“Peptide vaccine identification against *Mycobacterium tuberculosis* through combinatorial/integrated pan-genomics and reverse vaccinology approaches”

Author

Muhammad Saleh Sarwar

Regn Number: NUST201463583MASAB92514F

Supervisor

Dr. Hussnain A. Janjua

DEPARTMENT OF INDUSTRIAL BIOTECHNOLOGY

ATTA-UR-RAHMAN SCHOOL OF APPLIED BIOSCIENCES (ASAB)

NATIONAL UNIVERSITY OF SCIENCES AND TECHNOLOGY

ISLAMABAD

2017

Author

Muhammad Saleh Sarwar

Regn Number: NUST201463583MASAB92514F

A thesis submitted in partial fulfillment of the requirements for the degree of

MS INDUSTRIAL BIOTECHNOLOGY

Thesis Supervisor:

Dr. Hussnain A. Janjua

Thesis Supervisor's Signature: _____

ATTA-UR-RAHMAN SCHOOL OF APPLIED BIOSCIENCES (ASAB)

NATIONAL UNIVERSITY OF SCIENCES AND TECHNOLOGY,

ISLAMABAD

2017

Declaration

I certify that this research work titled "Peptide vaccine identification against Mycobacterium tuberculosis through combinatorial/integrated pangenomics and reverse vaccinology approaches" is my own work. The work has not been presented elsewhere for assessment. The material that has been used from other sources it has been properly acknowledged / referred.

Signature of Student

Muhammad Saleh Sarwar

NUST201463583MASAB92514F

Language Correctness Certificate

This thesis has been read by an English expert and is free of typing, syntax, semantic, grammatical and spelling mistakes. Thesis is also according to the format given by the university.

Signature of Student

Muhammad Saleh Sarwar

NUST201463583MASAB92514F

Signature of Supervisor

Copyright Statement

Acknowledgements

First of all, praise is due to almighty **ALLAH** with His compassion and mercifulness for giving me the strength to complete this endeavor.

This dissertation appears in its current form due to the assistance and guidance of several persons. I would therefore like to offer my sincere thanks to all of them. I would like to express my gratitude to my supervisor **Dr. Hussnain Ahmad Janjua** (Assistant Professor, Atta-ur-Rahman School ofApplied Biosciences, ASAB, NUST) who taught me how to perform in a field of which I had no experience before.

I am highly grateful to Dr. Amajd Ali (GEC member) for his expert opinion on this project. The pipline that we have used for the identification of peptide vaccine candidates against *Mycobacterium tuberculosis* was initially developed by Dr. Amjad Ali and co-authors (Hassan et al., 2016) on the discovery of peptide vaccine targets in *Acinetobacter baumanni*

I would also like to thank PhD student Afreenish Hassan and my fellow colleague Fatima Kanwal for their guidance throughout, without which this project wouldn't have been possible.

I would also like to thank my friends Behrose Naeem, Muhammad Tufail, Arshad Alam and Nazeef Qazi for their motivation and support throughout the project and making the time of research at University fun for me with their company.

Lastly, I would specially like to thank my parents and brother Arif Sarwar, without their prayers and support, I wouldn't have been in position of doing my MS degree.

Dedicated to my exceptional mother

Table of Contents

List of Figures

//

List of Tables

ABSTRACT

Mycobacterium tuberculosis has become an eminent healthcare concern for society because of its growing rates of morbidity and mortality in cases pulmonary and extra-pulmonary tuberculosis infections. The alarming situation of antibiotic resistance and lack of protective response of current vaccines requires cost–effective potent vaccine development. In the current study we have scrutinized the Mycobacterium tuberculosis genome utilizing integrated pangenomics and reverse vaccinology approaches.

