGGGAGGCGCGATGAGTCGACAGTGGCACTGGCTGGCAGCGACGCTGCTCCTGATCACCACCGCCGCGTGCAGTCGTCCGGGCACCGAGGAACCGGATTGCCCGACGAAAATAACCTTGCCGCCCGGTTGACCGCCGACACCACGGCGCGCAGCGACGCGGTGGTG	SAQQDACQLTLPRGGA**MSRQWHWLAATLLLITTAACSRPGTEEPDCPTKITLPPG**&PPTPRRAATRWC	**Y**	
S33	Rv1230c	AGAAGACATAGGTTTCCGACGGCTCACACACTGAGGACGGTCGAGTGCACATTGGGGGACGCTGGGGTGCACGCCCGGCCGTCGCTGCAGTGCGGCGGGGAGCTTGTCGCCTAACGCGGGCGCCGGCATTCGGCGTGGCAGCGATTGCCCCGTTGGTATTCGCCAGTGCGGTCGGCGGCGCGGCTCCGGTATTCCCCGGGAGAACCGCGCCGGTGCACGCCGTTATCACCCCGGTGGCCGCGGTCGCCGCGTCCGGCATAGACCTGTCCGGTCCGGTCGTCATCGCCATGAAGCGCCCGCCGACCAGCTTCCGCGTGGCGGTAGCTACCATATCGGCTCCACCACCACCGATGATCGTGAATTCGCCTGGTGCGCTTGGCATTCCGGCCATGGCACTGTCCGCCTACCGCAACGCCGAGCTGAAGATGGCCGCTGCCGCCCCTGGCTGTGGCGTCAGTTGGAACTTGCTGGCCGGGATTGGGCGCATCGAGTCGATGCACGCAAACGGCGGCGCCACCGACGCGCGCGGCACCGCGATCCAGCCGATCTACGGCCCAACGCTGGACGGCACCCTGCCAGGCAACGAGATCATCATCCAAAGCAGCGTCGGCAATCGCGTCACGTACGCCCGCGCGATGGGGCCAATGCAGTTCTTGCCCGGCACTTGGGCTCGGTACGCCACCGACGGCGATGACGACGGTGTGGCTGACCCGCAGAACCTGTTCGACTCCACGTTGGCCGCAGCCCGCTACCTGTGTAGCGGT	KT#VSDGSHTEDGRV**HIGGRWGARPAVAAVRRGACRLTRAPAFGVAAIAPLVFASAVGGAAPVFPGRTAPVHAVITPVAAVAASGIDLSGPVVIAMKRPPTSFRVAVATISAPPPPMIVNSPGALGIPAMALSAYRNAELKMAAAAPGCGVSWNLLAGIGRIESMHANGGATDARGTAIQPIYGPTLDGTLPGNEIIIQSSVGNRVTYARAMGPMQFLPGTWARYATDGDDDGVADPQNLFDSTLAAARYLCSG**	**Y**	**TM**
S34	Rv1254	GTCAGATCCGTATCGGGCCAGACTTCTCGCTAGTAGAATTGCCGGCGAAGCTGCCCCGCGCGACGCTCAAAAAGCTTGCACAGACCCGTATCTCGGGTGTGCTGATCGACCTTCGGCCATACCGGCCGCCCGACGCGGCGCGCCGGCATAATGGCGGCAAACCACGGCGGAAACACGTCGGATGACCCTGCCCAAGGAAAGAGCCGCCCAGGGCGGACTCGAGCGGATCGCCCACGTGGACCGGGTGGCGTCGTTGACCGGGATCCGTGCTGTTGCCGCATTGCTGGTCGTCGGCCAGTGTTGTCTTCACCCGGCGATAGGCGCCCAGCGCATCGGATTGCCGGTCGGAGAGGTAGTAGGCGGTGATCAGCTGTGTCCACAGCGGCTCCCGGTAGGG	SDPYRARLLASRIAGEAAPRDAQKACTDPYLGCADRPSAIPAARRGAPA#WRQTTAETRR**MTLPKERAAQGGLERIAHVDRVASLTGIRAVAALLVVG**QCCLHPAIGAQRIGLPVGEVVGGDQLCPQRLPVG	**Y**	

ID	Rv no.	DNA Insert Sequence	Peptide Fragment	SP	TM
S35	Rv1267c	ATCAGTGCGACATTGGCCGTCGGCACACTTGGTGACGTGACGAGCACAACGACGGCGACCACATTCGCATCTGTGACCATGAATTCACGTTCCAGATCAGCGCGGGGACG	QCDIGRRHTW&RDEH**NDGDHIRICDHEFTFQISAGT**	**N**	
S36	Rv1268c	CCTTCAAGACCGCCACCTTCGCGCTGGCCGCCGGTGCCGTTGCACTGGGATTGGCCAGCCCCGCCGACGCAGCGGCGGGCACCATGTATGGCGACCCGGCAGCCGCCGCCAAGTACTGGCGCCAGCAGACATACGACGACTGCGTCCTGATGTCGGCCGCGGACGTGATCGGTCTGAGTCGTGTCGACGAGTTCGAGCAACGACATGTACACGAGACGGTG	**FKTATFALAAGAVALGLASPADAAAGTMYGDPAAAAKYWRQQTYDDCVLMSAADVIG**LSRVDEFEQRHVHETV	**Y**	
S37	Rv1268c	CGACGATGACGACCAGCAAAATCGCCACCGCCTTCAAGACCGCCACCTTCGCGCTGGCCGCCGGTGCCGTTGCACTGGGATTGGCCAGCCCCGCCGACGCAGCGGCGGGCACCATGTATGGCGACCCGGCAGCCGCCGCCAAGTACTGGCGCCAGCAGACATACGACGACTGCGTCCTGATGTCGGCCGCGGACGTG	T**MTTSKIATAFKTATFALAAGAVALGLASPADAAAGTMYGDPAAAAKYWRQQTYDDCVLMSAADV**	**Y**	
S38	Rv1291c	CCGAGCTGTTCAAGGCCCGGTGCTGATTCTGAGGGCACGCAAACCAGCGCAACCTCCGATGACATCAGCACAAGGAGATCATCAATGTTCACTCGCCGTTTCGCCGCCTCCATGGTTGGCACCACCTTGACTGCCGCCACTTTGGGCCTGGCCGCACTCGGCTTCGCCGGGACCGCCAGCGCAAGCTCGACCGACGAAGCGTTC	PSCSRPGADSEGTQTSATSDDISTRRSS**MFTRRFAASMVGTTLTAATLGLAALGFAGTASASSTDEAF**	**Y**	
S39	Rv1366	TGGCCCTTTCCAGCCTAAACTTGGTCGTTGCGAGCACCGTCGCCTTGTTCTTGCTCGCCTCGCCAAGCGCCATCGCCCGCTGGCCGGGGTATTCGATCTCTGCACAGTGTTGCCACTCGGCTTCATGCATCGCCAGATCCGCTGAA	**SADLAMHEAEWQHCAEIEYPGQRAMALGEASKNKATVLATTKFRLERA**	**N**	

ID	Rv no.	DNA Insert Sequence	Peptide Fragment	SP	TM
S40	Rv1435c	TACGGCTATGCGGTTCTGCTGAACCAAATTGATTGCGGAAATATCTGCGCGCGAATAGAGGAGAGTTGACGCTGATGGCGATCGTCAATCGGTTCAACATCAAAGTAATTGCCGGCGCTGGGTTGTTTGCAGCTGCTATAGCGTTAAGCCCGGACGCGGCGGCTGACCCGCTGATGACGGGTGGCTACGCGTGCATCCAGGGCATGGCGGGCGACGCGCCGGTAGCCGCCGGCGACCCGGTGGCTGCCGGTGGACCGGCAGCCGCGGGAGCGTGCAGTGCCGCACTTACTGACATGGCTGGTGTTCCGTTCGTCGCGCCTGGGCCAGTGCCGGCAGCTGCACCGGTGCCCATCGGCGCACCGGTACCAATCCCTGGCGCACCG	RLCGSAEPN&L**RKYLRANRGELTLMAIVNRFNIKVIAGAGLFAAAIALSPDAAADPLMTGGYACIQGMAGDAPVAAGDPVAAGGPAAAGACSAALTDMAGVPFVAPGPVPAAAPVPIGAPVPIPGAP**	**Y**	
S41	Rv1447c	GAAGGTTTCGGTGGTGGAGTCCTCGGCGAACCCCTCCTCGTCGAGCAGCCCAACCACCTTCTCCCCGCCTTGCCAGCCGGCGGCGTACTGGCCGCGGCTGGTGGTCTGGTCGAGTGGCTCGGCAAGGCGGGTGGCCGAGAGCACCTTGA	**KVLSATRLAEPLDQTTSRGQYAAGWQGGEKVVGLLDEEGFAEDSTTETF**	**N**	
S42	Rv1478	GTTTTCACCCGATCAAACTGGCCTGGATCACCGCGGTGGTTGCCGGCCTGATGGTCGGTGTGGCAACGCCCGCCGATGCCGAACCCGGACAATGGGATCCCACGCTGCCGGCATTGGTCAGTGCGGGGGCGCCCGGAGATCCGCTGGCGGTAGCCAACGCGTCGTTGCAGGCCACCGCCCAGGCCACCCAG	**FHPIKLAWITAVVAGLMVGVATPADAEPGQWDPTLPALVSAGAPGDPLAVANASLQATAQATQ**	**Y**	
S43	Rv1522c	CATGGCCTGCAGGGTGTTGACTATCCCGCTCGAGCTGGCCACGGCCCCATTGATTTCGTTGCGTATTTGGGCGAGGGCGTCGGCCAACTGGTGCGCACCGCCGGTCAGCTGGTCCAGCTCGCCTCCGTGCTCTTCGAGCAGGGTGGTCGCTTCGTCGAGCTTGCCGCCCACTTCA	**EVGGKLDEATTLLEEHGGELDQLTGGAHQLADALAQIRNEINGAVASSSGIVNTLQAM**	**P**	**M**

ID	Rv no.	DNA Insert Sequence	Peptide Fragment	SP	TM
S44	Rv1566c	TGGTGGATCTGGCCAGCTATGGACGCGGCCGGCATCTGTTGTCAAAGTGTCTGTAGTGATTGGTGTGAATAAAGTGACTTCCGGCTGGAATGGCAGACGTGGTGGATCTGGCCAGCTATGGACGCGGCCGGCATCTGTTGTCAAAGTGTCTGTAGTGATTGGTGTGAATAAAGTGACTTCCGGCTGGAATGGCAGACGGTGAAGTCACGTGCTGAGGCACACACCACCCGAGAGTGCCAAGGAGGTTACAACCCCATGAAACGCAGCATGAAAAGCGGCTCCTTCGCGATCGGTCTGGCAATGATGCTCGCCCCGATGGTGGCCGCGCCCGGTCTTGCGGCCGCAGACCCGGCCACGCGGCCGGTGGATTATCAA	WWIWPAMDAAGICCQSVCSDWCE#SDFRLEWQTV**KSRAEAHTTRECQGGYNPMKRSMKSGSFAIGLAMMLAPMVAAPGLAAADPATRPVDYQ**	**Y**	
S45	Rv1566c	CATCTGTTGTCAAAGTGTCTGTAGTGATTGGTGTGAATAAAGTGACTTCCGGCTGGAATGGCAGACGGTGAAGTCACGTGCTGAGGCACACACCACCCGAGAGTGCCAAGGAGGTTACAACCCCATGAAACGCAGCATGAAAAGCGGCTCCTTCGCGATCGGTCTGGCAATGATGCTCGCCCCGATGGTGGCCGCGCCCGGTCTTGCGGCCGCAGACCCGGCCACGCGGCCGGTGGATTATCAACAGATCACCGACGTC	ICCQSVCSDWCE#SDFRLEWQTV**KSRAEAHTTRECQGGYNPMKRSMKSGSFAIGLAMMLAPMVAAPGLAAADPATRPVDYQQITDV**	**Y**	
S46	Rv1566c	GATGGTGGCCGCGCCCGGTCTTGCGGCCGCAGACCCGGCCACGCGGCCGGTGGATTATCAACAGATCACCGACGTCGTGATCGCGCGCGGGCTGTCGCAGCGCGGCGTGCCGTTCTCCTGGGCCGGCGGCGGCATCAGCGGCCCCACGCGCGGCACCGGTACCGGCATCAACACCGTCGGGTTCGACGCCTCCGGTTTGATCCAGTACGCCTATGCCGGTGCCGGGCTAAAGCTGCCGCGTTCTTCCGGCCAGATGTACAAGGTTGGGCAAAAGGTCCTGCCGCAGCAAGCGCGCAAGGGCGACCTGATCTTCTACGGCCCCGAAGGCACGCAAAGCGTCGCGTTATACCTCGGGAAGGGCCAGATGCTGGAGGTGGGCGACGTCGTCCAGGTTTCGCCGGTGCGCACCAACGGCATGACGCCTTACCTGGTCCGGGTTCTCGGGACCCAGCCGACGCCCGTCCAACAGGCGCCGGTCCAGCCAGCGCCGGTCCAGCAAGCGCCCGTCCAGCAAGCGCCCGTCCAACAGGCG	**MVAAPGLAAADPATRPVDYQQITDVVIARGLSQRGVPFSWAGGGISGPTRGTGTGINTVGFDASGLIQYAYAGAGLKLPRSSGQMYKVGQKVLPQQARKGDLIFYGPEGTQSVALYLGKGQMLEVGDVVQVSPVRTNGMTPYLVRVLGTQPTPVQQAPVQPAPVQQAPVQQAPVQQA**	**Y**	

ID	Rv no.	DNA Insert Sequence	Peptide Fragment	SP	TM
S47	Rv1613	CAAGTAGGCTCGGGCCGGTTTTCGATTCCTGCCGTGCAAACAACCG CGCGGCATTGATTGGTTACTTGCCGACCGGGTACCCGGACGTGCCA GCGTCGGTGGCCTTCGAAGCGAGCGGCTTGAACAGGTCGATGGCG CGGTCACCGGCTTCGATGTCGACCCCGGCCGACGCGTAGGTGATAC CCCGACTACCCGGGTCTTTTCCG	**SRLGPVFDSCRANNRAALIGYLPTGYPDVPASVA**FEASGLNRSMARSPASMST PADA#VIPRLPGSFP	**N**	
S48	Rv1621c	AGATGTTTTTCACCCCGGCCACCGGTGCCGTGATCGGGCTCTATGAC TTGAAGTCAATGGCCATTGTGGTGATCACACTGCCCCTGATACCGAT CTTCATGGTGCTGATCGGGCTGGCTACCACTAACCCCTCGGCGGCC GCGCTGGCGGCC	MFF**TPATGAVIGLYDLKSMAIVVITLPLIPIFMVLIGLATTNPSAAALAA**		
S49	Rv1621c	TGAAGTCAATGGCCATTGTGGTGATCACACTGCCCCTGATACCGATC TTCATGGTGCTGATCGGGCTGGCTACCACTAACCCCTCGGCGGCCG CGCTGGCGGCCATGACCGCCGTCCAGGCC	**KSMAIVVITLPLIPIFMVLIGLATTNPSAAALAAMTAVQ**		
S50	Rv1634	CGCTGATCAACTCGACCTTGCCCAAGTCGCTGTGGACCCGTGGCTCA GCACTGGTGTCGGCGATGTGGGGGGTCGCGACGCTGATCGGACCG GCGACCGGAGGCCTTTTCGCGCAGCTCGGGCTGTGGCGATGGGCG TTCGGCGTGATGACGTTGCTGACCGCGTTGATGGCCATGTTGGTGC CGGTCGCGCTCGGTGCCGGGGGGGTCGGCCCGGGCGGCGAGACG CCGGTG	**LINSTLPKSLWTRGSALVSAMWGVATLIGPATGGLFAQLGLWRWAFGVMTL LTALMAMLVPVALGAGGVGPGGETPV**	**N**	**TM**
S51	Rv1638	CAGTACGCGCAGCGGCAGCGCGCGGCAGCCGAACGGCTAGCTCAG ATCCGTGAGTCGATGCACACCGATGAGTGAGACCGCACCCAGTTCG CCGCCGTGAGCACAGACTTGACGCGCGGCGATAGTCCAGCTCAGGC GCTGACGTTGCGCCGTCTGTTCGACCGCGATGTGGCGGCCGAGGCA CCGCCGCCGACGACCTCAGCGAGAAACTTCCCGGTGTAGCTCGCCG GCACCGCGGCAACGTCCTCCGGAGTGCCTTGGGCGACAACGGTTCC GCCGCCGGCACCGCCCTCCGGGCCCAGGTCGATGATCCAATCCGAT GTCTTGATCACGTCCAGGTTATGTTCGATGACGATCACCGTATTGCC CTTGTCGACCAGGCCGTTGATCACGTTGAGCAGCTTGCGTATGTCGT CG	**DDIRKLLNVINGLVDKGNTVIVIEHNLDVIKTSDWIIDLGPEGGAGGGTVVAQ GTPEDVAAVPASYTGKFLAEVVGGGASAATSRSNRRRNVSA**&AGLSPRVKSV LTAANWVRSHSSVCIDSRI&ASRSAAARCRCAY	**N**	

ID	Rv no.	DNA Insert Sequence	Peptide Fragment	SP	TM
S52	Rv1700	MTCACGAAKAAKCCGACATGACGATGGGGTGGTATCCCATTGCCKAAGCGGCTCGCCGGGTGCTGCGTGSCAAAATCGTCAATTCCATTGCCATTGCCGGTGTTTTGGCCGTGCACGCGGTGACSACCGGGTTCSCCCAGCCACGCCCACTCSATACCGAATGGATCGACAGGCCAACGGCGTTCGCCGCGCGGARAGCCGAG	HEDMTMGWYPIA AARRVLR KIVNSIAIAGVLAVHAVTTGF**QPRPL TEWIDRPTAFAARAE**	**N**	
S53	Rv1733c	CAACCCGCGATCGTGAAGGAGCCACCATGATCACGTTTAGGCTGCGCTTGCCGTGCCGGACGATACTGCGGGTGTTCAGCCGCAATCCGCTGGTGCGTGGGACGGATCGACTCGAGGCGGTCGTCATGCTGCTGGCCGTCACGGTCTCGCTGCTGACTATCCCGTTCGCCGCCGCGGCCGGCACCGCAGTCCATGATTCCCGCAGCCACGTCTATGCCCACCAGGCCCAGACCCGCCATCCCGCAACCGCGACCGTGATCGATCACGAGGGGGTGATCGACAGCAACACGACCGCCACGTCAGCGCCGCCGCGCACGAAGATC	**TRDREGATMITFRLRLPCRTILRVFSRNPLVRGTDRLEAVVMLLAVTVSLLTIPF AAAAGTAVHDSRSHVYAHQAQTRHPATATVIDHEGVIDSNTTATSAPPRTKI**	**N**	**1xTM**
S54	Rv1804c	ACAAGAAGGCGTACTCGACCTGAAAGACGTTATCCACCATACGGATAGGGGATCTCAGTACACATCGATCCGGTTCAGCGAGCGGCTCGCCGAGGCAGGCATCCAACCGTCGGTCGGAGCGGTCGGAAGCTCCTATGACAATGCACTAGCCGAGACGATCAACGGCCTATACAAGACCGAGCTGATCAAACCCGGCAAGCCCTGGCGGTCCATCGAGGATGTCGAGTTGGCCACCGCGCGCTGGGTCGACTGGTTCAACCATCGCCGCCTCTACCAGTACTGCGGCGACGTCCCGCCGGTCGAACTCGAGGCTGCCTACTACGCTCAACGCCAGAGACCAGCCGCCGGCTGAGGTCTCAGATCAGAGAGTCTCCGGACTCACCGGGGCGGTTCAATTCCGGGAGAATCGTTGTCTATGAGAGTTGTGTCAACGCTACTCAGCATTCCGTTGATGATCGGCTTGGCGGTTCCGGCCCACGCGGGGCCCAGCGGTGACGACGCGGTCTTTCTTGCCTCGCTAGAGCGGGCAGGCATTACCTACAGCCACCCGGATCAAGCCATAGCATCGGGCAAG	QEGVLDLKDVIHHTDRGSQYTSIRFSERLAEAGIQPSVGAVGSSYDNALAETING LYKTELIKPGKPWRSIEDVELATARWVDWFNHRRLYQYCGDVPPVELEAAYYA QRQRPAAG&GLRSESLRTHRGGS**IPGESLSMRVVSTLLSIPLMIGLAVPAHAG PSGDDAVFLASLERAGITYSHPDQAIASGK**	**Y**	

ID	Rv no.	DNA Insert Sequence	Peptide Fragment	SP	TM
S55	Rv1808	ACCGCGCTTTGTTAATGTCGCTGGTCGCCACGAACATCTTCGGGCAGAACACACCGGCGATCGCGGCCACCGAGGCCCACTACGCGGAGATGTGGGCGCAAGATGCGGCCGCGATGTATGGCTATGCCGGCTCGTCGGCCACTGCGTCGCAGTTGGCGCCGTTCAGCGAGCCGCCGCAAACGACCAATCCGTCGGCAACGGCCGCTCAATCAGCCGTCGTCGCCCAGGCC	**RALLMSLVATNIFGQNTPAIAATEAHYAEMWAQDAAAMYGYAGSSATASQLAPFSEPPQTTNPSATAAQSAVVAQA**	**N**	
S56	Rv1813c	GGCTCAACCCTTTCAAACCGCTGGATTACCGACCGCAGAAAGGGGGCAGGACATGATCACAAACCTCCGACGCCGAACCGCGATGGCAGCCGCCGGCCTAGGGGCTGCTCTCGGGCTGGGCATCCTGCTGGTTCCGACGGTGGACGCCCATCTCGCCAACGGTTCGATGTCGGAAGTCATG	AQPFQTAGLPTAERGQD**MITNLRRRTAMAAAGLGAALGLGILLVPTVDAHLANGSMSEVM**	**Y**	
S57	Rv1886c	GTAACGACTTTGCGCCCGAATCGACATTTGGCCTCCACACACGGTATGTTCTGGCCCGAGCACACGACGACATACAGGACAAAGGGGCACAGGTATGACAGACGTGAGCCGAAAGATTCGAGCTTGGGGACGCCGATTGATGATCGGCACGGCAGCGGCTGTAGTCCTTCCGGGCCTGGTGGGGCTTGCCGGCGGAGCGGCAACCGCGGGCGCGTTCTCCCGGCCGGGGCTGCCGGTCGAGTACCTGCAGGTGCCGTCGCCGTCGATGGGCCGCGACATCAAGGTTCAGTTCCAGAGCGGTGGGAACAACTCACCTGCGGTTTATCTGCTCGACGGCCTGCGCGCCCAAGACGACTACAACGGCTGGGATATCAACACCCCGGCGTTCGAGTGGTACTACCAGTCGGGACTGTCGATAGTCATGCCGGTCGGCGGGCAGTCCAGCTTCTACAGCGACTGGTACAGCCCGGCCTGCGGTAAGGCTGGCTGCCAGACTTACAAGTGGGAAACCTTC	#RLCARIDIWPPHTVCSGPSTRRHTGQRGTG**MTDVSRKIRAWGRRLMIGTAAAVVLPGLVGLAGGAATAGAFSRPGLPVEYLQVPSPSMGRDIKVQFQSGGNNSPAVYLLDGLRAQDDYNGWDINTPAFEWYYQSGLSIVMPVGGQSSFYSDWYSPACGKAGCQTYKWETF**	**Y**	
S58	Rv1916	GGAAGCGCCGCATCTCCTCGTCGGTCATGCCGGTGGTGTCCCAGTTGAACGATGGTGAGAGGTTGTACGCCAGCATCTGGTCGGGGAACTCGGCATGGATCGCCTCGGCGAACTGTCGAGCGTCGGCGAGATCGGCGGTCTTGGTCTCCATCCAAAGAATGTCGGCAAACGGTGCCGCGGCCAGCGATTTGGCGATCGCATACGGTATGCCGCCGCGGATCTGCTGGGAGGCCTGCTCTTGCTGTTCGTAGTTGTTGGCGTCGCGAACCAGCCCGTCACGCACCCCG	GVRDGLVRDANNYEQQEQASQ**QIRGGIPYAIAKSLAAAPFADILWMETKTADLADARQFAEAIHAEFPDQMLAYNLSPSFNWDTTGMTDEEMRRF**	**N**	

ID	Rv no.	DNA Insert Sequence	Peptide Fragment	SP	TM
S59	Rv1926c	AATCTGGAGTTAGCAGCGGGTCATTTGCGGCTTAAGGTAATGACGTCGGCGAGGTTCGAACCAGGTAATCGCCCCAACAAGTAGTGGAGGTAGGGACCAATGAAGCTCACCACAATGATCAAGACGGCAGTAGCGGTCGTGGCCATGGCGGCCATCGCGACCTTTGCGGCACCGGTGGTGTTGACGGCGCCGCCGACGGGGCGGGCGCCG	NLELAAGHLRLKVMTSARFEPGNRPNK#WR#GP**MKLTTMIKTAVAVVAMAAIATFAAPV**VLTAPPTGRAP	**Y**	
S60	Rv1926c	TATCGAATCTGGAGTTAGCAGCGGGTCATTTGCGGCTTAAGGTAATGACGTCGGCGAGGTTCGAACCAGGTAATCGCCCCAACAAGTAGTGGAGGTAGGGACCAATGAAGCTCACCACAATGATCAAGACGGCAGTAGCGGTCGTGGCCATGGCGGCCATCGCGACCTTTGCGGCACCGGTCGCGTTGGCTGCCTATCCCATCACCGGAAAACTTGGCAGTGAGCTAACGATGACCGACACCGTTGGCCAAGTCGTGCTCGGCTGGAAGGTCAGTGATCTCAAATCCAGCACGGCAGTCATCCCCGGCTATCCGGTGGCCGGCCAGGTCTGGGAGGCCACTGCCACGGTCAATGCGATTCGCGGCAGCGTCACGCCCGCGGTCTCGCAGTTCAATGCCCGCACCGCCGACGGCATCAACTACCGGGTGCTGTGGCAAGCC	SNLELAAGHLRLKVMTSARFEPGNRPNKWRGP**MKLTTMIKTAVAVVAMAAIATFAAPVALAAYPITGKLGSELTMTDTVGQVVLGWKVSDLKSSTAVIPGYPVAGQVWEATATVNAIRGSVTPAVSQFNARTADGINYRVLWQA**	**Y**	
S61	Rv1926c	CGAGCGCGCGGCTCTCACCGGGTTCGGCGCCCGACGGGCCGGGTATGGCGGTCACCACACCATCCTGACATGCGGTTTGGCTCCCTTGCGCTGGTCGCCTACGACTCGGCCATCAAGCATTCATGGCCACGCCCGTCGTCGGTGCGCCGGCTACGGATGTGAACGGCATCGTCTGCCCCACGGTTCCTATCGAATCTGGAGTTAGCAGCGGGTCATTTGCGGCTTAAGGTAATGACGTCGGCGAGGTTCGAACCAGGTAATCGCCCCAACAAGTAGTGGAGGTAGGGACCAATGAAGCTCACCACAATGATCAAGACGGCAGTAGCGGTCGTGGCCATGGCGGCCATCGCGACCTTTGCGGCACCGGTCGCGTTG	RARGSHRVRRPTGRVWRSPHHPDMRFGSLALVAYDSAIKHSWPRPSSVRRLRM&TASSAPRFLSNLELAAGHLRLKVMTSARFEPGNRPNK#WR#GP**MKLTTMIKTAVAVVAMAAIATFAAPVA**L	**Y**	

ID	Rv no.	DNA Insert Sequence	Peptide Fragment	SP	TM
S62	Rv1980c	GACTCAGATATCGCGGCAATCCAATCTCCCGCCTGCGGCCGGCGGTGCTGCAAACTACTCCCGGAGGAATTTCGACGTGCGCATCAAGATCTTCATGCTGGTCACGGCTGTCGTTTTGCTCTGTTGTTCGGGTGTGGCCACGGCCGCGCCCAAGACCTACTGCGAGGAGTTGAAAGGCACCGATACCGGCCAGGCGTGCCAGATTCAAATGTCCGACCCGGCCTACAACATCAACATCAGCCTGCCCAGTTACTACCCCGACCAGAAGTCGCTGGAAAATTACATCGCCCAGACGCGCGACAAGTTCCTCAGCGCGGCCACATCGTCCACTCCACGCGAAGCCCCCTACGAATTGAATATCACCTCGGCCACATACCAGTCCGCGATACCGCCGCGTGGTACG	TQISRQSNLPPAAGGAANYSRRNFDV**RIKIFMLVTAVVLLCCSGVATAAPKTYCEELKGTDTGQACQIQMSDPAYNINISLPSYYPDQKSLENYIAQTRDKFLSAATSSTPREAPYELNITSATYQSAIPPRGT**	**Y**	
S63	Rv1980c	ATTTCGACGTGCGCATCAAGATCTTCATGCTGGTCACGGCTGTCGTTTTGCTCTGTTGTTCGGGTGTGGCCACGGCCGCGCCCAAGACCTACTGCGAGGAGTTGAAAGGCACCGATACCGGCCAGGCGTGCCAGATTCAAATGTCCGACCCGGCCTACAACATCAACATCAGCCTGCCCAGTTACTACCCCGACCAGAAGTCGCTGGAAAATTACATCGCCCAGACGCGCGACAAGTTCCTCAGCGCGGCCACATCGTCCACTCCACGCGAAGCCCCCTACGAATTGAATATCACCTCGGCCACATACCAGTCCGCGATACCGCCGCGTGGTACGCAGGCCGTGGTGCTCAAGGTCTACCAGAACGCCGGCGGCACGCACCCAACGACCACGTACAAGGCCTTCGATTGGGACCAGGCCTATCGCAAGCCAATCACCTATGACACGCTGTGGCAGGCTGACACCGATCCGCTGCCAGTCGTCTTCCCCATTGTGCAAGGTGAACTGAGCAAGCAGACCGGACAACAGGTATCGATAGCGCCG	FDV**RIKIFMLVTAVVLLCCSGVATAAPKTYCEELKGTDTGQACQIQMSDPAYNINISLPSYYPDQKSLENYIAQTRDKFLSAATSSTPREAPYELNITSATYQSAIPPRGTQAVVLKVYQNAGGTHPTTTYKAFDWDQAYRKPITYDTLWQADTDPLPVVFPIVQGELSKQTGQQVSIAP**	**Y**	
S64	Rv1981c	TTCTACTCTGGGTTCTACCTGCCGATGTACTGGTCGAGTCGGGCCAAGTTGACCAACACCGCCGACATGATCCGGCTGATCATCCGCGACGAGGCCGTGCACGGTTACTACATCGGCTATAAGTTCCAGCGTGGTCTGGCGTTGGTTGACGACGTCACGCGCGCCGAGCTCAAGGACTACACCTACGAGCTACTGTTCGAGCTCTACGACAACGAGGTGGAATACACCCAGGACCTCTACGACGAGGT	**FYSGFYLPMYWSSRAKLTNTADMIRLIIRDEAVHGYYIGYKFQRGLALVDDVTRAELKDYTYELLFELYDNEVEYTQDLYDE**F	**N**	

ID	Rv no.	DNA Insert Sequence	Peptide Fragment	SP	TM
S65	Rv1988	CCACGGCGACCACCCGCGCCCCCGCTCGCACTAGATGCGCCGTCAGTGCCCCTTCGCCGGCGCCGATGTCAAACACGAGCTCACCGGGCCGCACTGCGGCCGCGCTGACTACCCGCGCTGCCCATTCGTCATGGAGCCGGTGCCAGCCCCATGCCCGTCGCGACCGTCCGAGGGCGGACACGACGTACCGTCACTGCGTAGATGCCCACGCGCCCGACCGTAGCCCGCCACCGGCACTGCGATCAATCCAATTTCTCGGTTCAGGCAACCTTCTGGTCATCACCAGCCCCAGGGCTCTGGCGCCGTCCGCATCAACTCCGAGATGACGTTGGCCGTGACGACCCACTAGACCCACCTGGCAGTAGCCGCATTGTTGCAGTCGGCGAGCCTCAGTGCGCAGTCGCGTCTAGGTGCAAGGATATTGCCCGTTGAGCAGACAACTCGACGGCGGCGAGTAAGAACCGGTCAGCCCGCCTCTTAGGCCGCCCGTGGCTGAACCACCGGGGGCAATGATGCGATTCCAATTCGCTGGGCTGAGAACGTAGTGCGTGCCAGATCGTGCAACGG	VARSGTHYVLSPANWNRIIAPGGSATGGLRGGLTGSYSPPSSCLLNGQYPCT@TRLRTEARRLQQCGYCQVGLVGRHGQRHLGVDADGARALGLVMTRRLPEPRNWIDRSAGGGLRSGAWASTQ&RYVV**SALGRSRRAWGWHRLHDEWAARVVSAAAVRPGELVFDIGAGEGALTAHLVRAGARVVAV**	**N**	
S66	Rv2239c	CATGTATTGCCCGCCGAGATTGCCGAGGCGGCGCCCACAGCTGGGCTGATGCCGACCTCGTCGGTCAATCTGGGCAACTGGAGCGCCAGCAGATTGGTACAGCCGAAATAGTTGGACATGTA	**HVLPAEIAEAAPTAGLMPTSSVNLGNWSASRLVQP**K@LDMH	**N**	
S67	Rv2301	TCTACGCTTATGCAATGACAGATCACGACCATCCCACTGTGAATGATTTACTGACCCGCCGACTGCTCACCATGGGCGCGGCCGCCGCAATGCTGGCCGCGGTGCTTCTGCTTACTCCCATCACCGTTCCCGCCGGCTACCCCGGTGCCGTTGCACCGGCCACTGCAGCCTGCCCCGACGCCGAAGTGGTGTTCGCCCGCGGCCGCTTC	YAYAMTDHDHPTV**NDLLTRRLLTMGAAAAMLAAVLLLTPITVPAGYPGAVAPATAACPDAEVVFARGRF**	**Y**	
S68	Rv2376c	TGGCCGAGGACTGACCGCTCGCTCACAGCTACGACACAGACTTGCCCGGCGCGTGCACCGGTAGTTTGAACCAAACGCACAATCGACGGGCAAACGAACGGAAGAACACAACCATGAGATGGTGAAATCGATCGCCGCAGGTCTGACCGCCGCGGCTGCAATCGGCGCCGCTGCGGCCGGTGTGACTTCGATCATGGCTGGCGGCCCGGTCGTATACCAGATGCAGCCGGTCGTCTTCGGCGCGCCACTGCCGTTGGAC	GRGLTARSQLRHRLARRVHR#FEPNAQSTGKRTEEHNHE**MVKSIAAGLTAAAAIGAAAAGVTSIMAGGPVVYQMQPVVFGAPLPLD**	**Y**	

ID	Rv no.	DNA Insert Sequence	Peptide Fragment	SP	TM
S69	Rv2391	CGGCAACCCGCTCGACGTGCGGGAGCGCATCGAAAACATCTACGCCAAACAGGGTTTCGACAGCATCGACAAGACCGACCTGCGAGGGCGCTTTCGCTGCCGGTCGAGTTCGTCGGTGGTTACCCCGGGCGCGACCGCCTTGCCCGCCTCGGCCACCTTCGCGGCCTGAGCGCAGCTACGCATCCTGACGATCATCACCCCGCCCCCGGCTCACGCTTGGCCTCCGTGACCGCACGCATCGCCCGGTTGCGCGCACCG	**GNPLDVRERIENIYAKQGFDSIDKTDLRGRFR**CRSSSSVVTPGATALPASATFAA&AQLRILTIITPPPAHAWPP&PHASPGCAHR	**P**	
S70	Rv2693c	GCGATTCTGCGTCTGCTCTCCCACCTAGCGGTAGTCGCTGTAGCCGTAGTCGTCCAGCGGGACGGCAGCACCGGTGGCCGCACCGAAGTCCGGGCTGTAGTACTGATCCTCATACGACGGGATGGTGTACGCCGCAGCGCGGGCCTCCTCGGTGGGCTGCACCGCGACGTTGCTGTGGGCGGTGGTCTTCACACTGTCGATCCTGATCCGGCGGCCGATAGTCGGCTACTTGTGGAGCTGGCTCAGCGGGCGCGATCGCGCCTGGCGCGACGTGTCCCGCGCTGTCTT	RFCVCSPT@R@SL@P@SSSGTAAPVAAPKSGL@Y&SSYDGMVYAAARASSVGCTAT**LLWAVVFTLSILIRRPIVGYLWSWLSGRDRAWRDVSRAV**L	**P**	**TM**
S71	Rv2729c	CTGAAGAATATGTCGGCCATTTCACGTTGTTCCACCTCTCTTTGCGTCACGCTCTGTGGTGGCTGGGCAGCCTGGCCGCAGTCGCCAGTTTCACCCTGCAGGCCATTGCGCTGACGATGGGTTCGGTGGTGTTGGTGCAGTCGCTGCAGGCCACCGCACTGTTGTTCGCGCTGCTGATCGATGCTCGGTTGACTCATCACCGCTGTACTCCCAGGGAGTGGATGTGGGCGGTATTGCTGGCCGGCGCGGTGGCCGTCATCGTCATGTCGGGCAACCCGGCGGCCGGCACTACTCGGGCCCCCTTCTCGACCTGGGCGGTGGTTGCTGTCGTG	**EEYVGHFTLFHLSLRHALWWLGSLAAVASFTLQAIALTMGSVVLVQSLQATALLFALLIDARLTHHRCTPREWMWAVLLAGAVAVIVMSGNPAAGTTRAPFSTWAVVAVV**	**P**	**TM**
S72	Rv2768c	TTCAGAACGCGATCAACGGCGCGGTCAACACCACCGCCTGGTTCGTCATGGCCACCATCCCCAACGCGGTATTCCTCGGACACGCCTTTGCTGCCCTAAACCCGGCAACCGTGACCGCAGCCGCCGATGCCGTTCCAGCTGCCGCGGCAGCAGCTGGTTTGGCGCACACGGTGACGCCAGTGGGCGTTGGTGGGGCCTCGCTGACGGCGAGTCTGGGCGAGGCGTCCTCGGTCGGTGGCCTGTCGGTCCCGGCCGGTTGGTCGACCGCTGCTCCGGCCATGACGTCTGGTACCACGGCACTGGAGGGCTCGGTCTGGGCGGTCCCCGAGGAAGCCGGGCCA	**QNAINGAVNTTAWFVMATIPNAVFLGHAFAALNPATVTAAADAVPAAAAAAGLAHTVTPVGVGGASLTASLGEASSVGGLSVPAGWSTAAPAMTSGTTALEGSVWAVPEEAGP**	**N**	

ID	Rv no.	DNA Insert Sequence	Peptide Fragment	SP	TM
S73	Rv2878c	AGGGGGCGCCGGGATGTACTGAACCGTCACGGAGCGTATTACGTCATTAGGTCAAAACACGAACAAGTGAGGTCTGTCATGAGTCTTCGCCTGGTGTCCCCGATCAAGGCGTTTGCGGACGGCATTGTGGCCGTTGCTATCGCGGTTGTCCTGATGTTCGGTCTGGCCAATACACCGCGAGCGGTGGCAGCCGATGAACGTCTGCAGTTCACCGCA	RGRRDVLNRHGAYYVIRSKHEQVRSV**MSLRLVSPIKAFADGIVAVAIAVVLMFGLANTPRAVAADERLQFTA**	**Y**	
S74	Rv2878c	TCATTAGGTCAAAACACGAACAAGTGAGGTCTGTCATGAGTCTTCGCCTGGTGTCCCCGATCAAGGCGTTTGCGGACGGCATTGTGGCCGTTGCTATCGCGGTTGTCCTGATGTTCGGTCTGGCCAATACACCGCGAGCGGTGGCAGCCGATGAACGTCTGCAGTTCACCGCAACCACGCTCAGCGGTGCTCCC	IRSKHEQVRSV**MSLRLVSPIKAFADGIVAVAIAVVLMFGLANTPRAVAADERLQFTATTLSGAP**	**Y**	
S75	Rv2878c	CGAGCGCAGGGGGCGCCGGGATGTACTGAACCGTCACGGAGCGTATTACGTCATTAGGTCAAAACACGAACAAGTGAGGTCTGTCATGAGTCTTCGCCTGGTGTCCCCGATCAAGGCGTTTGCGGACGGCATTGTGGCCGTTGCTATCGCGGTTGTCCTGATGTTCGGTCTGGCCAATACACCGCGAGCGGTGGCAGCC	ERRGRRDVLNRHGAYYVIRSKHEQVRSV**MSLRLVSPIKAFADGIVAVAIAVVLMFGLANTPRAVAA**	**Y**	
S76	Rv2922A	CGTCCGTCGGCGTGGTTGGCCGCGTAACCGGTCAGGCCGAGCTCCAACGCTCGGCAGCGGGTCCACCAGCGG	**RWWTRCRALELGLTGYAANHADG**	**N**	
S77	Rv3009c	CGACGCGACCCAACTCCTGCGCGGACTTTCCGGCGCCAGTGACGTGATCCACGAATTCCTGAGCGAAAACGTGCTGGACGAACTGGCCATCACTCCTGCCCAGGTCGCAGCCGTGGTGGCATTGGTCGATGAGGGCAAGCTGTCCAACAGCTTGGCCCGCCAAGTCGTGGAGGGTGTGCTGGCCGGTGAAGGTGAGCCCGAACAGGTGATGACTGCGAGAGGGTTGGCGTTGGTCCGCGACGACTCGTTGACCCAGGCCGCGGTCGACGAGGCCCTGGCCGCAAATCCTGATGTGGCGGACAAGATTCGCGGCGGCAAGGTGGCCGCGGCCGGCGCGATTGTCGGTGCGGTGATGAAGGCGACCCGCGGACAGGCCGACGCGGCCCGAGTGCGCGAACTCGTTCTAGAGGCCTGCGGGCAGGGTTAGTTGCCGTCCGCGGGTAGCGGTTCATCACCGTCAGGGGAGTTGCGACGATACACGTTGAGTGTCCCGATGGTGGCCG	DATQLLRGLSGASDVIHEFLSENV**LDELAITPAQVAAVVALVDEGKLSNSLARQVVEGVLAGEGEPEQVMTARGLALVRDDSLTQAAVDEALAANPDVADKIRGGKVAAAGAIVGAVMKATRGQADAARVRELVLEACGQG**LPSAGSGSSPSGELRRYTLSVPMVA	**N**	