The pan-genome analysis of 47 completely sequenced strains avaiable at the time of analysis revealed 5,069 functional genes and 3,170 (62% of the pan-genome) were found to beconstituting the core-genome. The phylogenetic analysis intra and inter genome homology on the basis of geographical distribution. Among the conserved eight proteins found to have a potent antigenic potential namely, Esterase, Secreted antigen 85-C (85C), PPE family protein, ESX conserved component 5, lysine-N-oxygenase, ESX-2 secretion system 2, Exported repetitive partial and thiol peroxidase. All these vaccines fulfilled the essential criterias of vaccin candidates assortment including host non-homolohy, virulence, essentiality and conservation. The function annotation and protein-protein interaction analysis showed them to actively involve in significant biologica and molecular processes. These propitious vaccine candidates on epotope mapping generated antigenic 9-mer immunogenic T-cell epitopes.

The study, established upon integrated strategy of pan-genomics and reverse vaccinology revealed potential vaccine candidates against Mycobacterium tuberculosis. The inclusive analysis of all the completely sequenced genomes discovered eight putative antigens which could initiate substantial protective immune response against all Mycobacterium tuberculosis strains .

The antigenic epitopes identified in the vaccine candidates can be utilized for the development of a cost effective mutivalent peptide or recombinant vaccine against Mycobacterium tuberculosis.

CHAPTER 1: INTRODUCTION

Tuberculosis is ranked second in causing deaths from infectious diseases after those due to human immunodeficiency virus and is considered one of the oldest disease (WHO. 2015). In middle and low-income countries, TB in 2016 is still causing mortality and morbidity (WHO. 2015). In 2015, 1.5 million are estimated to have died from active TB disease out of the 9.6 million individuals that suffered from it. Six countries including Pakistan, South Africa, Nigeria, Indonesia, China and India, 60 % of these deaths (WHO. 2015). TB is heterogeously distributed throughout the world, for instance, tuberculosis frequency South Africa is >250-fold higher than in the USA (WH0. 2015). The rate of active TB disease development in infants that are exposed is much higher than in 2-10 years old children. However, in adolescence, the risk of TB rises and then it plateaus around the age of 25 and remains high throughout adult life (Marais et al., 2006). Men and women have different frequency of TB occurrence, it is developed twofold higher in men than in women (Dye et al., 2006). In addition, from all new cases that arises world wide of active TB, 10 % are developed in children (Swaminathan et al., 2010). *Mycobacterium tuberculosis* is the main cause of TB in humans (Pai et al., 2016)

Mycobacterium genus has been hypothesized to originate more than 150 million years ago and may have killed more than any other pathogen (Hayman et al., 1984 & Daniel et al., 2006). *M. tuberculosis* belongs to the Actinobacteria phylum and has a close relation to *Mycobacterium smegmatis* which is a saprophytic bacterium. *M. tuberculosis* is rod-shaped, non-motile and relatively large bacterium (Cole et al., 1998). The rods of this bacilli are 0.2-0.5 um in width and 2-4 micrometer in length (Attorri et al., 2000). *M. tuberculosis* needs oxygen for its survival and this is the reason for its presence, during TB infection, in the areas of lungs which

are well oxygenated (Hui-Zin et al., 2003). It divides every 15-20 hours and the only reservoir known of *M. tuberculosis* is human, for this reason, it is known as facultative intracellular parasite (Pai, et al., 2016 & Attorri et al., 2000).

M. tuberculosis cell wall contains peptidoglycan but have layers of mycolic acid on outside which makes it weekly stained or not stained at all in gram staining (Fu et al., 2002). Three different types of lipids make up the *M. tuberculosis* cell wall's lipid portion includes Wax-D Mycolic acids, cord factor and mycolic acids (Kaur et al., 2009). Mycolic acids molecules have hydrophobicity which has an effect on cell surface permeability as it forms a shell of around the organism (Brennan et al., 2008). This Mycolic acid shell is thought to be one of significant factor which determine virulence in *M. tuberculosis*. It is hypothesized that the survival of in the granuloma is due to this shell (Brennan et al., 2008). Cord factor inhibits migration of polymorphonuclear leukocytes thus is the cause of toxicity in host's cells. Strains *M. tuberculosis* which are highly virulent produces Cord factor in huge amount. The third unique lipid component of *M. tuberculosis's* cell wall, Wax-D, is the main component of Freund's complete adjuvant (Brennan et al., 2008).