ID	Rv no.	DNA Insert Sequence	Peptide Fragment	SP	TM
S78	Rv3103c	CTGAACGCAGAAATCATGAGGACAATGCATGCCCCGGAGCAGTAG CGTGAAGCTCAGCAACCAGAAACGGCACTGGCCGGGCTATCTGTTC GGCCGGATCCGCACGTCGACCCTGGTGTTGATCGCCGCGTTCCTGG CGGTGTGGTGGATTTACGAGACCTATAGGCCACAGGCACCAGGTCC TGGTGACTCT	&TQKS&GQCMP**RSSSVKLSNQKRHWPGYLFGRIRTSTLVLIAAFLAVWWIYE TYRPQAPGPGDS**	**N**	
S79	Rv3150	ATCTGCTCAAACAGCTTGGCAAGCAACAGATCACACCGGCGGGGAC GACKATCGTGTTCGCCGCCGCGCCGGTGATCGTCGCCGGGACAACG CTTTTGATCGCCGCGATCGCACCTCTGGKGGCCACCGGGTCACCCCT GGACCCCAGTGATCAGSCGCCACTGGGACGACCCGGAGTCGTGGA CCCTGGCCACTTATCAACGCCACGATCGCTATCGGGGCTATCAKGCG TTGCAGAAAGCCCTGACGATGCCGCCCGACGACGTGATCAGCATCG TCAA	ICSNSLASNRSHRRGR SCSPPRR&SSPGQRF&SPRSHLW PPGHPW**TPVI RHWDDPESWTLATYQRHDRYRGY ALQKALTMPPDDVISIVK**	**N**	
S80	Rv3162c	GCATTGTCGGGGTTTTGCGCCGCCGTCTACCTGGTGTGCCGATACG GGGCCGGTGGTGCGGGCGCGTGGTTCGGCAACGGTGGCGCCGGC GGCGTCGGCGGCGSCRCCRCCAT	**ALSGFCAAVYLVCRYGAG**GAGAWFGNGGAGGVGG	**P**	**TM**
S81	Rv3193c	GGTGTTACTTAAAGCCGTTGCTTATTGGCTGGATCGGTATGAGCTG CTGTCGCACACGCGTGGCGGCAAGCCGTTCACCGGTGCCGGGTACA CCGATATCAACGCCGTCCTGCCGGCGAAGCTGATTCTGATGGCGAT TGCGTTGATTTGCGCGGCCGCAGTGTTCTCGGCGATCGCCCTGCGG GACTTGCGGATTCCGGCGATCGGCCTGGTGTTGTTGCTGCTGTCGT CGCTGATTGTCGGCGCCGGCTGGCCGTTGATCGTCGAGCAAATCAG CGTCAAACCCAACGCTGCGCAAAAAGAGAGCGAATATATCAGCCGA AGTATCACCGCAACTCGGCAAGCCTATGGCCTGACGTCT	**VLLKAVAYWLDRYELLSHTRGGKPFTGAGYTDINAVLPAKLILMAIALICAAAV FSAIALRDLRIPAIGLVLLLLSSLIVGAGWPLIVEQISVKPNAAQKESEYISRSITA TRQAYGLTS**	**P**	**TM**

ID	Rv no.	DNA Insert Sequence	Peptide Fragment	SP	TM
S82	Rv3222c	TGTCCGGTGACATGCTCGGTGTTGGGTGGGGTCGTAAGCGATGCCGTCTTGCCAGCGGTGTCAGATGACGTAAAGCCAGGCGCGGGCGAGGATGCGTACCGCGTGCCGGTGGTCGTGGCCGCGGGCTCGGGTGCGGTTGTGCAGGTCGGCGGCCTAGAGGTTGGCTCGGCGGCTGTCGCCGGCGAAGTCGCAGACACCGTTGCGGAGTTGTTTGTCTGCCGCCCAAACGAACCCGACGTGGGTGACTTTGTCGGACTAGCCGGTGGAGCGGGCGACGCCGGCCAAGCAGGCCAGCAATTCGGGCTGGGCGTCGGCGTGCGGGGCGAGTCGTTCGGCGCTCGTCGGCGCTTGGCCCTGTCGACGGTCGGCGCGTCCGGGGCAACCGCCGGACTCCGCAAAACTCATGATGGACATCACGGCTGTCAAGCGCGCGGGGCGCTAACCCAGCGGCGCCTCTACATCGGCAATCCATCCGA	CPVTC**SVLGGVVSDAVLPAVSDDVKPGAGEDAYRVPVVVAAGSGAVVQVGGLEVGSAAVAGEVADTVAELFVCRPNEPDVGDFVGLAGGAGDAGQAGQQFGLGVGVRGESFGARRRLALSTVGASGATAGLRKTHDGHHGCQARGALTQRRLYIGNPS**D	**Y**	
S83	Rv3280	TCTTCAGCCAGGACGCCACGGTGTTTGGCGGCAGCCTTGGCGAGGTGTACGGCGAGAAAATCGTCAAGGTCCAGGAACTGGCGATCAAGACCGGCCGTCCGCTCATCGGCATCAACGACGGTGCTGGCGCGCGCATCCAGGAAGGTGTCGTCTCGCTGGGCCTG	**FSQDATVFGGSLGEVYGEKIVKVQELAIKTGRPLIGINDGAGARIQEGVVSLGL**	**N**	
S84	Rv3310	TCGACGACGGCGAAGTGTGGCTGCGCAACGCGCACATCCCGGAATACCGGCACGGCAGCTGGACCCTGACCAGGGTATACCGTGCCTTGGCGGTGATCGGTGTCCTGGCAGCATCGTTGCTGGCCTCATGGGTCGGCGCTGTCCCACAAGTGGGTCTGGCAGCGAGTGCCCTGCCGACCTTCGCGCACGTG	DDGEVWLRNAHIPEYRHGSWT**LTRVYRALAVIGVLAASLLASWVGAVPQVGLAASALPTFAHV**	**Y**	
S85	Rv3343c	CTGACAAGGAGCTCGCGGCGCAGACAATCATGGGCCACAACCTGATGGCGGTTCCCGTCGTCGATGCCGACAACCGGCTACTGGGCACCATCGACCGAATCCCAGTGGTTCTCGACGTCAACGCGCTGCTCGGCCCCATCAACGCCGGGTTGGTCATCCCGCCCGTCCCGGGATTCGGCAACACCACCGCGGTCCCGTCGTCGGGGTTCTTCAACATCGGCGGTGGCGGCGGGTTGTCGGGCTTCCACAACCT	DKELAAQTIMGHNLMAVPVVDADNRLLG**TIDRIPVVLDVNALLGPINAGLVIPPVPGFGNTTAVPSSGFFNIGGGGGLSGFHNL**	**N**	
S86	Rv3365c	CTAGCTCGAACAGGTTGGCACCGGCCCGCGGCAAACCTCAGCAACGCCCGCCGTCCTGGTCGCCGCGCAACTGGCCGGTCCGATGGAAAGTGTTCACGATCGCGCTTCTGCCGCTGGTAGTGGCGATGGTGTTAGCAGGATTGCGGGTCGAGGCTGCGATGGCCAGCACCAGCGGCCTGCGGCTGGTC	SSNRLA**PARGKPQQRPPSWSPRNWPVRWKVFTIALLPLVVAMVLAGLRVEAAMASTSGLRLV**	**Y**	**TM**

ID	Rv no.	DNA Insert Sequence	Peptide Fragment	SP	TM
S87	Rv3395A	GCTGCCTTGTGCAATCGCGTAAAACCACCTCGGTACTGGCGGCTGCCCTGCTGTTTTGCGGCCTGTTAGGCCCAGGGACGGCCCCACCGGCCACCGGTGGCGGGCCTGCCTGCCGGCCGGCAGAGCTCTTCGCCACCGACAACACCACCGATGGGTTCGAGCTACCGGCCGTTGCGACTATCGCACTAACCGGCACGGTGGTGACCGGATCGACCCTGGTCGACGGCGTGTTCTGGTCGAATGAGCGCCAGCAGATCGGCTACGAGCGCTCCCGTGAATTTCATCTGTGCGTTGTCGACGCGCCCACATTGCACAACGCCGCCGAGGCACTGCACCGCACCACGTCT	**CLVQSRKTTSVLAAALLFCGLLGPGTAPPATGGGPACRPAELFATDNTTDGFELPAVATIALTGTVVTGSTLVDGVFWSNERQQIGYERSREFHLCVVDAPTLHNAAEALHR**TTS	**Y**	**TM**
S88	Rv3410c	CCAGCGGACTGGGCCACGCCTAAAGACTTAGCTCTCTTTCGCGAGCGCGACCGCTTCGGTGCACTTCATTTCGCCGACAATCACGGCACCAAGGCCAGGGATTTCCAACGACGGCGATCCGGTTGCTGCAGGAGCTGCACGCGGCACCGCTAAATCCCGACCTGTTGGGTGCCGCGGTGGCTCGCATCCGCGAGGCCGGGGTGACCACCGCGGTGCGGGTGAGCCCGCAAAACGCCCAGTGGCTGACACCGGTACTGGTTGCGGCCGGTATTGACCTTGCTGGTCATCCAGGG	SGLGHA#RLSSLSRARPLRCTSFRRQSRHQGQGFPT**TAIRLLQELHAAPLNPDLLGAAVARIREAGVTTAVRVSPQNAQWLTPVLVAAGIDLLVIQ**	**N**	
S89	Rv3469c	TGATCGGGATCGCTTGGTGCGCGCTGTTCCCTATCATCTGGGCGCTGTCGGGCTCCCTGAAGGCGGACGGCGAGGCACTCGATCTGGCAATTCCCGGCCCCGCGACCGAAACCCATCAGCGTTCCATCCAGGAAATCGGCCCCGGCGTCGAATGCCTCCAAGGTGTTGGCGACGGCCATGGCGAGGTTGTTGTGCCCGTGGAAGCCGACGGAGACATCGCTGGCACCGCG	**RGASDVSVGFHGHNNLAMAVANTLEAFDAGADFLDGTLMGFGRGAGNCQIECL**AVRLQGARQRPDDREQRAPSDPD	**N**	
S90	Rv3616c	GGCAAGCAATTTGAGGAGTTGAGTCGCGTTGATCAGCGTTTTCACGACCAAGTAGGCAAGCGCGCCGCCCACTACGGCCATCGCGCCCGCGCAAAACGGCGCCTGGAAGGCGGCCGATAGGGCGTGCCCGACGACCGGGATGTAGGTCAGGTCCACAGCCACCGGGCGCACGAACTCGAGACCTTTCTTGGCGCCCTCCAGGATGTCGCGGGTCGTCTGGACCGCGTTGGCCTGGTCGTGGATCAGGCTGATGAGCTGACGATCGAGGTCTGCCAGTTCCTGGAAAAAATTCACGTGGTTGCGGTTTTTGCCGGCGTATTTGTCCGCGGCCGAACCTAACCAGCCATCACCCGGAAAC	**FPGDGWLGSAADKYAGKNRNHVNFFQELADLDRQLISLIHDQANAVQTTRDILEGAKKGLEFVRPVAVDLTYIPVVGHALSAAFQAPFCAGAMAVVGGALAYLVVKTLINATQLLKLLA**	**N**	

ID	Rv no.	DNA Insert Sequence	Peptide Fragment	SP	TM
S91	Rv3646c	TTGGACATCGTCGAGTTGACCCCAGACGGCCATGCCACCAACCCGCCGGCCCGCTACACCGAGGCGTCGCTGGTCAAAGCGCTCGAGGAGCTGGGCATCGGCCGCCCGTCGACCTACCCGTCGATCCCGGCGACCCGGCGAGCGCTACTCCTCACCGCTTTGGGTGGCCTGCTGATTGCCGGGCTGGTCACCGCGATTCCCGCCGTCGGCCGCGCGCCGGAGCGGCTGGCCGGCTACATCGCCAGCAAT	**LDIVELTPDGHATNPPARYTEASLVKALEELGIGRPSTY**PSIPATRRALLLTALGGLLIAGLVTAIPAVGRAPERLAGYIASN	**P**	
S92	Rv3703c	CACCGCCAGCGCGCGGGGCTGACCGCTCCGCAATTCTGGAGGTCGGGCGGGCGGACGCGGACCCGGTTCGGGCACGTCGAGGACATTCCCGCCGATGAGCCGGT	**HRQRAGLTAPQFWRSGGRTRTRFGHVEDIPADEP**	**N**	
S93	Rv3712	AGGAACCCGCCCGTTGACCGCGATGACCACCCCGTCTGCATGCTTGTCGACCATCGCCAGCGCTTCCTGCCAGCCGGCCGGGTTTTTGGCCAGCAGGATCCGGGCTTGGTGCGCGCCGATACGAACGGTCCGGTAGCGTCCGGCGACCTCGTCGACCTGGCAGACGGCGGCGACGGCCACAGCCGGATCGGCGCCGAGGGCGACTGCGGCGGCCACGGCTTGGGCGGCGTTGCCGCGATTCACCGAGCCTGGCAGTGCCAGCCGCATCGGCAGCGCCAGCCCGTCGGGCCCATACAGCGTGGCGTCGTCGAACCACCAGTG	**HWWFDDATLYGPDGLALPMRLALPGSVNRGNAAQAVAAAVALGADPAVAVAAVCQVDEVAGRYRTVRIGAHQARILLAKNPAGWQEALAMVDKHADGVVIAVNGRVP	**P**	
S94	Rv3734c	TGGCCACCCACCTTGATGATCCAGCCGACCGGCTGAACGCCATCCACGCTTCGATGCGCGGTAATAAGAACGTACTATCGCAGCTGCCCCGCGCCCAGGCGTTGGCGGTGTCACTGCTGCTGTTGAGCCCAGCTGCGCTGAACACCCTGCCGGGCTTGGCCAAGGCAACGCCACCGCCGTTTAATGTGTGCATCTCGAACGTGCCCGGTGCGCGCGAACCGCTGTACTTCAACGGCGCCAGGATGGTCGGCAACTATCCAATGTCGTTGGTGCTCGACGGACAAGCGCTCAACATCACCCTGACCAGCACCGCCGAT	**ATHLDDPADRLNAIHASMRGNKNVLSQLPRAQALAVSLLLLSPAALNTLPGLAKATPPPFNVCISNVPGAREPLYFNGARMVGNYPMSLVLDGQALNITLTSTAD**	**N**	
S95	Rv3793	ACACCCGTGTTGAACCCACCCGCGTTGAACAACCCCGTGTCGCCGTCCGGCTACCATCTACCTCCGTGCGCGACGCGGGAGCAAACTACCGGATCGCCCGGTACGTCGCTGTGGTGGCGGGTCTGCTAGGCGCTGTGCTGGCCATCGCCACCCCACTGCTGCCGGTCAACCAGACCACCGCG	TRVEPTRVEQPRV**AVRLPSTSVRDAGANYRIARYVAVVAGLLGAVLAIATPLLPVNQTTA**	**N**	**TM**

ID	Rv no.	DNA Insert Sequence	Peptide Fragment	SP	TM
S96	Rv3835	AACATTGGTCATGCGGCCCCGCCGACCCTGTCAGAATGGAGCGG**ATGTTGGACGCGCCCGAGCAGGACCCCGTCGATCCCGGCGACCCGGC**GAGCGCTACTCCTCACCGCTTTGGGTGGCCTGCTGATTGCCGGGCTGGTCACCGCGATTCCCGCCGTCGGCCGCGCGCCGGAGCGGCTGGCCGGCTACATCGCCAGCAATCCGGTGCCCAGCACTGGCGCCAAGATCAACGCTTCGTTCAACCGCGTCGCCAGTGGTGACTGCTTGATGTGGCCGGACGGCACGCCGGAGTCTGCCGCCGGACGGTATCGCTTCGGCAACCGCAGCGGTGCAGGCAGCCTGGGCCGTACTGGTCGCCTATTTCAACCGATTACG	HWSCGPADPVRMER**MLDAPEQDPVDPGDPAS**ATPHRFGWPADCRAGHRDSRRRPRAGAAGRLHRQQSGAQHWRQDQRFVQPRRQW&LLDVAGRHAGVCRRTVSLRQPQRCRQPGPYWSPISTDY	**Y**	
S97	Rv3859c	TGCTGCAGCTTGAACACAGTCTCCGGGTTGAACAGGTGGTACTCGCCCTCGCGGCGCCACTGGTATTCCCCACCCACCTCGAGTTCGCGGTGAGCGCGTTCGTCCGGCCGGTCCAGATAGGCCAGCCGGTGCCGGGCTGCGACATCGGCCGCGATGTCATCCAGGGTGATCCCGCCGGTGGGGCAGGTAAGCCCGGTGAAGTATTCGTCGAGCACTTGCTCGGAGATGCCGACAGCCTGGAACAGTTGCGCACCGGTGTAGGAGGCCAGCGTCGAGATGCCCATCTTCGACATCACTTTCAGCACACCCTTACCTGCGGCTTTGATGTAGTTGTTCAGCGCCGCCGTACGGTCGATGCCCTCGATAACACCGCGGTCGAGCATGTCCTCGATCGACTCGAACACCAGGTAGGGGTTGATCGCGGCCGCGCCGAATCCGACCAGCGCGGCCATGTGGTGCACCTCGCGGGCATCACCGGACTCGACCACCAGACCCACTTGGGTGCGGGTCCGTTCCCGAACCAGGTGGTGGTGCACTCCCGCAACGGCGAGCAGCGACGGTATCGGAGCCATTTCCTCGTCGGACTCGCGGTCGGACAAGATGATGATCCGAGCGCCGTCGGCGATTGCCGCCGCCGCCGCGCCACGTACCTCTTCCAGCGCGGCAGCCAGCCCAGCACCTCCCTCGGAGACCCGGTACAGACAGCGAATCACCTTGGACCGCAATCCGTGTGG	**PHGLRSKVIRCLYRVSEGGAGLAAALEEVRGAAAAAIADGARIIILSDRESDEEMAPIPSLLAVAGVHHHLVRERTRTQVGLVVESGDAREVHHMAALVGFGAAAAINPYLV**	**N**	

ID	Rv no.	DNA Insert Sequence	Peptide Fragment	SP	TM
S98	Rv3870	GGTAGGTCGGCACCGCGTTCGCGGTACTTCTCGTATTCGGCCACGCCGGACAGGGCTCCGGCCGCGCCGACTTTCATCCCGGCCTGTCGGAGGATCGACTGGCGCCGATCGAGTTCTCCGGTCAACACCTCGCCCATCCGGCTGACGAGCTCGAGCCCGCACAGCCGACGCTCACGGTGCCGTCCGCACAGCCG	RLCGRHRERRLCGL**ELVSRMGEVLTGELDRRQSILRQAGMKVGAAGALSGVAEYEKYRERGADL**	**N**	**TM**

Table S2. 2: Tuberculist function category distribution of 86 identified proteins.

Functional Category	Rv Number	Gene Name	Description
cell wall and cell processes	Rv0116c	*ldtA*	Probable L,D-transpeptidase LdtA
	Rv0179c	*lprO*	Possible lipoprotein LprO
	Rv0203	-	Possible exported protein (Heme-binding protein)
	Rv0236c	*aftD*	Possible arabinofuranosyltransferase AftD
	Rv0320	-	Possible conserved exported protein
	Rv0425c	*ctpH*	Possible metal cation transporting P-type ATPase CtpH
	Rv0559c	-	Possible conserved secreted protein
	Rv0603	-	Possible exported protein
	Rv0676c	*mmpL5*	Probable conserved transmembrane transport protein MmpL5
	Rv0842	-	Probable conserved integral membrane protein
	Rv1029	*kdpA*	Probable potassium-transporting ATPase a chain KdpA
	Rv1200	-	Probable conserved integral membrane transport protein
	Rv1228	*lpqX*	Probable lipoprotein LpqX
	Rv1230c	-	Possible membrane protein
	Rv1291c	-	Conserved hypothetical secreted protein
	Rv1435c	-	Probable conserved proline, glycine, valine-rich secreted protein
	Rv1522c	*mmpL12*	Probable conserved transmembrane transport protein MmpL12
	Rv1634	-	Possible drug efflux membrane protein
	Rv1733c	-	Probable conserved transmembrane protein
	Rv1926c	*mpt63*	Immunogenic protein Mpt63 (antigen Mpt63/MPB63)
	Rv1980c	*mpt64*	Immunogenic protein Mpt64 (antigen Mpt64/MPB64)
	Rv2301	*cut2*	Probable cutinase Cut2
	Rv2376c	*cfp2*	Low molecular weight antigen CFP2
	Rv2693c	-	Probable conserved integral membrane alanine and leucine rich protein

Functional Category	Rv Number	Gene Name	Description
	Rv2729c	-	Probable conserved integral membrane alanine valine and leucine rich protein
	Rv2878c	*mpt53*	Soluble secreted antigen Mpt53 precursor
	Rv3162c	-	Possible integral membrane protein
	Rv3193c	-	Probable conserved transmembrane protein
	Rv3310	*sapM*	Acid phosphatase (acid phosphomonoesterase)
	Rv3395A	-	Probable membrane protein
	Rv3616c	*espA*	ESX-1 secretion-associated protein A, EspA
	Rv3793	*embC*	Integral membrane indolylacetylinositol arabinosyltransferase EmbC
	Rv3835	-	Conserved membrane protein
	Rv3870	*eccCa1*	ESX conserved component EccCa1. ESX-1 type VII secretion system protein.
conserved hypotheticals	Rv0192	-	Conserved hypothetical protein
	Rv0455c	-	Conserved protein
	Rv0822c	-	Conserved protein
	Rv1118c	-	Conserved protein
	Rv1156	-	Conserved protein
	Rv1157c	-	Conserved ala-, pro-rich protein
	Rv1268c	-	Hypothetical unknown protein, probably secreted protein
	Rv1366	-	Hypothetical protein
	Rv1638	-	Conserved hypothetical protein
	Rv1804	-	Conserved protein
	Rv1813c	-	Conserved hypothetical protein
	Rv2239c	-	Conserved hypothetical protein
	Rv3103c	-	Conserved hypothetical protein
	Rv3222c	-	Conserved hypothetical protein
	Rv3365c	-	Conserved protein
	Rv3703c	-	Conserved hypothetical protein

Functional Category	Rv Number	Gene Name	Description
information pathways	Rv1700	-	NUDIX hydrolase
	Rv1981c	*nrdF1*	Ribonucleoside-diphosphate reductase (beta chain) NrdF1
	Rv3009c	*gatB*	Probable glutamyl-tRNA(GLN) amidotransferase (subunit B) GatB (Glu-ADT subunit B)
	Rv3646c	*topA*	DNA topoisomerase I TopA
intermediary metabolism and respiration	Rv0125	*pepA*	Probable serine protease PepA
	Rv0255c	*cobQ1*	Probable cobyric acid synthase CobQ1
	Rv0983	*pepD*	Probable serine protease PepD (serine proteinase)
	Rv1161	*narG*	Respiratory nitrate reductase (alpha chain) NarG
	Rv1254	-	Probable acyltransferase
	Rv1447c	*zwf2*	Probable glucose-6-phosphate 1-dehydrogenase Zwf2
	Rv1613	*trpA*	Probable tryptophan synthase, alpha subunit TrpA
	Rv1621c	*cydD*	Probabletransport transmembrane ATP-binding protein ABC transporter CydD
	Rv1916	*aceAb*	Probable isocitrate lyase AceAb
	Rv1988	*erm(37)*	Probable 23S rRNA methyltransferase Erm(37)
	Rv2391	*sirA*	Ferredoxin-dependent sulfite reductase SirA
	Rv2922A	*acyP*	Probable acylphosphatase AcyP
	Rv3150	*nuoF*	Probable NADH dehydrogenase I (chain F) NuoF
	Rv3410c	*guaB3*	Probable inosine-5'-monophosphate dehydrogenase GuaB3
	Rv3469c	*mhpE*	Probable 4-hydroxy-2-oxovalerate aldolase MhpE (HOA)
	Rv3712	-	Possible ligase
	Rv3859c	*gltB*	Probable ferredoxin-dependent glutamate synthase [NADPH] (large subunit) GltB
lipid metabolism	Rv0101	*nrp*	Probable peptide synthetase Nrp (peptide synthase)
	Rv0129c	*fbpC*	Secreted antigen 85-C FbpC (85C)
	Rv0436c	*pssA*	Probable CDP-diacylglycerol--serine O-phosphatidyltransferase PssA
	Rv0675	*echA5*	Probable enoyl-CoA hydratase EchA5

Functional Category	Rv Number	Gene Name	Description
	Rv0824c	*desA1*	Probable acyl-[acyl-carrier protein] desaturase DesA1
	Rv1886c	*fbpB*	Secreted antigen 85-B FbpB (85B) (antigen 85 complex B)
	Rv3280	*accD5*	Probable propionyl-CoA carboxylase beta chain 5 AccD5
	Rv3734c	*tgs2*	Putative triacylglycerol synthase (diacylglycerol acyltransferase) Tgs2
PE/PPE	Rv1808	*ppe32*	PPE family protein PPE32
	Rv2768c	*ppe43*	PPE family protein PPE43
	Rv3343c	*ppe54*	PPE family protein PPE54
regulatory proteins	Rv1267c	*embR*	Probable transcriptional regulatory protein EmbR
virulence, detoxification, adaptation	Rv1478	-	Possible invasion protein
	Rv1566c	-	Possible Inv protein
unkown	MT3042	-	Uncharacterized protein

Table S2. 3: Summary of significantly enriched Gene Ontology (GO) terms and gene list.

Type	Category	Term	Count	*P*-value	Gene list
GO					
	Biological Process				
		response to antibiotic	5	0,04	*Rv0129c, Rv1634, Rv1886, Rv1988 and Rv3793*
	Cellular Component				
		extracellular region	28	0	*Rv0116c, Rv0125, Rv0129c,Rv0179, Rv0192, Rv0203, Rv0436c, Rv0455c, Rv0559c, Rv0676c, Rv0983, Rv1157c, Rv1161, Rv1230c, Rv1435c, Rv1566c, Rv1804c, Rv1886c, Rv1926c, Rv1980c, Rv2301, Rv2379c, Rv3103c, Rv3193c, Rv3310, Rv3616c, Rv3835*
	Molecuar Function				
		transferase activity, transferring acyl groups	4	0,01	*Rv0116c, Rv0129c, Rv0192, Rv1886c*

Table S3. 1: Tuberculist functional categories of active-TB selected ORFs and their DNA sequence derived peptide sequences.

Functional category	Rv no.	Description	Peptide sequence
cell wall and cell processes	Rv0092	cation transporter P-type ATPase A **ctpA** Cation-transporting ATPase	MRAARRHAESLGETAVFVEVDGEPCGVIAVADAVKDSARDAVAALADR
	Rv0203*	hypothetical exported protein. HEME binding according to uniprot	ETNQVMTAVLQQQVGPGSVASLKAHFEANPKVASDLHALSQPLTDLSTRCSLPI SGLQAIGL
	Rv0361	Probable conserved membrane protein	AKETQVIVTAHEAATEVFQTNQAPTTPVARTTPVATIDAPTCTATIGSILANRLLR SEATAVEIC
	Rv0362	Mg2+ transport transmembrane protein mgtE	**PNMTVSQAVASVRERASGLRSDARTT**VQTAG&RTPRA
	Rv0724	Possible protease IV SppA (endopeptidase IV) (signal peptide peptidase). Involved in digestion of the cleaved signal peptides. This activity is necessary to maintain proper secretion of mature proteins across the membrane. Conserved in M. tuberculosis, M. leprae, M. bovis and M. avium paratuberculosis; **predicted to be essential for in vivo survival and pathogenicity (See Ribeiro-Guimaraes and Pessolani, 2007).**	ITGSIGVITGKLVVRDLKDRLGVGSDAVRTNANAD
	Rv1009	Probable resuscitation-promoting factor RpfB. Predicted possible vaccine candidate (See Zvi et al., 2008).	&SNGWGL@VSEGPFCFLFAV@VVECCRGRGYSALTLLTKLHQTQSP**MLRLVV GALLLVLAFAGGYAVAACKTVTLTVDGTAMRVTTMKSRVIDIVEEN**

Functional category	Rv no.	Description	Peptide sequence
	Rv1029	Probable potassium-transporting ATPase a chain KdpA	SGEATMSHFTQMTGLAVQNFVSASAGMCVLAALIRGLARKRASTLGNFWVDL ARTVLRNIGDDIAEDVEIDLSRIDAITRNVPKKTVIRPGEGLNMVLI
	Rv1269c	Conserved probable secreted protein	TPMTTMITLRRRFAVAVAGVATAAATTVTLAPAPANAADVYGAIAYSGNGSW GRS
	Rv1271c*	Conserved hypothetical secreted protein	NTKDEAFIAQMESIGVTFSSPQVATQQAQLVCKKLASGETGTEIAEEVLSQTNLT TKQAAYFVVDATKAYCPQYASQLT
	Rv1435c*	Probable conserved proline, glycine, valine-rich secreted protein	LTIAISVNSPLFARRYFRNQFGSAEPHSRIEFLFDHRLNCQHPMGNMSPAAPGRF QMV
	Rv1733c*	Probable conserved transmembrane protein. Predicted possible vaccine candidate (See Zvi et al., 2008).	QFTAMIATTRDREGATMITFRLRLPCRTILRVFSRNPLVRGTDRLEAVVMLLAVT VSLLTIPFAAAAGTAVHDSRSHVY
	Rv1799*	Probable lipoprotein LppT, has possible signal peptide and appropriately positioned PS00013 Prokaryotic membrane lipoprotein lipid attachment site.	IPPPYATPPTTTNTPYPPTPPDPPTPAHRPPDPRRRWTIGRSAGSLLPARRRRAR HWLGVNEASRISTKFPRNISRRPQWS@HLIGVRSGQYIET&EE&S**MSVKSKNG RLAARVLVALAALFAMIALTGSACLAEGPPLGRNPQGAPAPVGGTVIVA**
	Rv1860	Alanine and proline rich secreted protein Apa (fibronectin attachment protein) (immunogenic protein MPT32) (antigen MPT-32) (45-kDa glycoprotein) (45/47 kDa antigen). Corresponds to spots 1860 identified in short term culture filtrate. Identified in immunodominant fractions of M. tuberculosis H37Rv culture filtrate.	SLVTVAVPATANADPEPAPPVPTTAASPPSTAAAPPAPATPVAPPPPAAANTPN AQPGDPNAA

Functional category	Rv no.	Description	Peptide sequence
	Rv1910c	hypothetical exported protein. phosphatidylethanolamine-binding protein	STTFLSRSSCHREPRESKRHRR@HRPPADRPGSSAHSKADAAASLARWSKPWL LQLRLATMINMESTVAHAFHRFALAILGLALPVALVAYGGNGDSRKAAPLAPKA AALGRSMPETPTGDVLTISSPAFADGAPIPEQYTCK
	Rv1926c	Immunogenic protein Mpt63 (antigen Mpt63/MPB63) (16 kDa immunoprotective extracellular protein). Predicted possible vaccine candidate (See Zvi et al., 2008).	WFSAIPRRSPP#PAASSVRLKVMTSARFEPGNRPNK@WR@GP**MKLTTMIKT AVAVVAMAAIATFAAPVALAAYPITGKLGSELTMTDTVGQ**
	Rv1926c	Immunogenic protein Mpt63	WR@GP**MKLTTMIKTAVAVVAMAAIATFAAPVALAAYPITGKLGSELTMTDT VGQVVLADVDNQLTVGTDLDQGSFVTAGLDADDHR**
	Rv1979c	Possible conserved permease. FUNCTION Unknown; possibly involved in transport of amino acid across the membrane. With B-cell epitopes. http://iai.asm.org/content/70/12/6996.long	LLGSNKIAASDDTVKLAAAIGNATFRTIIVVGALISMFGINVAASFGAPRLWTALA
	Rv1980c	Immunogenic protein Mpt64. mRNA identified by microarray: downregulated during starvation (see Betts et al., 2002), and possibly down-regulated by hrcA|Rv2374c (see Stewart et al., 2002).	LWQADTDPLPVVFPIVQGELSKQTGQQVSIAPNAGLDPVNYQNFAVTNDGVIF FFNPGELLPEAAGPTQVLVPRSAIDSMLA

Functional category	Rv no.	Description	Peptide sequence
	Rv1980c	immunogenic protein MPB64/MPT64	DRSASLSLVRHRRQQRDALCLSSTQISRQSNLPPAAGGAANYSRRNFDVRIKIFM LVTAVVLLCCSGVATAAPKTYCEELKGTDTGQACQIQMSDPAYNINISLPSYYPD QKSLENYIAQTRDKFLSA
	Rv2152c	Probable UDP-N-acetylmuramate-alanine ligase MurC. Involved in cell wall formation; peptidoglycan biosynthesis.	IATV**AGVSGASVAEHVTVPMRYVPDFSAVAQQVAAAASP**
	Rv2339	Probable conserved transmembrane transport protein MmpL9. Unknown. Thought to be involved in fatty acid transport.	ILLCWLGFTVFVSVAVPPLEAIGETRAVAVAPDDAQSMRAMRRAGKVFNEFDS NSIAMVVLESDQPLGEKAHRYYDHLVDTLVLDQ
	Rv2673	Possible conserved integral membrane protein. Possible arabinofuranosyltransferase AftC; Involved in the biosynthesis of the mycobacterial cell wall arabinan	MRWGRALEYLKITYGWSLLLIVTFTVLYFRYLGSQQWPFPRSLMVIASQFDGDY
	Rv2875	Major secreted immunogenic protein Mpt70. Generally found as a monomer; homodimer in culture fluids.	RRNHVGGSETAQAAVSGSEGVNGMKVKNTIAATSFAAAGLAALAVAVSPPAA MAVMGASALPAAPP
	Rv2878c	soluble secreted antigen mpt53 precursor	R#KLARKLY#QFSRTILNAEIKRTMHAPERRGRRDVLNRHGAYYVIRSKHEQVRS V**MSLRLVSPIKAFADGIVAVAIAVVLMFGLANTPRAVAADERLQFTATTLSGA PFDGASLQGKPAVLWFWTPWCPFCNAEAPSLSQVAAANPA**

Functional category	Rv no.	Description	Peptide sequence
	Rv2921c	Probable cell division protein FtsY	DELDKVKRVVTRRASVDEVLLVLDATIGQNGLAQARVFAEVVDISGAV
	Rv2942	Conserved transmembrane transport protein MmpL7. Involved in translocation of phthiocerol dimycocerosate (dim) in the cell wall.	RATITSGSADGQRRSPRLTNLLVVAAWVAAAVIANLLLTFTQAEPHDTSPALLPQ DANRVSPSFTSSH
	Rv3036c	Probable conserved secreted protein TB22.2 **highly similar to** secreted immunogenic protein **MPT64**/MPB64. **spot 3036c identified in short term culture filtrate.** With bit of Rv3553	TRFRGRAAQCRQV@TDVADDLAVDDPRRPDHAPRQGIDLVTGADGGKHPDA APTPLVCEDD@T**MRYLIATAVLVAVVLVGWPAAGAPPSCAGLGGTVQAGQI CHVHASGPKYMLDMTFPVD**
	Rv3264c	D-alpha-D-mannose-1-phosphate guanylyltransferase ManB	IWPAGRDGTEL**ATHQVDAVVLVGGKGTRLRPLTLSAPKPMLPTAGLPFLTHLL SRIAAAGIEI**
	Rv3330	Probable penicillin-binding protein DacB1	**NALGGCGIITAPGSAPAPGDVSAEAWLVADLDSGAVIAARDPHGRHRPASVI** LNELAPVG IC YFDAVL QLLPTG N
	Rv3395A	Probable membrane protein	TQQVSVRTNTKSTQNDT#**RPACRPAELFATDNTTDGFELPAVATIALTGTVVTT**
	Rv3810*	exported repetitive protein precursor pirG (cell surface protein). mRNA identified by microarray analysis and up-regulated after 96h of starvation (see Betts et al., 2002).	PTDPITVPVA
conserved hypotheticals	Rv0455c	Conserved protein, Function unknown	LPASRRPKSATVTTMSRLSSILRAGAAFLVLGIAAATFPQSAAADSTEDFPIPRRM IATTCDAEQYLAAVRDTSPVYYQRYMIDFNNHANLQ

Functional category	Rv no.	Description	Peptide sequence
	Rv0740*	Conserved hypothetical protein. This region is a possible MT-complex-specific genomic island (See Becq et al., 2007).	RKNLHVWAIVGNCYIEIMPMGTRVELSKLADVALDIGRSVGCSAYENDFTLPDIP TQ
	Rv1157c	Conserved ala-, pro-rich protein, H37Rv-infected guinea pig lungs at 90 days but not 30 days (See Kruh et al., 2010).	NTTVASTWSVCKGLAAVVITSAAAFALCPNAAAEPATPQ
	Rv1268c	Hypothetical protein	KGPTMTTSKIATAFKTATFALAAGAVALGLASPADAAAGTMYGDPAAAAK
	Rv3752c	Identified by mass spectrometry in M. tuberculosis H37Rv-infected guinea pig lungs at 90 days but not 30 days (See Kruh et al., 2010).	HTWCPPLRRRGWPIPVQRHRRPARRWRS&RRLAAGVLGDGWRLEGTTLAVT VEPCTMC
	Rv1357c*	Conserved hypothetical protein	RDHSTDPQPFVSVNVSASTICDPGFLVLVEGVLGETGLPAHALQLELAEDARLS
	Rv1813c	Conserved hypothetical protein	IYLSHTWAHDTQAPESRTEPSTVGGQPPYSDDLIDHRPGRRASNRGRLSTQPH& PFDASTARKSAQPFQTAGLPTAERGQD**MITNLRRRTAMAAAGLGAALGLGILL VPTVDAHLANGSMSE**
	Rv2414c	Conserved hypothetical protein. Identified by mass spectrometry in M. tuberculosis H37Rv-infected guinea pig lungs at 30 days but not 90 days (See Kruh et al., 2010).	STSPTVTALTSREFRAAGLTHLTAVSGANVTIVCAAALVSARLIGPR
	Rv2426c	Conserved hypothetical protein	WNHAKQILRIQAGSGDWEATKTDVFSEEFLLQRPLLTAIRRTEPTVLLIDETDKA DIEIEGLLLEVLSDFAVTVPELGTLTATRAPFVLLTSNATRELSEALW

Functional category	Rv no.	Description	Peptide sequence
	Rv3038c	conserved hypothetical protein	TEEFTAAMLGWPLRTFECTVPPGRLGWGWARFAFTSWKTLGWVDSGVETGS VIGGQFDSMLAKLIVHGADRAEALARARRALNEFG
	Rv3054c	Conserved hypothetical protein	QMVPVWYAPAPK FAPPARTTPQGRYIPEVWSSTAAWPNPAMVRSGAPMM**AASDAHAALVVTPE YNGSIPAVIKNAIDWLSRPFGDGALKDKPLAVIGGSMGRYGGVWAHDETR**
	Rv3188	Conserved hypothetical protein	SPPCGRGASEGPSAAPLRAPPLSSWSISGSVYGSRTGSCGGTT**MAVTLDRAVEA SEIVDALKPFGVTQVDVAAVIQVSDRAVRGWRTGDIRPERYDRLAQLRDLVLL LSDSLTPRGVGQWLHAKNRLLDGQRPVDL**
	Rv3218	Conserved protein	HPGPLPSSSTGRPGPHRHTA@AHCRPPQRADPARIVTRPKPIAVLLNGPTRPAS SHGLSSCV**MRAVLIVNPTATATTPAGRDLLAHALESRLQLTVEHTNHRGHGTE LGQAAVADGVDLVVVHGGDGTVSAVVNGMLGRPGTTPVRPVPAVAVVPG GSANVLARALGI**
	Rv3256c	conserved hypothetical protein	AETAGTILASTLGAGA
	Rv3863	hypothetical alanine rich protein	**DDSRISQTHLRAVSDDGRWRIVGNIPRGMFVGGRRGSSVTVSDKTLIRFGDP PGGKALTFEVVRPSDSAAQHGRVQPSADLSDDPAHNAAPVAPDPGVVRAG AAAAARRRELDISQRSLAADGIINA**VP@PASRSSRRSGVVRMRSSVPEALSRC MVI
	Rv3900c*	Conserved hypothetical alanine rich protein. Identified in the cell membrane fraction of M. tuberculosis H37Rv using 2DLC/MS (See Mawuenyega et al., 2005).	RGGSFGSVEGAGDLASPQSVGAGGFSGSGVQAACSQPAPRAIGASSRHASAGP VRPAPVVTTPAAATPPVIATGPRWRCPAGRCRRRPSDRAYRLRRLGNRLRPGW &RGNR