Currently, the only licensed vaccine available to avoid TB disease is Bacillus Calmette–Guérin (BCG) and vaccination of more than 90% of infants globally is done with it (Global routine vaccination coverage, 2014. & Zwerling et al., 2011). Practices and policies of BCG are available online at BCG Wolrd Atlas (http://www.bcgatlas.org) (Zwerling et al., 2011). In 1921, BCG vaccine on humans was used for the first time and since then it has been assessed in various observational studies and interventional trials. In adults, the BCG vaccine efficacy in clinical trials in adults against pulmonary TB has been stated to be 0-80 % (Mangtani et

al., 2014 & Roy et al., 2014). This is unclear that why such variability exists in efficacy of BCG vaccine (Mangtani, et al. 2014). The efficacy, however, of the BCG in infants and <5 years of old children have been reported in case-control studies to be 50-80 % (Trunz et al., 2006).

Morbidity and mortality as a result of TB can be high <5 years old childeren, so BCG is invaluable for this age group in preventing active TB disease. However, its efficacy is not certain in adults and adolescent which suffer more from pulmonary and transmissible Tuberculosis disease than infants and childerns (Barreto et al., 2011 & Tuberculosis Research Centre, 1999). Additionally, pediatric BCG vaccine efficacy meta-analysis has shown that there is generally up to 10-year protection period (Abubakar et al., 2013). BCG vaccine in most countries is administered once, at birth, and its efficacy suggest that it is unlikely to extend its protection consistently into adolescence thus it is doubtful that present form of BCG vaccine might significantly contribute in controlling TB epidemic globally (Zwerling et al., 2011). BCG vaccines cannot be administered to children and infants suffering from HIV as the live attenuated bacteria causes other disease in them (Madhukar et al., 2016).

Even though efficacy of BCG is variable but still its induction of protective immunity against TB is proven, however the mechanism of protection is still not well illustrated. TB infection is not contained by 10 % of infected population and preventing disease development in these individuals, is indeed, the current vaccination research main goal. For assessing the ability of a vaccine to reach these goals, trial designs which are normal can be used (Ellis et al., 2015). Adults and adolescents mostly spread *M. tuberculosis* infection with active pulmonary TB disease, therefore most of the newly developing vaccines are designed focusing on these age groups. However, due to partial efficacy of BCG vaccines in newborns and its HIV-exposed

newborns non-recommendation makes it important to for develop a new improved vaccine for infants (Madhukar et al., 2016).

It has been reported that, in first 20 years, 30 million cases of active TB disease would be averted if a vaccine having efficacy of 60 % is administered to only 20 % of adults and adolescents across the globe. If the same vaccine is delivered to 90 % infants than, in total, 35 million active TB disease cases could be averted (AERAS, 2014). Another study concludes that effect of vaccine targeting adult and adolescents on the global TB burden in time period of 2024-2050 is much greater than vaccines which targets infants (Knight et al., 2014).

Study has shown that *M. tuberculosis* T cell epitopes of humans are hyper conserved evolutionary (Inaki et al., 2010). 21 strains of *M. tuberculosis* were taken, sequenced and their comparison showed that *M. tuberculosis* non-essential genes conservation is less than that of essential genes and variation in the sequence is few in the most 491 human T cell epitopes which are experimentally confirmed (Inaki et al., 2010).

To date, only one reverse vaccinology study has been performed on *M. tuberculosis*, in which, proteome of strain *M. tuberculosis H37Rv* was taken and analyzed for promiscuous T cell epitopes through prediction software NERVE (New Enhanced Reverse Vaccinology Environment) (Gloria et al., 2015). However, analysis of core genome can give us vaccine targets that are better which can help in overcoming evasion mechanisms of microorganism (Scarselli et al., 2008). No reverse vaccinology study has been yet performed on core genome of all completely sequenced genome strains of *M. tuberculosis* to obtain T cell epitopes which can be effective for all the strains.