Functional category	Rv no.	Description	Peptide sequence
	Rv3909	Conserved protein	AQRSRAGAVTALQLGWAALARVTSAIGVVAGLGMALTVPSAAPHALAGEPS TPFVQ
information pathways	Rv0630c	**exonuclease V beta chain recB** Involved in homologous recombination.	APIPAIYTGDTDVFASQAAKDWLCLLEAFDAPQRSGLVRAAACTMFFGETAESL AAE
	Rv2836c	Possible DNA-damage-inducible protein F DinF. Function unknown; induction by DNA damage. drug transmembrane transporter activity.	VRPLRYVVAGFGSSALLCPL
	Rv3646c	DNA topoisomerase I TopA	GHQEVVFSATGRTLTFPGFLKAYVETVDELVGGEADDAERRLPHLTPGQRLDIVE LTPDGHATNPPARYTEASLVKALEELGIGRPSTYSSIIKTIQDR
	Rv3834c	SERYL-tRNA synthetase SerS (serine--tRNA ligase) (SERRS) (serine translase)	SRGEDPALVDALLTADAARRAVISTADSLRAEQKA
intermediary metabolism and respiration	Rv2277c	Possible glycerolphosphodiesterase. Start of protein uncertain, encoded by neighbouring IS6110 as given, is intact in Mycobacterium tuberculosis CDC1551. Required for growth in C57BL/6J mouse spleen, by transposon site hybridization (TraSH) in H37Rv (See Sassetti and Rubin, 2003). Found to be deleted (partially or completely) in one or more clinical isolates (See Tsolaki et al., 2004).	SSWLALSGTVRYCAPRIAHVSTVGKA&RGPMVKTAMLGAVALVIALGGTCGVA DALPLGQTDDPMIVAHRAGTRDFPENTVLAITNAVAAGVDGMWLTVQVSSDG V
	Rv0013	Possible anthranilate synthase component II TrpG (glutamine amidotransferase). MEMBRANE	CAAAHTPLLGVCLGHQAIGVAFGATVD

Functional category	Rv no.	Description	Peptide sequence
	Rv0327c*	**Cytochromes P450** are a group of heme-thiolate monooxygenases. They oxidize a variety of structurally unrelated compounds, including steroids, fatty acids, and xenobiotics.	QLHPIVISIRLHQTFDYQKGAAMASTLTTGLPPGPRLPRYLQSVLYLRFREWFLPA MHRKYGDVFSLRVPPYADNLVVYTRPEHIKEIFAADPRSLHAGEGNHPVVFDPQ P#TVVRRRPSPNVN&GRRIGRNLALSSRSAPFGLTQS
	Rv0509	Probable glutamyl-tRNA reductase HemA (GLUTR)	VICDLGMPRDVDPAVARLPCVWVVDVDSVQHEPSAHAAAADVEAARHIVAAE VASYLVV
	Rv0509	glutamyl-tRNA reductase hemA	RRIRESGVPAEALALDRTGV@RIRAEADL&GRPVGWWLWFR&F&L&KDGKR# #GGYDRKCR&KRATV&R#RQT&FCRY&LRCCYRWFHWY FR
	Rv1130	conserved hypothetical protein. Possible methylcitrate dehydratase PrpD. Predicted possible vaccine candidate (See Zvi et al., 2008). Involved in the catabolism of short chain fatty acids (SCFA) via the tricarboxylic acid (TCA)(acetyl degradation route) and via the 2-methylcitrate cycle I (propionate degradation route).	GKVGIEAVDRAMRGE
	Rv1133c	5-methyltetrahydropteroyltriglutamate-homocysteine methyltransferase metE.	RDGHDAVADEIASSRAAIASRKRDPRLHNGQIRARIEAIVASGAHRGNAAQRRA SQDARLHLPPLPTTTIGSYPQTSALTVTQ
	Rv1393c	Probable monoxygenase. Function unknown; probably involved in cellular metabolism.	QAFVEATFPIAAHYFAVFPLAKHMESAGRRYLRQQVHDP

Functional category	Rv no.	Description	Peptide sequence
	Rv1526c	Probable glycosyltransferase. Unknown; thought to be involved in cellular metabolism. Found to be deleted (partially or completely) in one or more clinical isolates (See Tsolaki et al., 2004)	AVPPNLIEFVESAGLTGVA
	Rv1703c†	Probable catechol-O-methyltransferase. Mycobacteria with Segniliparus rugosus ATCC BAA-974	SPLALELGTYLGYGALRIARAAPEARVYSVELAE
	Rv1876	Probable bacterioferritin BfrA. Involved in iron storage (may perform analogous functions in iron detoxification and storage as that of animal ferritins); ferritin is an intracellular molecule that stores iron in a soluble, nontoxic, readily available form. The functional molecule, which is composed of 24 chains, is roughly spherical and contains a central cavity in which the polymeric ferric iron core is deposited.	MQDNWGFTELAAHTRAESFDEMRHAEEITDRILLLDGLPNYQRIGSLRIGQTLD GAREA
	Rv2070c	Precorrin-6X reductase CobK. Involved in cobalamin biosynthesis.	AARCPLCPGPIGPVRIGGFGGVEGLRGWLREERIDAVVDATHPF
	Rv2332	Probable [NAD] dependent malate oxidoreductase Mez	LPXVQNLRAISTTVAEAVYRAAXQXGVASRTHD
	Rv2499c	Possible oxidase regulatory-related protein	HAGDRESDDAVSACRVAGSTVGRRILQ

Functional category	Rv no.	Description	Peptide sequence
	Rv2537c	3-dehydroquinate dehydratase AroD (AROQ) (3-dehydroquinase) (type II dhqase)	VRLVQGKAGSQTELGLKAVVRQSDSEAQLLDWIHQAADAAEPVILNAGGLT
	Rv2861c	Methionine aminopeptidase MapB (map) (peptidase M). predicted to be essential for in vivo survival and pathogenicity (See Ribeiro-Guimaraes and Pessolani, 2007). Essential gene for in vitro growth of H37Rv, by sequencing of Himar1-based transposon mutagenesis (See Griffin et al., 2011)	WTAQFEHTLLVTDTGVEILTCL
	Rv3157	Probable NADH dehydrogenase I (chain M) NUOK. Involved in aerobic|anaerobic respiration [catalytic activity: NADH + ubiquinone = NAD(+) + ubiquinol]. **mRNA identified by** microarray analysis and down-regulated after 4h, 24h and 96h of starvation	AGAVLIILLPPGRRRLAKWAGMVVSVLTLAVSIVVAAEFKPSAEPYQFVEKHSWI PAFGAGYTLGVDGIAVVLVLLTTVLIPLLLVAGWNDATDA
	Rv3318	Probable succinate dehydrogenase (flavoprotein subunit) SdhA. Identified by mass spectrometry in M. tuberculosis H37Rv-infected guinea pig lungs at 30 and 90 days (See Kruh et al., 2010).	AYELATGDIHVFHAKAVVIATGGSGRMYKTTSNAHTLTGDGIGIVFRKGLPLED MEFHQFHPTGLAGLGILISEAVRGEGGRLLNGEGERFMERYAPTIVDLAPRDIVA RSMVLEVLEG
	Rv3423c	**Alanine racemase Al**r. Provides the D-alanine required for cell wall	VRPGIAVYGLSPVPALGDMGLVPAMTVKCAVCAQGW*SRRRG-GVSF

Functional category	Rv no.	Description	Peptide sequence
		biosynthesis. Transforms L-alanine to D-alanine	
	Rv3762c	Possible hydrolase. Function unknown; probably involved in cellular metabolism.	YFPRFRALCMAENATHNLHNLLTLRGALVRDPRAWSGYLTEAIDTFADRTDVVF ASHHWPTWGREKIVEFLSQQRDMYSYLHDQTLRLLNQGYTGVEIAEMFQLPPA LQRAWHTH
lipid metabolism	Rv1886c	esterase, putative, antigen 85-B (fbpB)	RPFTFTHAHLRLPDTQSAGLLYPHRTCRPPTGNQAHTRGTHHERGSDTEMRRL CARIDIWP**PHTVCSGPSTRRHTGQRGTGMTDVSRKIRAWGRRLMIGTAAAV VLPGLVGLAGGAATAGA**FSRPGLPVEYLQV
	Rv1886c†	Secreted fibronectin-binding protein antigen 85-B fbpB . Involved in cell wall mycoloylation. Proteins of the antigen 85 complex are responsible for the high affinity of mycobacteria to fibronectin. Mycobacteria only maize Colletotrichum graminicola	PHTVCSGPSTRRHTGQRGTGMTDVSRKIRAWGRRLMIGTAAAVVLPGLVGLA GGAATAGA
	Rv2048c	Polyketide synthase Pks12. Involved in biosynthesis of mannosyl-beta-1-phosphomycoketide (MPM). Required for growth in C57BL/6J mouse spleen, by transposon site hybridization (TraSH) in H37Rv (See Sassetti and Rubin, 2003).	AHRLHGLPEAEQHAVLLGLVRLHIATVLGNITPEAID
	Rv2252	Diacylglycerol kinase. Involved in synthesis of phosphatidylinositol mannosides (PIMS).	ICPNADHSDGLLDITMAQSDSRTKLLRLFPTIFKGAHVELDEVSTTRAKT

Functional category	Rv no.	Description	Peptide sequence
	Rv2941	Fatty-acid-AMP ligase FadD28 (fatty-acid-AMP synthetase) (fatty-acid-AMP synthase). Involved in phthiocerol dimycocerosate (dim) biosynthesis. Thought to be involved in the release and transfer of mycoserosic acid from mas onto the DIOLS.	ISYMLPRSPIVRIVDSDTCIECPDGTVGEIWVHGDNVANGYWQKPDESERTFGG KIVTPSPGTPEGPWLRTGDSGFVTDGK
	Rv3720	Possible fatty acid synthase. Function unknown, but involved in lipid metabolism.	LANVVRSIGVEHILPIAP
	Rv3803c	Secreted MPT51/MPB51 antigen protein FbpD (MPT51/MPB51 antigen 85 complex C) (AG58C) (mycolyl transferase 85C) (fibronectin-binding protein C) (85C). **May have a role in host tissue attachment, whereby ligands may include the serum protein fibronectin and small sugars.**	MKGRSALLRALWIAALSFGLGGVAVAAEPTAKAAPYENLMVPSPSMGRDIPVA FLAGGPHA
PE/PPE	Rv0978c*	PE-PGRS family protein PE_PGRS17	MSFVNVAPQLVSTAAADAARIGSAINTANTAAAATTQVLAAAQDEVSTAIAALF GSHGQH
	Rv1087	PE-PGRS family protein PE_PGRS21. Identified by mass spectrometry in M. tuberculosis H37Rv-infected guinea pig lungs at 90 days but not 30 days (See Kruh et al., 2010).	TPGGQVGDGDGGAGGAGGNGGASGAGGWLLGTGGAGGAGGNGGNG

Functional category	Rv no.	Description	Peptide sequence
	Rv1450c	PE-PGRS family protein PE_PGRS27	MSLVIVAPETVAAAALDVARIGSSIGAANAAAAGSTTSVLAAGADEVSAAIATLF GSHAREYQAISTQVAAFHDRFAQTLSAAVGSYVSAEATNAAPLATLEHNVLNAL NAPTQALLGRP
	Rv1468c*	PE-PGRS family protein PE_PGRS29. Identified by mass spectrometry in M. tuberculosis H37Rv-infected guinea pig lungs at 30 and 90 days (See Kruh et al., 2010).	DNGGVGALGANGGAGGTGGWLFGNGGAGGNSGGGGGAGGIGGSAVLFGA GGA
	Rv1768*	PE-PGRS family protein **PE_PGRS31** Identified by mass spectrometry in M. tuberculosis H37Rv-infected guinea pig lungs at 30 and 90 days (See Kruh et al., 2010). **Signal peptide**	HSYAVAEAATAQSVQQDLLNLINAPTQALLGRPLIGNGANGLPGT
	Rv1806*	PE family protein PE20. This region is a possible MT-complex-specific genomic island (See Becq et al., 2007).	AFVLVCPDALAIAAGQLRHVGSVIAARNAVAAPATAELAPAAADEVSALTATQF NFHAAMYQAVGAQAIAMNEAFVAMLG
	Rv1818c	PE-PGRS family protein PE_PGRS33. Function unknown. Seems to influence both cell surface interactions among mycobacteria and the interactions of bacteria with macrophages. host cell surface receptor binding (UniProt info).	NGAPGTGANGGDGGILIGNGGAGGSGAAGMPGGNGGAAGLFG
	Rv2741	PE-PGRS family protein PE_PGRS47. Predicted to be an outer membrane protein (See Song et al., 2008). This region is a possible MT-complex-specific genomic island (See Becq et	GDGGESDNGDGGNGGVGGKAGLVGEGGNGGDG

Functional category	Rv no.	Description	Peptide sequence
		al., 2007). Identified by mass spectrometry in M. tuberculosis H37Rv-infected guinea pig lungs at 90 days but not 30 days (See Kruh et al., 2010).	
	Rv3343c*	PPE family protein PPE54	STGLFNAGGFNTGVVNAGSYNTGSFNAGQANTGGFNPGSVNTGWLNTGDTN
	Rv3347c**	PPE family protein PPE55. function unknown. NO BCG Predicted possible vaccine candidate (See Zvi et al., 2008). Identified by mass spectrometry in M. tuberculosis H37Rv-infected guinea pig lungs at 30 and 90 days (See Kruh et al., 2010).	ALTQSPIINLGLADVGNYN
	Rv3508*	PE-PGRS family protein PE_PGRS54. Function unknown.	NGGAGGAGGTPTGSGTEGTGGDGGDAGAGGNGGSATGVGNGGNGGDGGN GGDGGN
regulatory proteins	Rv0890c	Probable transcriptional regulatory protein (probably LuxR-family). Identified by mass spectrometry in M. tuberculosis H37Rv-infected guinea pig lungs at 90 days but not 30 days (See Kruh et al., 2010).	RATTLQVFLRDQLLLEFSRPGGDAIAARQFANDAVDATNGWHRMVALTIRARV ATARGEPELARDDAHAALACGGGMH
	Rv3058c*	Possible transcriptional regulatory protein (probably TetR-family)	PFAELSVRAISLRAGVARSGFYFYFDSKY
	Rv3849	ESX-1 transcriptional regulatory protein EspR. Involved in transcriptional mechanism. Regulates	AARLNRLFDTVYPPGRGPHTSAEVIAA

Functional category	Rv no.	Description	Peptide sequence
		transcription of genes required for ESX-1 system.	
unknown	MT0066.2*	Hypothetical protein not yet annoated in H37Rv as recently recorded	LMRVRWICAGNRGRKHKRRCTTQYRSTQASKLQLHFKLRQTLNRLGGLQAMV SACG
	MT1330.1*	hypothetical protein	**ASSVELALAVPAKPSAARPKVAAVKVVPTMEAAKRRVNIDDLLVLMSSEVAL VCVPSESAPGLDRFS**INPWQYRYRFDGVPTVQGERSAMVVPQAQSGAVAVD GVSPGQKNGSHH@E
	Rv2348c**	Hypothetical protein. All info from Tuberculist. mRNA identified by microarray analysis and up-regulated after 4h of starvation (see Betts et al., 2002).	DAVVAKRAESGMLGGLSVPLSWGVAVPPDDYDHWAPAPEDGADVDVQAAE GAD
virulence, detoxification, adaptation	Rv1563c	Maltooligosyltrehalose synthase TreY. Involved in trehalose biosynthesis (protective effect).	WRHGLCGYRRFFSITSLAGLRQEDRAVFDASHAEVARWFTEGLVDGVRVDHLD GLSDPSGYLAQLRELLGPNAWIVVEKILAVDEALEPTLPVDGSTGYDVLRE
	Rv2429	Alkyl hydroperoxide reductase D protein AhpD (alkyl hydroperoxidase D). Involved in oxidative stress response.	HCLVAHEHTLRTVGVDREAIFEALKAAAIVS
	Rv2231A†	toxin VapC16. function unknown. *MTB*C with only Frankia sp. CN3	LTMACTACPTIWTLRCQTTCSNAFTGEALPHRHPRLAADAVNETRAIVQDVRN SILLSAASAWEIAINYRLGKLPPPEPSASYVPE
	Rv3922c	Possible hemolysin function unknown. Putative membrane protein insertion efficiency factor	RVTGRASXRGLIFXIXXYRHMLSPLRPASCRFVPTCSQYAVDALTEYGLLRGSWLT MIXLAKCGPWHRGGWDPIPEGLTTGRSCXTDVDGANDDWNPASK

Table S3. 2: Tuberculist functional categories of latent-TB selected ORFs and their DNA sequence derived peptide sequences.

Functional category	**Rv no.**	**Description**	**Peptide sequence**
cell wall and cell processes	Rv0012†	Probable conserved membrane protein. Function unknown	PTPCPENGETMIDRRRSAWRFSVPLVCLLAGLLLAATHGV
	Rv0050	Probable bifunctional penicillin-binding protein 1A/1B PonA1 (murein polymerase) (PBP1): penicillin-insensitive transglycosylase (peptidoglycan TGASE) + penicillin-sensitive transpeptidase (DD-transpeptidase). Involved in peptidoglycan synthesis (at the final stages), cell wall formation. Synthesis of cross-linked peptidoglycan from the lipid intermediates.	KATMDGALKGTSNETFPKPTEVGGYAGVPPPPPPSEVPPSETVIQPTVEIAK
	Rv0179c	Possible lipoprotein LprO, function unknown. Prokaryotic membrane lipoprotein lipid attachment site.	SNAAEPSRSCY&S&QCDQCA#CWDGC**MWIRAERVAVLTPTASLRRLTAC YAALAVCAALACTTGQPAARAADG**
	Rv0203*	Possible exported protein, function unknown. HEME binding	GGGNSQAAGPSQTRRKPVANRKESSMKTGTATTRRR**LLAVLIALALPGAA VALLAEP**
	Rv0318c*	Probable conserved integral membrane protein	**ASLVTFPVLAGILG***GVVPSVRTPSAAM***VSGVQHFAAGIVMAAVAXEVL**P DLRSRGP
	Rv0412c	Possible conserved membrane protein	VTVELAHPSTEPLGSRSPAEPAHPRRWFISTTPGR**IMTIGIVLAALGVASAF ATSTTIEH**
	Rv0412c	Possible conserved membrane protein. Function unknown	ISTTPGR**IMTIGIVLAALGVA**TAVPSTAISART
	Rv0507	Probable conserved transmembrane transport protein MmpL2. Function Unknown. Thought to be involved in fatty acid transport.	TPPPHGVKAYVTGPAALNADQAEAGDKSIAKVTAITSMVIAAMLLVIYRSVI TAVLV

Functional category	Rv no.	Description	Peptide sequence
	Rv0559c	Possible conserved secreted protein, function unknown. A core mycobacterial gene; conserved in mycobacterial strains (See Marmiesse et al., 2004).	KGTKL**AVVVGMTVAAVSLAAP**AQADDYDAPFNNTIHRFGIYGPQDYNAW
	Rv0680c	Probable conserved transmembrane protein	VKWNTV**AASLAAGVITIAVALAA**
	Rv1009	Probable resuscitation-promoting factor RpfB. Thought to promote the resuscitation and growth of dormant, nongrowing cells. Has little to no effect on actively-growing cells.	SNGWGL@VSEGPFCFLFAV@VVECCRGRGYSALTLLTKLHQTQSP**MLRLVVGALLLVLAFAGGYAVAACKTVTLTVDGTAMRVTTMKSRVIDIVEENGFSVDDRDDLYPAAGVQVHDADTIVLRRSRPL**
	Rv1029	Probable potassium-transporting ATPase a chain KdpA. One of the components of the high-affinity ATP-driven potassium transport (or KDP) system.	YSGEATMSHFTQMTGLAVQNFVSASAGMCVLAALIRGLARKRASTLGNFWVDLARTVLRNIGDDIAEDVEIDLSRIDAITRNVPKKTVIRPGEGLNMVLI
	Rv1183	Probable conserved transmembrane transport protein MmpL10. Function Unknown. Thought to be involved in fatty acid transport.	**VVGCWVALALVLPMAVPSLAEMAQRHPVAVLPADAPSSVAVRQMAEAFHESGSENILVVL**
	Rv1184c	Possible exported protein. Function unknown. A core mycobacterial gene; conserved in mycobacterial strains (See Marmiesse et al., 2004). Predicted to be an outer membrane protein (See Song et al., 2008). This region is a possible MT-complex-specific genomic island (See Becq et al., 2007). PE-PPE domain-containing protein	YARKPPNDSCFPMLSRTRFSMQRQMKRV**IAGAFAVWLVGWAGGFGTAIAAS*EPAYPWAPGPP***
	Rv1269c	Conserved probable secreted protein. Function unknown	*TLRRRFAVAVAGVATAAATTVTLAPAPANA***ADVYGAIAYSGNGSWGRSWDYPTRAAAEATAVKSCGYSDCKVLTSFTACGAVAAN RAYQ**
	Rv1271c*	Conserved hypothetical secreted protein	FTTAVGAAAIGLAVATAGTAGANTKDEAFIAQME

Functional category	Rv no.	Description	Peptide sequence
	Rv1273c	Probable drugs-transport transmembrane ATP-binding protein ABC transporter. Thought to be involved in active transport of drugs across the membrane (export): multidrugs resistance by an export mechanism.	QILMAVLMATMTLAVLPRASVCAERITEVLSTPADPVSSDDIRGRTAGALHA
	Rv1291c	Conserved hypothetical secreted protein. Function unknown	LRPKNSRSMAKKAASSMRSR@RSSSSNSKQSSSAGSPSTKNLSRPGADSEGTQTSATSDDISTRRSSMFTRRFA***ASMVGTTLTAATLGLAALGFAG*****TASASSTDEAFLAQLQADGITPPSAARAIKDAHAVCDALDEGHSAKAVIKAVAK**
	Rv1424c*	Possible membrane protein. This region is a possible MT-complex-specific genomic island (See Becq et al., 2007).	SDPRPPTETPTRPKTGHSPRCPTSSRRPPAEKLVFPVLLGILTLLLSACQTASASGYNEPRGYDRATLKLV
	Rv1635c	Probable mannosyltransferase. Probable conserved transmembrane protein. Possibly involved in the biosynthesis of lipoarabinomannan (lam). A core mycobacterial gene; conserved in mycobacterial strains (See Marmiesse et al., 2004).	RADPKVTRIMSASTLEQPAAARTGPGASGRWGNQNRRAGP
	Rv1639c	Conserved hypothetical membrane protein. Function unknown	DLTVMHPTLFSAFVDIAGDFYPNAGNKTQTIVRLFGGNEDAWSAFDPTTVITRHGSYTGLSGWFAISSPGPPSP
	Rv1779c	Possible integral membrane protein. Function unknown.	SLRIRLGSRDVICVISRAVVNNTTRGRFDSVCRAGACQA&PAGGVFSL**MCAHEYAEQRSAVSGIEGLLTWLG**
	Rv1863c	Probable conserved integral membrane protein. abortive infection protein. the abortive infection (Abi) systems provide protection by the abortion of an existing phage infection. Typically, these Abi systems target a crucial step of phage multiplication such as replication, transcription or	RWWGFRGVAVAGSVLFGLWHIATSLGLTSSNV

Functional category	Rv no.	Description	Peptide sequence
		translation, and lead to the death of the infected cell.	
	Rv1884c	Probable resuscitation-promoting factor RpfC. Thought to promote the resuscitation and growth of dormant, nongrowing cell. Factor that stimulates resuscitation of dormant cells. Has peptidoglycan (PG) hydrolytic activity. Stimulates growth of stationary phase M.bovis (a slow-growing Mycobacterium). Has little to no effect on actively-growing cells. Predicted possible vaccine candidate (See Zvi et al., 2008).	TRIAKPLIKSAMAAGLVTASMSLSTAVAHAGPSPNWDAGAQ
	Rv1926c	Immunogenic protein Mpt63 (antigen Mpt63/MPB63) (16 kDa immunoprotective extracellular protein). Predicted possible vaccine candidate (See Zvi et al., 2008).	LRLKVMTSARFEPGNRPNK@WR@GP**MKLTTMIKTAVAVVAMAAIATFAAPVALAAYPITGKLGSE**
	Rv1980c	Immunogenic protein Mpt64 (antigen Mpt64/MPB64)	LSLVRHRRQQRDALCLSSTQISRQSNLPPAAGGAANYSRRNFDVRIKIFMLVTAVVLLCCSGVATAAPKTYCEELKGTDTGQACQIQMSDPAYN
	Rv1984c	Probable cutinase precursor CFP21. Hydrolyzes cutin. Shown to have esterase and lipase activity.	PRSLVRIVGVVVATTLALVSAPAGGRAAHADPCSDIAVVFARGTHQA
	Rv2152c	Probable UDP-N-acetylmuramate-alanine ligase MurC. Involved in cell wall formation; peptidoglycan biosynthesis. CYTOPLASM?	IATV**AGVSGASVAEHVTVPMRYVPDFSAVAQQVAAAASP**

Functional category	Rv no.	Description	Peptide sequence
	Rv2180c	Probable conserved integral membrane protein. NB: Rv1129c part of C82	
	Rv2181*	Alpha(1->2)mannosyltransferase Mannosyltansferase responsible for the addition of alpha(1->2) branches to the mannan core in the biosynthesis of lipomannan (LM) and lipoarabinomannan (lam)	QNIAGALARLTIGDDERFALWVAGSLLVLAATIWAMRRVLRAGEPTLAVIC VALFGLVVSPVSWSHH
	Rv2376c	Low molecular weight antigen CFP2 (low molecular weight protein antigen 2) (CFP-2). Function not known (putative secreted protein); may play a role in the development of protective immune responses. A core mycobacterial gene; conserved in mycobacterial strains (See Marmiesse et al., 2004).	VKSIAAGLTAAAAIGAAAAGVTSIMAGGPVVYQMQPVVFGAPLPLDP
	Rv2434c	Probable conserved transmembrane protein	**QIGREHLEQVVMNKPMLLQELGRVIDERQRKAQQAIRRDLHQSPAAAG EHRGPARR**&RAVGHGWPSDRSVSAR
	Rv2721c	Possible conserved transmembrane alanine and glycine rich protein. Function unknown	STLIGRTLLGLAATAVTAVLLAPTVAASPMGDAEDAMMAAWEKAGGDTS TLGVRKGD
	Rv3265c	dTDP-RHA:a-D-GlcNAc-diphosphoryl polyprenol, a-3-L-rhamnosyl transferase WbbL1 (alpha-L-rhamnose-(1->3)-alpha-D-GlcNAc(1->P)-P-decaprenyl). Probably involved in cell wall arabinogalactan linker formation: Uses dTDP-L-rhamnose as substrate to insert the rhamnosyl residue into the cell wall. Seems to be essential for mycobacterial viability.	AVNRTIAQLGEMAGDAGEPWVDDWVIVA

Functional category	Rv no.	Description	Peptide sequence
	Rv3310	Acid phosphatase (acid phosphomonoesterase) (phosphomonoesterase) (glycerophosphatase) sapM. Involved in cellular metabolism: acting on ester bonds.	PRPYTPG**MLRGIQALSRPLTRVYRALAVIGVLAASLLASWVGAVPQVGLA ASALPTFAHVVIVVEENRSQAAIIGNKS**
	Rv3390	Probable conserved lipoprotein LpqD	NIGSVTAPMEFGSIAMQLSSRLENHRKPFPRTVSTGQAIA**MAKRTPVRKA CTVLAV**
	Rv3476c	Probable dicarboxylic acid transport integral membrane protein KgtP (dicarboxylate transporter). Involved in active transport of dicarboxylic acid across the membrane. Responsible for the translocation of the substrate across the membrane.	SGSLRELATHYWKPLLLCFLVTLGGTVAFYTYSVNAP
	Rv3476c	Probable dicarboxylic acid transport integral membrane protein KgtP (dicarboxylate transporter). Involved in active transport of dicarboxylic acid across the membrane. Responsible for the translocation of the substrate across the membrane.	WTTACAYDHLIPGRGVGVLLDDGSQVALFRLDDGSVHAVGNVDPFSGAA VMSRGIVGDRGGRAM*VPSRSS***IGVAAPILLILCRLVQGFATGG***EYGTSATY MSEAATRERRGYFS***
	Rv3629c	Probable conserved integral membrane protein, funfction unknown.	RL**IIAAILQRMSPFWQRMFLTIGILIAVFGMRLVFPLAIIWTTAGLDPVRA MELALRPPAHGALEFADGS**
	Rv3682	Probable bifunctional membrane-associated penicillin-binding protein 1A/1B PonA2. Involved in peptidoglycan synthesis (at the final stages), cell wall formation. Synthesis of cross-linked peptidoglycan from the lipid intermediates. **Supposedly involved in stationary-phase survival.**	AGWDLPMSGKTGTTEAHRSAGFVGFTNRYAAANYIYDDSSSPTDLCSGPL RHCGSGDLYGGNEPSRTWFAAMKPIANNFGEVQLPPTDPRYVDGAPGSR VP**SVAGLDVDAARQRLKDAGFQVADQTNSVNSSA**K

Functional category	Rv no.	Description	Peptide sequence
	Rv3682	Probable bifunctional membrane-associated penicillin-binding protein 1A/1B PonA2 (murein polymerase) [includes: penicillin-insensitive transglycosylase (peptidoglycan TGASE) + penicillin-sensitive transpeptidase (DD-transpeptidase)]. Involved in peptidoglycan synthesis (at the final stages), cell wall formation. Supposedly involved in stationary-phase survival.	**SVAGLDVDAARQRLKDAGFQVADQTNSVNSSA**E
	Rv3737	Probable conserved transmembrane protein, function unknown.	GIATLVAVAAYLIAGQDPTALVATGIVVLLSGMTLVGSMQDAVTGYMLTALARLGDALFLTAGIVVGILISLRGVTNAGIQIELHVDATTTLATPGMPLPILVAVSGAALSGVCLTIASYA
	Rv3779	Probable conserved transmembrane protein alanine and leucine rich, function unknown. A core mycobacterial gene; conserved in mycobacterial strains (See Marmiesse et al., 2004).	VQYALIVLAAIGGLILLVKKIWWPLAVWLLLIVMNVDAGTPLGGPIG
	Rv3793	Integral membrane indolylacetylinositol arabinosyltransferase EmbC. Involved in the biosynthesis of the mycobacterial cell wall arabinan and resistance to ethambutol. A core mycobacterial gene; conserved in mycobacterial strains (See Marmiesse et al., 2004).	RIAVRLPSTSVRDAGANYRIARYVAVVAGLL
conserved hypotheticals	Rv0004	Conserved hypothetical protein, Function unknown. Belongs to superfamily DUF721; this family contains several actinomycete proteins of unknown function.	SEDEHWDRVGSGWPRPGRDG**TLDEARAAARARGQDAGRGRVASVASGRVAGRRRS**
	Rv1069c†	Conserved protein, Function unknown.	LFRFLVGQVDRIAPFRVSAAIVVVLLVVLTITLLNGVVLKFAMNSMNSTFAAVNDNAIAKI

Functional category	Rv no.	Description	Peptide sequence
	Rv1157c	Conserved ala-, pro-rich protein, Function unknown. Has hydrophobic stretch at N-terminus suggestive of secretion signal. First start taken. A core mycobacterial gene; conserved in mycobacterial strains (See Marmiesse et al., 2004). structural constituent of cell wall	*NTTVASTWSVCKGLAAVVITSAAAFALCPNAAA***DPATP**
	Rv1157c	Conserved secreted ala-, pro-rich protein, function unknown. A core mycobacterial gene; conserved in mycobacterial strains (See Marmiesse et al., 2004).	*TWSVCKGLAAVVITSAAAFALCPNAAA***DPATPQPNPTQQLPG**WPGSPNR RRWRCTCRSGGLSPSPDQVS#T
	Rv1268c	Hypothetical protein, Predicted to be an outer membrane protein.	RKGPT**MTTSKIATAFKTATFALAAGAVALGLASPADAAAGTMY**
	Rv1352	Conserved protein, function unknown.	LHDREEN**MARTLALRASAGLVAGMAMAAITLAPGARAETGEQFPGDGV FLVGTDIAPGTYRTEGPSNPLILVFGRVSELSTCSWSTRSAPEVSNENIVDT NTSMGPMSVVIPPTVAAFQHYGHRTVVD**
	Rv1357c*	Conserved hypothetical protein. Function unknown.	RDHSTDPQPFVSVNVSASTICDPGFLVLVEGVLGETGLPAHALQLELAEDAR LS
	Rv1804c*	Conserved protein. Function unknown	LYQYCGDVPPVELEAAYYAQRQRPAAG&GLRSESLRTHRGGS**IPGESLSMR VVSTLLSIPLMIGLAVPAHAGPSGDDAVFLASLERAGITYSHPDQAIASGK AVCALVESGESGLQVVNE**
	Rv1810*	conserved hypothetical protein. Identified in immunodominant fractions of M. tuberculosis H37Rv culture filtrate (See Covert et al., 2001). DNA microarrays show higher level of expression in M. tuberculosis H37Rv during Mg2+ starvation (See Walters et al., 2006).	VLTSASGGSPSTAVNVIGR#GLCCTDGRILTRKVCRALAADLPQDAMQLQR TMGQCRP**MRMLVALLLSAATMIGLAAPGKADPT**GDDAAFLAALDQAGI TTLTQATP#V

Functional category	Rv no.	Description	Peptide sequence
	Rv1810*	Conserved protein. Function unknown. Identified in immunodominant fractions of M. tuberculosis H37Rv culture filtrate using 2D-LPE, 2D-PAGE, and LC-MS or LC-MS/MS (See Covert et al., 2001).	**MRMLVALLLSAATMIGLAAPGKADPT**
	Rv2018*	conserved hypothetical protein. Putative antitoxin VapB45. Possibly the antitoxin component of a type II toxin-antitoxin (TA) module. Its cognate toxin is VapC45.	LATWADGYERRPANAPAVQGQPIITALPHPTGSHARLPFVGIAEAYV
	Rv2414c	Conserved hypothetical protein. Identified by mass spectrometry in M. tuberculosis H37Rv-infected guinea pig lungs at 30 days but not 90 days (See Kruh et al., 2010).	TVTALTSREFRAAGLTHLTAVSGANVTIVCAAALVSARLIGP
	Rv2532c	Hypothetical protein. Function unknown	PFRSRPTRFGRPAGQGAVDRRLAGLCQRGAGPGHAGDPAAACGRSSGA WGGV**MTRLELRVVVAAVLAATVVLGAVVCAAYGLTI**
	Rv2955c*	Conserved protein. This region is a possible MT-complex-specific genomic island (See Becq et al., 2007). O-methyltransferase activity	MAPWYWRGLQVTLEPGSAIAWIVRLTGGFEETEIDIAAALYS
	Rv3212	Conserved alanine valine rich protein. Function unknown. A core mycobacterial gene; conserved in mycobacterial strains (See Marmiesse et al., 2004). Predicted to be an outer membrane protein (See Song et al., 2008).	ERRTKTDIAAAATIAVVVAVAASLIWWTSDARATISRPAAV
	Rv3218	Conserved protein. Function unknown	ASPAMPWMHA@PPVISGSAAASGLNAI**AVAVVPGGSANVLARALGISA DPIAATNQLIQLLDDYGRHQQWRRIGLIDCGERWAVF**
	Rv3333c	Hypothetical proline rich protein function unknown	IMFTGIASHAGALGAALVVLIGAAILHDGPAAADPNQDDRFLALLEKKEIPA VANVPRVIDAAHK

Functional category	Rv no.	Description	Peptide sequence
	Rv3354*	Conserved hypothetical protein. Function unknown. A core mycobacterial gene; conserved in mycobacterial strains (See Marmiesse et al., 2004).	NPVDDAFIAALNNAGVNYGDPVDAKALGQSVCPILA
	Rv3365c	conserved hypothetical protein. Function unknown. integral membrane sensor signal transduction histidine kinase	VRWK**VFTIALLPLVVAMVLAGL**RVEAAMASTSGLRLVAARA MIP
	Rv3572	Hypothetical protein, Predicted to be an outer membrane protein.	VAGANLRTPDGPTGFPPGLWARQTTEIRSTNRLAYLDAHATSQFERVMKA GGSDVIT
	Rv3896c	Conserved hypothetical protein. Function unknown.	LRQDFIPSPVGGPINTLVSATLPAEFDALGARSTPSISSGPQTRSNAVTVALP QVNLRCQLVNTCERVASKSTCCVAVIYSLAESRCWD#LGEEDSMSFVTTQP EALAAA**AAIQKALDIKGVHDPAARARWTRGMDLVARRESNYNANAINH WDSNAARGTPS**
information pathways	Rv0058†	Probable replicative DNA helicase DnaB. Participates in initiation and elongation during chromosome replication; it exhibits DNA-dependent ATPase activity. The intein is an endonuclease (potential).	DFYRPAHQNVYDAILDLYGRGEPADAVTVAAELDRRGLLRR
	Rv0429c	Probable polypeptide deformylase Def (PDF) (formylmethionine deformylase). Removes the formyl group from the N-terminal met of newly synthesized proteins	CLSVPGESFPTGRAKWARVTGLDADGSPVSIEGTGLFARMLQHE
	Rv0630c*	Probable exonuclease V (beta chain) RecB. Involved in homologous recombination.	IPAIYTGDTDVFASQAAKDWLCLLEAFDAPQRSGLVRAAACTMFFGETAES LAAEI
	Rv1329c	Probable ATP-dependent helicase DinG. Probable helicase involved in DNA repair and perhaps also replication.	&PRPV**SESVSMSVPELLAIAVAALGGTRRRGQQEMAAAVAHAFETGEHL VVQAGTGTG**

Functional category	Rv no.	Description	Peptide sequence
	Rv1547	Probable DNA polymerase III (alpha chain) DnaE1 (DNA nucleotidyltransferase)DNA polymerase III is a complex, multichain enzyme responsible for most of the replicative synthesis in bacteria.	TTKVMIIVVMFFSHTYSSYGADIVDDAVVLVNAKVAVRD
	Rv2101	Probable helicase HelZ. Has helicase activity.	IRLRGQWVALDTEQLRRGLEFLERKPTGRKTTAEILALI
	Rv2839c	Probable translation initiation factor if-2 InfB. If-2, one of the essential components for the initiation of protein synthesis in vitro, protects formylmethionyl-tRNA from spontaneous hydrolysis and promotes its binding to the 30S ribosomal subunits. It is also involved in the hydrolysis of GTP during the formation of the 70S ribosomal complex.	RQEYDSMQAPVVGGVRLPHGNGET
	Rv3459c	30S ribosomal protein S11 RpsK	TSARKGQKTRRREKKNVPHGAAHIKSTFNNTIVTITDPQGN
	Rv3923c	Ribonuclease P protein component RnpA	ALPSSRHVSSARLEQQLRCGLRRAVELAGSDR&VCLGKEL
insertion seqs and phages	Rv1041c*	Probable is like-2 transposase. Possibly required for the transposition of an insertion element. This region is a possible MT-complex-specific genomic island (See Becq et al., 2007).	LRSKKIKHTIPERQDQIDRRKAKGSAGGRPPAFDAALYGLRNTVERGFHRLK QWRGIATRYDKYALTYLGGVLLACAVIHARVGTPKLGDTP@PRPASVHPG RDSARQTVALSSRSAPIGDYL
	Rv2944†	Possible transposase for insertion sequence element IS1533	SNKAFGRWGEVFGGDDVVAAAMIDRLVHHAE
intermediary metabolism and respiration	Rv0327c*	Possible cytochrome P450 135A1 Cyp135A1. Cytochromes P450 are a group of heme-thiolate monooxygenases. They oxidize a variety of structurally unrelated compounds, including steroids, fatty acids, and xenobiotics. This region is a possible MT-complex-specific genomic island	QLHPIVISIRLHQTFDYQKGAA**MASTLTTGLPPGPRLPRYLQSVLYLRFREW FLPAMHRKYGDVFSLRVPPYADNLVVYTRPEHIKEIFAADPRSLHAGEGN H**PVVFDPQP#TVVRRRPSPNVN&GRRIGRNLALSSRSAPFGLTQS

Functional category	Rv no.	Description	Peptide sequence
		(See Becq et al., 2007).	
	Rv0565c	Probable monooxygenase. Function unknown; probably involved in cellular metabolism.	DVLATLVADPGFQARFT**MSVTPNAGCVDVVIVGAGISGLGAAYRIIERNP QLTYTILERRARIGGTWDLFRYPGVRSDSSIFTLSFPYEPWTREEGIADGAH IREYLTDMAHKYGIDRHIEFNSYVRAADWDSSTDTWTVTFEQNGVHKHY RSRFVFFGSGYYKYDEGYTPDFGGIEKFGGAVVHPQHWPEDLDYTGKKIV VIGSGATAV**
	Rv0688	Putative ferredoxin reductase. Ferredoxins are iron-sulfur proteins that transfer electrons in a wide variety of metabolic reactions.	ASWRDPMGHQARVEHWSNVADQARVVVPAMLGTDVPTGVVVPYFWS DQYDVKIQCLGEPHATDVVHLVEDDGRK
	Rv0769	Probable dehydrogenase/reductase. Function unknown; probably involved in cellular metabolism.	VSAGVSGSGWLR@PWGWC**AAAVAKQIVADGGTAIHVPVDVSDEDSAK AMVDRAVGAFGGIDYLVNNAAIYGGM**
	Rv1133c	Probable 5-methyltetrahydropteroyltriglutamate--homocysteine methyltransferase MetE	TEVIALQERLGLDVLVHGEPERNDMVQYFAEQLAGFFATQNGWVQSYGS RCVRPPIL
	Rv1161	Respiratory nitrate reductase (alpha chain) NarG. Nitrate reduction [catalytic activity: nitrite + acceptor = nitrate + reduced acceptor].	FSPAIDPPWETRSDFDAFAAIARAFSALAKRHLGTRTDVVLTALQHD