To do analysis of Pan-genome of all available strains of *M. tuberculosis* on NCBI was the first aim of our study. Secondly, we wanted to assess how much alarming the antibiotic

resistance phenomenon is in *M. tuberculosis* thus we analyzed the core genome of the pathogen for antibiotic resistance genes. Thirdly, we wanted to do identify candidate for vaccine which are conserved utilizing approaches of *in silico*.

CHAPTER 2: LITERATURE REVIEW

2.1 Epidemiology

It is estimated that in 2015, 9.6 million individuals suffered from active TB disease and 1.5 million died out of them (WHO, 2015). Six countries including Pakistan, South Africa, Nigeria, Indonesia, China and India accumulated approximately 60 % of these deaths (WHO, 2015). TB is heterogeously distributed throughout the world, for instance, tuberculosis frequency South Africa is >250-fold higher than in the USA (WH0,2015). The rate of active TB disease development in infants that are exposed is much higher than in 2-10 years old children. However, in adolescence, the risk of TB rises and then it plateaus around the age of 25 and remains high throughout the adult life (Marais et al., 2006). Men and women have different frequency of TB occurrence, it is developed twofold higher in men than in women (Dye et al., 2006). In addition, from all new cases that arises world wide of active TB, 10 % are developed in children (Swaminathan et al., 2010).

HIV infection is the strongest of the major known factors for TB (Havlir et al., 2008). Of all new active TB disease cases, 12 % occurs in HIV-positive people and 25 % of TB related deaths occurs in them too. Out of these 25 % deaths, majority (75 %) occurs in Africa (Getahun et al., 2015). Apart from HIV-positive individuals, other risk factors that are accountable for the remaining fraction of TB cases includes malnutrition, air pollution (Lonnroth et al., 2010), excessive alcohol use (Rehm et al., 2009), type 2 diabetes mellitus (Jeon et al., 2008) and

smoking (Bates et al.,2007). In order to control TB, these behavioral and social factors need to be addressed (Lonnroth et al., 2010).

Drug resistance in *M. tuberculosis* has emerged and it is distributed hetero generously throughout the world and hence has become a major concern in controlling the infection (Madhukar et al., 2016). It is estimated that the prevalence of MDR-TB is at 5 % but it varies from >20% in countries of former Soviet Union to 1 % in areas of western Europe, sub-Saharan Africa, North America and Africa (WHO. 2014). In recent years, one of the major concern is the development of drug-resistant TB in china, where there is resistance to either rifampicin or isoniazid in 1/4th of the active TB, and India, where totally-resistant strains has emerged (Zhao et al., 2014). Throughout the world in 2015, 480 000 people, in total, have been estimated to develop TB with multidrug-resistance (MDR-TB) (WHO. 2016).

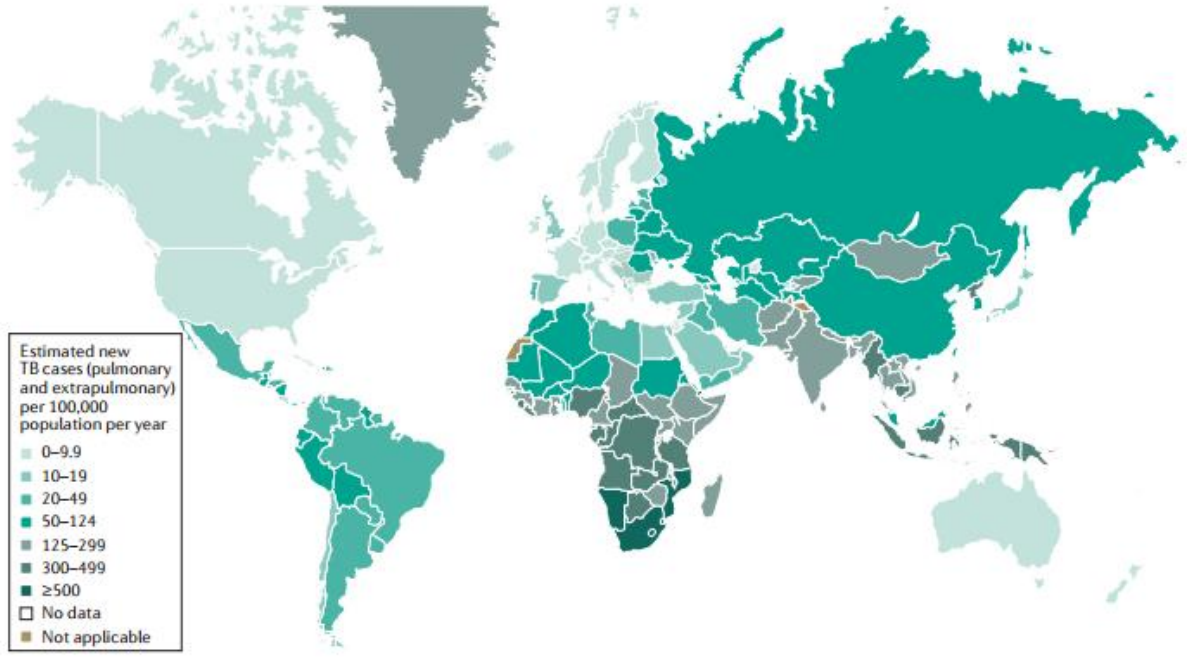


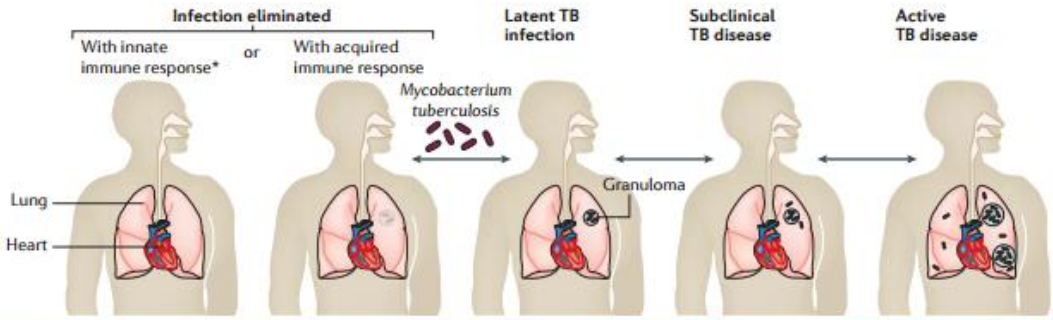
Figure 1| Active TB disease (extra pulmonary and pulmonary) global incidence.
Active tuberculosis (TB) disease lowest rates are observed in countries with High-income — New Zealand, Australia, USA, Canada and most countries in western Europe— typically less than 10 cases per 100,000 populations per year. On contrary, higher rates of TB are found in countries with lower-income. (WHO. 2015)

2.2 Tuberculosis

Robert Koch, in 1882, discovered for the first time the agent which causes an airborne infectious tuberculosis (TB) disease (WHO. 2015). In middle and low-income countries, Tuberculosis in 2016 is still causing of mortality and morbidity (WHO. 2015). *M. tuberculosis* is mainly a pulmonary pathogen but it doesn't cause disease only in lungs. Additionally, its spectrum is dynamic from infection, which is asymptomatic, to disease, which can be life-threatening (Barry et al., 2009 & Esmail et al., 2014) (FIG 1). From perspective of public and clinical health, the disease non-transmissible and asymptomatic state of the disease is classified in the latent Tuberculosis infection (LTBI) and others which have transmissible state of the infection are classified as having the active TB disease. Active TB disease general symptoms includes are fever, weight loss, lack of appetite, fatigue, fever and persistent cough and people with advanced stage of the disease can experience hemoptysis, which is coughing with blood. In some exceptional cases, a patient with active TB disease (culture-positive) experience no symptoms and they are described to have subclinical TB2 (Esmail et al., 2014).