Functional category	Rv no.	Description	Peptide sequence
	Rv1223	Probable serine protease HtrA (DEGP protein). Possibly hydrolyzes peptides and/or proteins (seems to cleave preferentially after serine residue). Conserved in M. tuberculosis, M. leprae, M. bovis and M. avium paratuberculosis; predicted to be essential for in vivo survival and pathogenicity (See Ribeiro-Guimaraes and Pessolani, 2007).	LSRSCRT#**PAGRFTKVAAAVADSVVTIESVSDQEGMQGS**
	Rv1703c	Probable catechol-O-methyltransferase	SP**LALELGTYLGYGALRIARAAPEARVYSVELAE**
	Rv1777†	cytochrome P450 144 cyp144. Cytochromes P450 are a group of heme-thiolate monooxygenases.	VRRSPKGSPGAVLDLQRRVDQAVSADHAELMTIAKDANTFFGAESVQDPY PLYERMRAAGSVHRIANSDFYAVCGWD
	Rv2095c	Proteasome accessory factor C PafC	MVCAPTGAGKTVVGEFAVH
	Rv2249c	Probable glycerol-3-phosphate dehydrogenase GlpD1. Involved in aerobic respiration and oxidation of glycerol.	ARRSADLTALADGGALDVIVIGGGITGVGIALDAATRGLTVALVEKQ
	Rv2277c	Possible glycerolphosphodiesterase, function unknown.	SSWLALSGTVRYCAPRIAHVSTVGKA&RGP**MVKTAMLGAVALVIALGGT CGVADALPLGQTDDPMIVAHRAGTRDFPENTVLAITNAVAAGVDGMW LTVQVSSDGV**
	Rv3119*	Probable molybdenum cofactor biosynthesis protein E MoaE1	MANVVAEGAYPYCRLTDQPLSVDEVLAAVSGPEQGGIVIFVGNVRDHNAG HDV
	Rv3157	Probable NADH dehydrogenase I (chain M) NUOK (NADH-ubiquinone oxidoreductase chain M). Involved in aerobic|anaerobic respiration	FESPAANANRLRP#LRVIDAGRSGVSGGGAAGGAAVV**NNVPWLSVLWLV PLAGAVLIILLPPGRRRLAKWAGMVVSVLTLAVSIVVAAEFKPSAEPYQFV EKHSWIPAFGAGYTLGVDGIAVVLVLLTTVLIPLLLVAGWNDATDA**

Functional category	Rv no.	Description	Peptide sequence
	Rv3248c	Probable adenosylhomocysteinase SahH.	TRFETDKDKWTKIAESVKGVTEETTTGVLRLYQFAAAGDLAFPAINVNDSVTK
	Rv3273	Probable transmembrane carbonic anhydrase (carbonate dehydratase) (carbonic dehydratase). Generates CO(2) and H(2)O from H(2)co(3), and possibly involved in transport of sulfate across the membrane.	LVVFRNHHPARRSAESAGYPEADQLSIVNVAVQVERLTRHPILATAVAAADLQVIGIFFD
	Rv3309c	Probable uracil phosphoribosyltransferase Upp. Involved in pyrimidine salvage pathway	LVGQRRPMRPATMLPGHRSPPRHGIPWSGAG&EPGFRGACREY**MDGVDRSRGWTHPYQPPFRGPSHDCYIGFNAVQVHVVDHPLAAARLTTLRDERTDN**
	Rv3318	succinate dehydrogenase flavoprotein subunit sdhA	AYELATGDIHVFHAKAVVIATGGSGRMYKTTSNAHTLTGDGIGIVFRKGLPLEDMEFHQFHPTGLAGLGILISEAVRGEGGRLLNGEGERFMERYAPTIVDLAPRDIVARSMVLEVLEG
	Rv3324c	Probable molybdenum cofactor biosynthesis protein C 3 MoaC3. Thought to be involved in the biosynthesis of molybdopterin.	GIE**AVTVTLEPQGADRLSIAATVTTVART**
	Rv3368c	Possible oxidoreductase. Function unknown; probably involved in cellular metabolism	TLNLSVDEVLTTTRSVRKRLDFDKPVPRDVLMECLELALQAPTGSNSQGWQWVFVEDAAKKKAIADVYLANARGYLSGPAPEYPDGDTRGERMGRVRDSAT
	Rv3601c	Probable aspartate 1-decarboxylase precursor PanD	LRTMLKSKIHRATVTCADLHYVGSVTIDADLMDAADPA&GWGHRRWIREPA
lipid metabolism	Rv0222	Probable enoyl-CoA hydratase EchA1. Identified by mass spectrometry in M. tuberculosis H37Rv-infected guinea pig lungs at 30 days but not 90 days (See Kruh et al., 2010).	AGLSVELTVVESAMAQA&HETPDEYEASMSSESDAANTEPEVLVEQRDRIL
	Rv0405*	Probable membrane bound polyketide synthase Pks6. Polyketide synthase possibly involved in lipid synthesis.	PAPSVSDPTAAGQIGASDGGAELLASSGFAARLAGRSADEQLAAAIEVVCEHAAAVLGRDGAAGLDAGQAFADSGFNSLSAVELRNRLTA

Functional category	Rv no.	Description	Peptide sequence
	Rv1141c*	Probable enoyl-CoA hydratase EchA11 (enoyl hydrase). Could possibly oxidize fatty acids using specific components	**ANTKVALMPDGGASALVAAATGRIRAMRLALLAEQLPAREALAWGLISAVYPDSDFEAEVDKVIS**QVAMAVTAATVIPAMAPAALAATVV
	Rv1142c	Probable enoyl-CoA hydratase EchA10. Could possibly oxidize fatty acids using specific components.	MHMALLPDRVPAAEALSWGLVSAVYPAADFDAEVDKLISRLLAGPALAIAKTKNAINAATLTELAPTLLRELDGQALLLRTD
	Rv1886c†	Secreted antigen 85-B FbpB (85B) (antigen 85 complex B). Involved in cell wall mycoloylation. Proteins of the antigen 85 complex are responsible for the high affinity of mycobacteria to fibronectin. Predicted possible vaccine candidate (See Zvi et al., 2008).	**TVCSGPSTRRHTGQRGTGMTDVSRKIRAWGRRLMIGTAAAVVLPGLVGLAGGAATAGA**FSRPGLPVEYLQV
	Rv3141	Probable NADPH quinone oxidoreductase FadB4. Involved in lipid degradation. Thought to be differentially expressed within host cells (see Triccas et al., 1999). mRNA identified by microarray analysis and up-regulated after 96h of starvation (see Betts et al., 2002).	**DALAQQWSQLERLLRSGKLPPPEPVVYPLDQAAAAIASLENRTAKGKVVLRVRD**#RPSRDASPACSGQFAASSLVAVGVGY EFL
	Rv3229c	Possible linoleoyl-CoA desaturase (delta(6)-desaturase). Thought to be involved in lipid metabolism	APYLSDHGGAGVAHRVARIG**HHLYPDLPSNRLHEISVRVREVCDRYDLPYTTGSFLVQYGKT**
	Rv3280	Probable propionyl-CoA carboxylase beta chain 5 AccD5 (pccase). Key enzyme in the catabolic pathway of odd-chain fatty acids, isoleucine, threonine, methionine, and valine.	LFSQDATVFGGSLGEVYGEKIVKVQELAIKTGRPLIGINDGAGARIQEGVVSLGL

Functional category	Rv no.	Description	Peptide sequence
	Rv3506	Fatty-acid-CoA synthetase FadD17. Function unknown, but supposed involvement in lipid degradation.	GLNPVRRGAALAGDIAKADCQLVLTGSGSAEVPADVEHINVDSPEWTDEV AAHRDTEVRFRSADLADLFMLIFTSGTSGDPKAVKCSHRKVAIAGVTITQRF SLGRDD
PE/PPE	MT2423	PPE family protein. Rv2353c is halve size of MT2423	SGNTGDTNVGSGNIGNTNLGGG
	Rv0278c*	PE-PGRS3 family protein. NOVEL!!	PAGSTQPAGTAGTADCSELAAPAGPARTSGPVGTAGTADCSE
	Rv0304c*	PPE family protein PPE5This region is a possible MT-complex-specific genomic island (See Becq et al., 2007).	RNVGIFDGGNSNSGSFNVGFQNTGFGNSGAGNTGFFNAGDSNTGFANA GNVNTGFFNGGDINTGGFNGGNVNTGFGSALTQAGANSGFGNLGTGNS GWGNSDPSGTGNSGFYNTSTSDLATPAFNSGLANISTSIAGLLRDSTGTMV LNLGLANHGTLNVGIANLGDYNIGFANLGSANFGSANIGGNNIGGANTGIF DIGLANLGSYNIGFGNFGDDNLGFGNLGSYNVGFGNLGNDNLGFANTGS NNIGFANTGSNNIGIGLTGDGQIGFGSLNSGSGNIGLFNSG
	Rv0915c	PPE family protein PPE14. Function unknown. Possibly a protective antigen involved with the early control of infection.	STLIVEPWMGPAAAAMAAAATPYVGWLAATAALAKETATQARAAAEAFG TAF
	Rv0915c	PPE family protein PPE14. Function unknown. Possibly a protective antigen involved with the early control of infection.	EPWMGPAAAAMAAAATPYVGWLAATAALAKETATQARAAAEAFGTAFA M
	Rv1396c	PE-PGRS family protein PE_PGRS25. Function unknown. This region is a possible MT-complex-specific genomic island (See Becq et al., 2007).	TTSLLIPTGFRGTVMSFLFAQPEMLGAAATDLASIGSAISTANAAAAAATTR VLAAGADEVSAAVAALFS

Functional category	Rv no.	Description	Peptide sequence
	Rv1806*	PE family protein PE20. Function unknown. This region is a possible MT-complex-specific genomic island (See Becq et al., 2007).	**AFVLVCPDALAIAAGQLRHVGSVIAARNAVAAPATAELAPAAADEVSALT ATQFNFHAAMYQAVGAQAIAMNEAFVAMLGASADSYAATEAANIIA**S APASRSAPANAKARPSQPRASVCQPSCRNND
	Rv3144c*	PPE family protein PPE52. Function unknown. Identified by mass spectrometry in M. tuberculosis H37Rv-infected guinea pig lungs at 90 days but not 30 days (See Kruh et al., 2010).	TAAQSFASVTAGLAGQAW
	Rv3345c*	PE-PGRS family protein PE_PGRS50. Function unknown	
	Rv3508*	PE-PGRS family protein PE_PGRS54. Function unknown. Identified by mass spectrometry in M. tuberculosis H37Rv-infected guinea pig lungs at 90 days but not 30 days (See Kruh et al., 2010). mRNA identified by DNA microarray analysis and possibly down-regulated by hspR|Rv0353 (see Stewart et al., 2002).	NGGAGGAGGTPTGSGTEGTGGDGGDAGAGGNGGSATGVGNGGNGGD GGNGGDGGN
	Rv3595c	PE-PGRS family protein PE_PGRS59. Function unknown.	
regulatory proteins	Rv2258c	Possible transcriptional regulatory protein. Possibly involved in transcriptional regulation	ESRR&RCRAKR&GGAAQ**MSGALETTEEFGNRFVAAIDSAGLAILV**
unknown	C3*	null_M.afr	GGNSELHDAIHRFLLICLPR
	MT1330.1*	Hypothetical protein	AALGGVIPSACSCARN**ASSVELALAVPAKPSAARPKVAAVKVVPTMEAAK RRVNIDDLLVLMSSEVALVCVPSESAPGLDRFS**
	MRGA327*	response regulator receiver domain-containing protein. Histidine kinase?	RRSSRVPTSASGV**AATAPNCGEA**
virulence, detoxification,	Rv0353	Probable heat shock protein transcriptional repressor HspR (MerR family). Involved in	SLHDVELLRQVQHLSQDEGVNLAGIKRIIELTSQVEALQSRLQEMAEELAVL RANQRQ

Functional category	Rv no.	Description	Peptide sequence
adaptation		transcriptional regulation (repression) of heat shock proteins	
	Rv1566c	Possible Inv protein. Function unknown	T**MMLAPMVAAPGLAAADPATRPVDYQQITD**VVIARGLSQRGVPFSWAGGGISGPT
	Rv1566c	Possible Inv protein, function unknown	RAEAHTTRECQGGYNPMKRSMKSGSFAIGLA**MMLAPMVAAPGLAAADPATRPVDYQQITD**
	Rv1839c*	Possible antitoxin VapB13, M. xenopi	MSKRLQVLLDPDEWEELREIARRHRTTVSEWV

Table S3. 3: List of active-TB UniProt identifiers retrieved from UniProtKB as WEGO input data.

UniProt ID	Gene Ontology IDs						
P9WNF5	GO:0005576						
P9WLS9	GO:0005886	GO:0016021					
P9WFN5	GO:0005576						
P9WFL9	GO:0005886						
O05436	GO:0005886						
O53740	GO:0005576						
O05899	GO:0005886						
P96214	GO:0005618	GO:0005886					
Q79FP2	GO:0005618						
O06311	GO:0005576	GO:0005618	GO:0005887				
O06823	GO:0005576						
O05445	GO:0005886						
P9WIQ7	GO:0005576	GO:0005886	GO:0009405	GO:0016021	GO:0044119		
P9WJU7	GO:0005887	GO:0006855	GO:0006869	GO:0009405	GO:0044119	GO:0052572	GO:0071555
Q79FB3	GO:0005576	GO:0009405	GO:0009986	GO:0044164			
Q79FU2	GO:0009405						
Q6MWY2	GO:0040007	GO:0052170					
Q6MWX9	GO:0052556						
P9WQA9	GO:0008784	GO:0009252	GO:0030170	GO:0030632	GO:0040007		
P9WP89	GO:0009236	GO:0016994					
P9WMP7	GO:0006782	GO:0008883	GO:0040007	GO:0050661			

UniProt ID	Gene Ontology IDs						
P9WK19	GO:0005506	GO:0006555	GO:0016151	GO:0030145	GO:0035551	GO:0050897	GO:0070006
P9WLF0	GO:0006629	GO:0008889					
L7N6A5	GO:0009058	GO:0016779					
O69728	GO:0018741	GO:0018909	GO:0046983				
O69719	GO:0002100	GO:0008270	GO:0040007	GO:0052717			
P0CV93	GO:0004540	GO:0046872					
I6XD69	GO:0016491	GO:0031177					
P9WQN7	GO:0001968	GO:0005576	GO:0016746	GO:0035375			
P9WQM5	GO:0005886	GO:0015171	GO:0016021				
P71922	GO:0005524	GO:0005886	GO:0016887				
P71616	GO:0015238	GO:0015297	GO:0016021				
O06312	GO:0005886	GO:0015095	GO:0016021	GO:0046872			
P71662	GO:0005618	GO:0016491					
P95105	GO:0005829	GO:0010181	GO:0052873				
P95100	GO:0000976	GO:0003700	GO:0005829	GO:0005886			
O53380	GO:0004175	GO:0009002	GO:0016021				
P9WMZ7	GO:0016021	GO:0016758	GO:0040007	GO:0045227	GO:0071555		

UniProt ID	Gene Ontology IDs								
P9WQB5	GO:0004601	GO:0005829	GO:0005886	GO:0006979	GO:0015036	GO:0032843	GO:0045454	GO:0051920	
P9WIR7	GO:0005576	GO:0009267	GO:0009986	GO:0050840	GO:0052559				
P9WPX7	GO:0003855	GO:0005829	GO:0009073	GO:0009423	GO:0040007				
P9WPQ9	GO:0004322	GO:0005576	GO:0005829	GO:0005886	GO:0006826	GO:0008199	GO:0010039	GO:0033214	
P9WPU1	GO:0005524	GO:0005576	GO:0005618	GO:0005886	GO:0006825	GO:0016021	GO:0019829	GO:0046872	
P9WP29	GO:0003951	GO:0004143	GO:0005524	GO:0005576	GO:0005618	GO:0005886	GO:0008654	GO:0009247	
	GO:0003677	GO:0005576	GO:0005618	GO:0005829	GO:0005886	GO:0006351	GO:0006355	GO:0009295	GO:0009405
P9WJB7	GO:0032993	GO:0043565	GO:0050708	GO:0052572					
O69687	GO:0005618	GO:0005886	GO:0008168	GO:0008610					
P9WQ59	GO:0004321	GO:0005524	GO:0005886	GO:0006633	GO:0008610	GO:0016874	GO:0044119	GO:0052170	GO:0052572
	GO:0070566	GO:0071770							
P9WGD9	GO:0005525	GO:0005618	GO:0005737	GO:0005886	GO:0006614				
P9WK25	GO:0004470	GO:0004471	GO:0005829	GO:0006090	GO:0006108	GO:0008948	GO:0046872	GO:0051287	
P9WG65	GO:0005576	GO:0005623	GO:0016209	GO:0016491	GO:0045454	GO:0055114			
P9WIF5	GO:0005576	GO:0005618	GO:0009279	GO:0009405	GO:0009986	GO:0042594	GO:0046789	GO:0052167	

UniProt ID	Gene Ontology IDs							
O06582	GO:0001101	GO:0003994	GO:0005618	GO:0005829	GO:0006099	GO:0008203	GO:0019629	GO:0019679
P9WQ21	GO:0005829	GO:0005992	GO:0030980	GO:0047470				
P9WN35	GO:0000162	GO:0004049	GO:0005886	GO:0005950	GO:0006541			
P9WFT7	GO:0004828	GO:0005524	GO:0005737	GO:0005886	GO:0006434	GO:0040007	GO:0097056	
P9WG49	GO:0000287	GO:0003677	GO:0003917	GO:0005618	GO:0005829	GO:0005886	GO:0006265	GO:0040007
P9WMG1	GO:0003677	GO:0005618	GO:0005829	GO:0006351	GO:0006355			
P9WLV1	GO:0005975	GO:0008194	GO:0016758	GO:0030259	GO:0043231			
P95072	GO:0005618	GO:0005886	GO:0006465	GO:0008233	GO:0016021			
I6YAZ1	GO:0005737	GO:0008757	GO:0032259					

Table S3. 4: List of latent-TB UniProt identifiers retrieved from UniProtKB as WEGO input data.

UniProtID	Gene Ontology IDs		
P9WIN7	GO:0005576	GO:0005618	
P9WIL9	GO:0005618		
P9WM15	GO:0005576		
P9WLX7	GO:0005618	GO:0005886	
P9WKL3	GO:0005576	GO:0005618	
O50416	GO:0005576	GO:0005886	
P95021	GO:0016021		
P95152	GO:0016020		
I6XF52	GO:0016021		
L0T243	GO:0005887		
P94973	GO:0016021		
O53930	GO:0005576	GO:0005829	GO:0016021
O05854	GO:0005829	GO:0005886	GO:0016021
O06624	GO:0005576		
I6Y4F1	GO:0016021		
O53514	GO:0016021		
O53417	GO:0005829	GO:0005886	GO:0005887
O53953	GO:0005576		
O07423	GO:0005576		
O50383	GO:0005576		
O07222	GO:0005576		
O69704	GO:0005737	GO:0005829	GO:0016021

UniProtID	Gene Ontology IDs						
P9WJV7	GO:0005618	GO:0005886	GO:0006810	GO:0016021	GO:0044119		
P9WJU1	GO:0005576	GO:0005618	GO:0005829	GO:0005886	GO:0006869	GO:0016021	
P71915	GO:0016021	GO:0055085					
P96258	GO:0005576	GO:0005887	GO:0040007				
P9WJ51	GO:0006355						
P9WIJ3	GO:0006412	GO:0031365	GO:0035601	GO:0040007	GO:0042586	GO:0043686	GO:0046872
P9WJR5	GO:0006777	GO:0061799					
P9WGZ3	GO:0000049	GO:0004526	GO:0008033	GO:0040007			
P9WLF1	GO:0006629	GO:0008889					
P96404	GO:0004300	GO:0006635					
O86335	GO:0008610	GO:0009405	GO:0016788	GO:0031177	GO:0044119		
O06541	GO:0003824	GO:0008152					
O53464	GO:0003677						
I6YCF3	GO:0004386	GO:0005524					
O50397	GO:0016491						
I6YGX2	GO:0008658						
A0A089QH62	GO:0016301						
P9WPL1	GO:0004497	GO:0005506	GO:0005618	GO:0016705	GO:0020037		
P9WQJ1	GO:0005524	GO:0005618	GO:0005886	GO:0016021	GO:0042626		
P9WGQ9	GO:0005886	GO:0016491					
Q6MX47	GO:0016021	GO:0046873					

UniProtID	Gene Ontology IDs								
P95137	GO:0005618	GO:0005829	GO:0005886	GO:0008171					
O06152	GO:0005576	GO:0016021	GO:0016757						
P95185	GO:0005886	GO:0016491							
I6XHB8	GO:0016021	GO:0022857							
O53762	GO:0004499	GO:0005886	GO:0050660	GO:0050661					
O50440	GO:0005618	GO:0005886	GO:0006629	GO:0016021	GO:0016746				
P9WP43	GO:0005576	GO:0005618	GO:0019626	GO:0034338	GO:0047372	GO:0051793	GO:0052651	GO:0052689	
P9WNZ3	GO:0005886	GO:0016213	GO:0042759	GO:0046872					
P9WMR3	GO:0003677	GO:0003678	GO:0004519	GO:0005524	GO:0005618	GO:0005829	GO:0006268	GO:0006269	GO:0006314
	GO:0006974	GO:0016539	GO:0040007	GO:1990077					
P9WNT7	GO:0003677	GO:0003887	GO:0005618	GO:0005737	GO:0005886	GO:0006260	GO:0008408	GO:0040007	
P9WMR5	GO:0003677	GO:0005524	GO:0005886	GO:0006281	GO:0006310	GO:0008026			
P9WNL5	GO:0005829	GO:0005886	GO:0009247	GO:0016021	GO:0040007	GO:0046677	GO:0052636	GO:0071555	GO:0071766
O53551	GO:0004467	GO:0005524	GO:0005829	GO:0005886	GO:0008610	GO:0042759	GO:0071766	GO:0102391	
P9WN81	GO:0005618	GO:0005886	GO:0006071	GO:0006072	GO:0009331	GO:0052591			
P71707	GO:0005886	GO:0008360	GO:0008658	GO:0008955	GO:0009002	GO:0009252	GO:0016021	GO:0030288	GO:0071456

UniProtID	Gene Ontology IDs								
	GO:0071555								
P9WQH7	GO:0003989	GO:0004658	GO:0005524	GO:0005618	GO:0005886	GO:0009317	GO:0015977	GO:0043234	
P9WKK1	GO:0003743	GO:0003924	GO:0005525	GO:0005576	GO:0005618	GO:0005737	GO:0005886	GO:0040007	
P9WJR3	GO:0005829	GO:0006777	GO:0030366	GO:0032324					
P9WJQ3	GO:0001101	GO:0005576	GO:0005618	GO:0005886	GO:0008940	GO:0009055	GO:0009325	GO:0030151	GO:0042128
	GO:0043602	GO:0051539							
P9WIL3	GO:0004068	GO:0005618	GO:0005829	GO:0006523	GO:0009405	GO:0015940			
P9WMZ9	GO:0000026	GO:0005886	GO:0008654	GO:0009247	GO:0009405	GO:0016021	GO:0046488		
P9WH65	GO:0000028	GO:0000462	GO:0003735	GO:0005886	GO:0006412	GO:0022627	GO:0040007	GO:0048027	GO:0070181
O07747	GO:0005576	GO:0009405	GO:0010628	GO:0010629	GO:0016787	GO:0040010			
P9WGV3	GO:0004013	GO:0005576	GO:0005618	GO:0005829	GO:0005886	GO:0006730	GO:0009087	GO:0019510	GO:0033353
	GO:0035375	GO:0035635	GO:0040007	GO:0044650	GO:0046085	GO:0070403			
O53361	GO:0003993	GO:0004438	GO:0004805	GO:0005576	GO:0005618	GO:0006742	GO:0009405	GO:0044161	GO:0046854
	GO:0050189	GO:0050192							
P9WFF3	GO:0004845	GO:0004849	GO:0005525	GO:0005829	GO:0005886	GO:0006206	GO:0006223	GO:0043097	GO:0044206

UniProtID	Gene Ontology IDs						
P9WMY3	GO:0005886	GO:0016758	GO:0040007	GO:0045226	GO:0102096		
O06542	GO:0003824	GO:0005829	GO:0005886	GO:0008152			
O06291	GO:0004252	GO:0005618	GO:0005886	GO:0005887	GO:0040007		
O53532	GO:0005618	GO:0005737	GO:0005886	GO:0008757	GO:0032259	GO:0071456	
P95034	GO:0005623	GO:0006124	GO:0008860	GO:0009055	GO:0045454	GO:0050660	GO:0051287
P96878	GO:0004089	GO:0005886	GO:0005887	GO:0008270	GO:0008271	GO:0015116	GO:0015976
P96287	GO:0005524	GO:0005737	GO:0006271				
P96360	GO:0003677	GO:0004803	GO:0005618	GO:0005886	GO:0006313		
P72045	GO:0005576	GO:0009247	GO:0016021	GO:0016049	GO:0016758		
Q93IG6	GO:0000160	GO:0004673	GO:0004871	GO:0005524	GO:0005829	GO:0005886	GO:0016021
O06302	GO:0000984	GO:0005886	GO:0045892				

Table S4. 1: List of 97 B-cell epitopes predicted using BCPred from immunogenic *Mtb* peptides recognized by humoral response of TB patients.

Gene ID	Locus Tag	Selected Peptide Sequence	BCPred B-Cell Epitope	Pos
886796	Rv0179c	MWIRAERVAVLTPTASLRRLTACYAALAVCAALACTTGQPAARAADG	ACTTGQPAARAADGREMLAQ	35
886478	Rv0361	**AKETQVIVTAHEAAT**EVFQT**NQAPTTP**VARTTPVATIDAPTCTATIGSILANRLLRSEATAVEIC	APGFDAKETQVIVTAHEAAT	70
			NQAPTTPPRMPTGMPPKTAV	95
887569	Rv0559c	KGTKLAVVVGMTVAAVSLAAPAQADDYDA**PFNNTIHRFGIYGPQDYNAW**	PFNNTIHRFGIYGPQDYNAW	31
885077	Rv0978c	MSFVNVAPQLVSTAAADAA**RIGSAINTANTAAAATTQVL**AAAQDEVSTAIAALFGSHGQH	RIGSAINTANTAAAATTQVL	20
886048	Rv1009	SPMLRLVVGALLLVLAFAGGYAVAACKTVTLTVDGTAMRVTTMKSRVIDIVEE**NGFSVDDRDDLYPA AGVQVH**DADTIVLRRSRPL	NGFSVDDRDDLYPAAGVQVH*	52
887033	Rv1268c	KGPTMTTSKIATAFKTATFALAAGAVALGL**ASPADAAAGTMYGDPAAAAK**	ASPADAAAGTMYGDPAAAAK	27
887039	Rv1269c	TLRRRFAVAVAGVATA**AATTVTLAPAPANAADVYGAI**AYSGNGSWGRSWDYPTRAAAEATAVKSC GYSDCKVLTSFTACGAVAANXRAYQ	ATTVTLAPAPANAADVYGAI	23
887019	Rv1271c	NTKDEAFIAQMESIGVTFSSPQVATQQAQLVCKKLASGETGTEIAEEVLSQTNLTTKQAAYFVVDATK AYCPQYASQLT	KDEAFIAQMESIGVTF	3[a]
			AQMESIGVTFSSPQVA	9[a]
			AQLVCKKLASGETGTE	28[a]
			GTEIAEEVLSQTNLTT	41[a]
			EVLSQTNLTTKQAAYF	47[a]
			NLTTKQAAYFVVDATK	53[a]
			YFVVDATKAYCPQYAS	61[a]
		FTTAVGAAAIGLAVATAGTAGANTKDEAFIAQME	TTAVGAAAIGLAVATA	2[a]
			GLAVATAGTAGANTKD	11[a]
			AGTAGANTKDEAFIAQ	17[a]
886975	Rv1291c	MFTRRFAASMVGTTLTAATLGLAALGFAGTASASSTDEAFLAQLQADGITPPSAARAIKDAHAVCDA LDEGHSAKAVIKAVAK	FTRRFAASMVGTTLTA	2[a]
			ASMVGTTLTAATLGLA	8[a]
			AGTASASSTDEAFLAQ	28[a]
			TDEAFLAQLQADGITP	36[a]
			ADGITPPSAARAIKDA	46[a]

Gene ID	Locus Tag	Selected Peptide Sequence	BCPred B-Cell Epitope	Pos
			PSAARAIKDAHAVCDA	52[a]
			AHAVCDALDEGHSAKA	61[a]
			ALDEGHSAKAVIKAVA	67[a]
886889	Rv1329c	VSESVSMSVPELLAIAVAALGGTRRRGQQEMAAAVAHAFETGEHLVVQAGTGTG	ESVSMSVPELLAIAVA	3[a]
			GGTRRRGQQEMAAAVA	21[a]
			QQEMAAAVAHAFETGE	28[a]
			AFETGEHLVVQAGTGT	38[a]
886745	Rv1396c	TTSLLIPTGFRGTVMSFLFAQPEMLGAAATDLASIGSAISTANAAAAAATTRVLAAGADEVSAAVAALFS	IPTGFRGTVMSFLFAQ	6[a]
			SFLFAQPEMLGAAATD	16[a]
			LGAAATDLASIGSAIS	25[a]
			GSAISTANAAAAAATT	36[a]
			AAAAATTRVLAAGADE	45[a]
			VLAAGADEVSAAVAAL	53[a]
886685	Rv1424c	SDPRPPTETPTRPKTGHSPRCPTSSR**RPPAEKLVFPVLLGILTLLL**SA**CQTASASGYNEPRGYDRATL**KLV	RPPAEKLVFPVLLGILTLLL*	42
			CQTASASGYNEPRGYDRATL	64
886653	Rv1435c	LTIAISVNSPLFARRYFRNQFGSAEPHSRIEFLFDHRLNCQHPMGNMSPAAPGRFQMV	NSPLFARRYFRNQFGS	8[a]
			NQFGSAEPHSRIEFLF	19[a]
			HSRIEFLFDHRLNCQH	27[a]
			FDHRLNCQHPMGNMSP	34[a]
			PMGNMSPAAPGRFQMV	43[a]
886605	Rv1450c	MSLVIVTPETVAAAASDV**ARIGSSIGVANSAAAGSTTS**VLAAGADEVSAAIATLFGSHAREYQAISTQVAAFHDRFAQTLSAAVG**SYVSAEATNAAPLATLEHNV**LNALNAPTQALLGRPLI	ARIGSSIGAANAAAAGSTTS	19
			SYVSAEATNAAPLATLEHNV	86
885588	Rv1804c	MRVVSTLLSIP**LMIGLAVPAHAGPSGDDAVF**LASLERAGITYSHPDQAIASGKAVCAL**VESGESGLQVVNE**	LMIGLAVPAHAGPSGDDAVF	12
			VESGESGLQVVNELRTRNPG	59
885591	Rv1810	TRKVCRALAADLPQDAMQLQRTMGQCRPMRMLVALLLSA**ATMIGLAAPGKADPTGDDAA**FLAALDQAGITTLTQATP#V	ATMIGLAAPGKADPTGDDAA	24

Gene ID	Locus Tag	Selected Peptide Sequence	BCPred B-Cell Epitope	Pos
		MRMLVALLLSA**ATMIGLAAPGKADPT**		
885546	Rv1813c	MITNLRRRTAMAAAGLGAALGLGILLVPTVDAHLANGSMSE	RRTAMAAAGLGAALGL	7[a]
			AAGLGAALGLGILLVP	13[a]
			GLGILLVPTVDAHLAN	21[a]
885896	Rv1860	SLVTVAVPA**TANADPEPAPPVPTTAASPP**STAAAPPAPATP**VAPPPPAAANTPNAQPGDPN**AA	TANADPEPAPPVPTTAASPP	36
			TAAAPPAPATPVAPPPPAAA	57
			PGDPNAAPPPADPNAPPPPV	83
			VAPPPPAAANTPNAQPGDPN*	68
885759	Rv1884c	TRIAKPLIKSAMAAGLVTASMSLSTAVAHAGPSPNWDAGAQ	GNASATSGDMSSMTRIAKPL*	25
885785	Rv1886c	MTDVSRKIRAWGRRLMIGTAAAVVLPGLVGL**AGGAATAGAFSRPGLPVEYL**QV	AGGAATAGAFSRPGLPVEYL	32
		MTDVSRKIRAWGRRLMIGTAAAVVLPGLVGL**AGGAATAGA**		
885897	Rv1910c	VAHAFHRFALAILGLALPVALV**AYGGNGDSRKAAPLAPKAAA**LGRSMPETPTGDVL**TISSPAFADGA PIPEQYTCK**	AYGGNGDSRKAAPLAPKAAA	23
			TISSPAFADGAPIPEQYTCK	57
885334	Rv1926c	WFSAIPRRSPP#PAASSVRLKVMTSARFEPGNRPNK@WR@GPMKLTTMIKTAVAVVAMAAI**ATF AAPVALAAYPITGKLGS**EL**TMTDTVGQ**	ATFAAPVALAAYPITGKLGS*	20
		WR@GPMKLTTMIKTAVAVVAMAAI**ATFAAPVALAAYPITGKLGS**EL**TMTDTVGQVVL**ADVDNQL TVGTDLDQGSFVTAGLDADDHR	TMTDTVGQVVLGWKVSDLKS*	42
		LRLKVMTSARFEPGNRPNK@WR@GPMKLTTMIKTAVAVVAMAAI**ATFAAPVALAAYPITGKLGS** E		
885925	Rv1980c	LSLVRHRRQQRDALCLSSTQISRQSNLPPAAGGAANYSRRNFDVRIKIFMLVTAVVLLCCSGV**ATAAP KTYCEELKGTDTGQA**CQIQMSDPAYN	ATAAPKTYCEELKGTDTGQA	21
		LWQAD**TDPLPVVFPIVQGELSKQTG**QQVS**IAPNAGLDPVNYQNFAVTND**GVIFFF**NPGELLPEAA GPTQVLVPRS**AIDSMLA	TDPLPVVFPIVQGELSKQTG	152
		DRSASLSLVRHRRQQRDALCLSSTQISRQSNLPPAAGGAANYSRRNFDVRIKIFMLVTAVVLLCCSGV **ATAAPKTYCEELKGTDTGQA**CQIQMSDPAYNINISLPSYYPDQKSLENYIAQTRDKFLSA	IAPNAGLDPVNYQNFAVTND	176
			NPGELLPEAAGPTQVLVPRS	202
885813	Rv1984c	PRSLVR**IVGVVVATTLALVSAPAGGR**AAHADPCSDIAVVFARGTHQA	IVGVVVATTLALVSAPAGGR*	9
			SAPAGGRAAHADPCSDIAVV	22
885515	Rv2376c	VKSIAAGLTAAAAIGAAAAGVTSIM**AGGPVVYQMQPVVFGAPLPL**DP	AGGPVVYQMQPVVFGAPLPL	29

Gene ID	Locus Tag	Selected Peptide Sequence	BCPred B-Cell Epitope	Pos
888584	Rv2499c	HAG**DRESDDAVSACRVAGS**TVGRRILQ	DRESDDAVSACRVAGS	4[a]
			SACRVAGSTVGRRILQ	12[a]
887184	Rv2878c	R#KLARKLY#QFSRTILNAEIKRTMHAPERRGRRDVLNRHGAYYVIRSKHEQVRSVMSLRLVSPIKAFA DGIVAVAIAVVLMFGLANTPRAVAADERLQF**TATTLSGAPFDGASLQGKPA**VLWFWTPWCP**FCNA EAPSLSQVAAANPA**	TATTLSGAPFDGASLQGKPA	45
			FCNAEAPSLSQVAAANPAVT*	75
887636	Rv2944	**SNKAFGRWGEVFGGDDV**VAAAMIDRLVHHAE	IVTSNKAFGRWGEVFGGDDV	178
887320	Rv3036c	TRFRGRAAQCRQV@TDVADDLAVDDPRRPDHAPRQGIDLVTGADGGKHPDAAPTPLVCEDD@T MRYLIATAVLVA**VVLVGWPAAGAPPSCAGLGG**TVQAGQIC**HVHASGPKYMLDMTFPVD**	VVLVGWPAAGAPPSCAGLGG	13
			HVHASGPKYMLDMTFPVDYP	41
887931	Rv3212	ERRTKTDIAAAATIAVVVAVAASLIW**WTSDARATISRPAAV**	WTSDARATISRPAAVAVPTP	31
888906	Rv3218	HPGPLPSSSTGRPGPHRHTA@AHCRPPQRADPARIVTRPKPIAVLLNGPTRPASSHGLSSCVMRA**VL IVNPTATATTPAGRDLLA**HALESRLQLTVEHTNHRGHGTELGQAAVADGVDLVVVHGGDGTVSAV VNGML**GRPGTTPVRPVPAVAVVPGG**SANVLARALGI	VLIVNPTATATTPAGRDLLA	4
		ASPAMPWMHA@PPVISGSAAASGLNAI**AVAVVPGG**SANVLARALGISADPIAATNQLIQLLDDYG RHQQWRRIGLI**DCGERWAVF**	GRPGTTPVRPVPAVAVVPGG	76
			DCGERWAVFNAGMGVDAEVV	137
888696	Rv3256c	A**ETAGTILASTLGAGA**	ETAGTILASTLGAGAAEPIV*	68
887632	Rv3333c	IMFTGIASHAGALGAALVVLIGAA**ILHDGPAAADPNQDDRFLAL**L**EKKEIPAVANVPRVIDAAHK**	ILHDGPAAADPNQDDRFLAL	24
			EKKEIPAVANVPRVIDAAHK*	45
888114	Rv3345c	GKGGNGG**QGGIGGAGERGADGAG***PNANGANGEN**GGSGGN***GGDGGAGGNG**GAGGKA	QGGIGGAGERGADGAG	8[a]
			GERGADGAGPNANGAN	15[a]
			PNANGANGENGGSGGN	24[a]
			GGSGGNGGDGGAGGNG	34[a]
			GGDGGAGGNGGAGGKA	40[a]
3205044	Rv3395A	TQQVSVRTNTKSTQNDT#RPAC**RPAELFATDNTTDGFELPAV**ATIALTGTVVTT	RPAELFATDNTTDGFELPAV	36
887227	Rv3572	VAGAN**LRTPDGPTGFPPGLWARQTT**EIRSTNRLAYLDAHATSQFERVMKAGGSDVIT	LRTPDGPTGFPPGLWARQTT	64
885751	Rv3682	**AGWDLPMSGKTGTTEA**HRSAGFVGFTNRYAAAN**YIYDDSSSPTDLCSGPLRHC**G**SGDLYGGNEPS RTWFAAMKP**IANNFGEV**QLPPTDPRYVDGAPGSRVPS**VAGLDVDAARQRLKDAGFQVADQTNS VNSSAK	SAGAAGWDLPMSGKTGTTEA*	600
			YIYDDSSSPTDLCSGPLRHC	637
			SGDLYGGNEPSRTWFAAMKP	658

Gene ID	Locus Tag	Selected Peptide Sequence	BCPred B-Cell Epitope	Pos
			QLPPTDPRYVDGAPGSRVPS	686
886121	Rv3803c	VHTALHDGGGHMKGRSALLRALWIAALSFGLGGVAVAAEPTAK**AAPYENLMVPSPSMGRDIPV**AFLAGGPHA	AAPYENLMVPSPSMGRDIPV	33
886216	Rv3896c	LRQDFIPSPVGGPINTLVSATLPAEFDALGARSTPSISSGPQTRSNAVTVALPQVNLRCQLVNTCERVASKSTCCVAVIYSLAESRCWD#LGEEDSMSFVTTQPEALAAAAAIQKALDIKGVHDPAARARWTRGMDLVA**RRESNYNANAINHWDSNAAR**GTPS	RRESNYNANAINHWDSNAAR	213
886256	Rv3922c	RVTGRASARGLIFVIQVYRHMLSPLRPASCRFVPTCSQYAVDALTEYGLLRGSWLTMIRLAKC**GPWHRGGWDPIPEGLTTGRS**CQT**DVDGANDDWNPASK**	GPWHRGGWDPIPEGLTTGRS	76
			DVDGANDDWNPASKRGERES	99

*AAP Epitope

[a]ABCPred used for prediction and start position on provided protein peptide

APPENDIX B: ETHICS APPROVAL

UNIVERSITY OF KWAZULU-NATAL
INYUVESI YAKWAZULU-NATALI

RESEARCH OFFICE
BIOMEDICAL RESEARCH ETHICS ADMINISTRATION
Westville Campus
Govan Mbeki Building
Private Bag X 54001
Durban
4000
KwaZulu-Natal, SOUTH AFRICA
Tel: 27 31 2604769 - Fax: 27 31 260-4609
Email: BREC@ukzn.ac.za
Website: http://research.ukzn.ac.za/Research-Ethics/Biomedical-Research-Ethics.aspx

19 December 2017

Mr TE Chiliza
UKZN Westville
Discipline of Microbiology
P/Bag X54001
Durban
4000
chilizat@ukzn.ac.za

Dear Mr Chiliza

PROTOCOL: Selection and identification of novel *Mycobacterium tuberculosis* phage-displayed biomarkers by immunoscreening against patient's serum samples. REF: BE236/13.

RECERTIFICATION APPLICATION APPROVAL NOTICE

Approved: 13 November 2017
Expiration of Ethical Approval: 12 November 2018

I wish to advise you that your application for Recertification received on **06 December 2017** for the above protocol has been **noted and approved** by a sub-committee of the Biomedical Research Ethics Committee (BREC) for another approval period. The start and end dates of this period are indicated above.

If any modifications or adverse events occur in the project before your next scheduled review, you must submit them to BREC for review. Except in emergency situations, no change to the protocol may be implemented until you have received written BREC approval for the change.

This approval will **be ratified** by **a** full Committee at its next meeting taking place on **13 February 2018.**

Yours sincerely

Mrs A Marimuthu
Senior Administrator: Biomedical Research Ethics