TB standard treatment is comprised of four first-line antibiotics which includes and pyrazinamide, ethambutol, rifampicin and isoniazid. Multidrug-resistant TB (MDR-TB) exists, which is defined to be TB having resistance to at least rifampicin and isoniazid, and drug resistance to all of the four antibiotics can occur. MDR-TB is virtually reported in all countries (WHO, 2015). The TB which is not only resistant to rifampicin and isoniazid but also to second-line three injectable amino glycosides are called as Extensively drug-resistant TB disease (WHO,

2015). Drug resistant, drug-sensitive, LTBI and active TB have different tests for diagnosis and their treatment options varies from one another (Madhukar et al., 2016).



TST	Negative	Positive	Positive	Positive	Usually positive
IGRA	Negative	Positive	Positive	Positive	Usually positive
Culture	Negative	Negative	Negative	Intermittently positive	Positive
Sputum smear	Negative	Negative	Negative	Usually negative	Positive or negative
Infectious	No	No	No	Sporadically	Yes
Symptoms	None	None	None	Mild or none	Mild to severe
Preferred treatment	None	None	Preventive therapy	Multidrug therapy	Multidrug therapy

Figure 2| TB Spectrum— from infection of *Mycobacterium tuberculosis* to active TB disease.
Even though dynamic of disease of tuberculosis (TB) can be viewed as continuum from infection of *M. tuberculosis* to active infectious disease, for simplicity, in public health and clinical settings categorization of patients are done as either having latent TB infection (LTBI) or active disease of TB. Depending on changes in comorbidities and immunity of host, the condition of TB can advance or reverse. After exposure, elimination of *M. tuberculosis* can occur either through responses of innate immunity or acquired T cell immune responses. Those individuals who got rid of the infection through responses of innate or acquired immunity without priming of T cell or memory cells (denoted by *) can have negative results of interferon- γ release assay (IGRA) or tuberculin skin test (TST). Whereas, the IGRA or the TST results would be positive in those individuals who retained a strong memory T cell response after eliminating the pathogen. LTBI treatment for such individuals is not beneficial. If, however, the elimination of pathogen doesn't occur, the bacteria remain in a quiescent or latent state which can be detected as positive results for IGRA and TST. For such patients, recommended preventive LTBI therapy regimens would be beneficial (isoniazid mostly for 6-9 months). Symptoms might not be reported in subclinical TB patients, but culture test for them will be positive. Symptoms for patients experiencing active TB disease includes fever, cough and weight loss, and molecular tests, culture and sputum smear can usually confirm its diagnosis. Immune suppression caused by active TB disease or some other underlying disease such as AIDS may cause negative results of the IGRA and the TST in patients suffering from active TB disease. Recommended treatment should be given to those patients which are suffering from subclinical or active TB disease, that includes four drugs intensive phase, followed by a longer two drugs continuation phase (Madhukar et al., 2016)

2.3 Mechanisms/pathophysiology

2.3.1 Microbiology

Constant transmission of infection of *M. tuberculosis* (Firdessa et al. 2013) and reactivation of LTB1 (Reed et al., 2009), globally, are responsible for the disease of TB. Most of the TB cases are due to *M. tuberculosis* or *Mycobacterium africanum* which closely related organism to *M. tuberculosis*. *M. tuberculosis* complex zoonotic members such as *Mycobacterium caprae* or *Mycobacterium bovis* are responsible for minority of cases of TB disease (Bos et al., 2014). The only known reservoir of *M. tuberculosis* is human so it acts both as a symbiont and a pathogen which makes it necessary to understand its host-pathogen interactions (Comas et al., 2013)

2.3.2 Host–pathogen interactions

There has been substantial variability in genetics among isolates obtained throughout the world which suggests accumulation of genetic drift associated with migration pattern of humans or different lineages varied pathogenicity (Warner et al., 2015). On the basis of studies related to epidemiology, hyper virulent strains have been proposed and they are suggested to exists (Reed et al., 2004). If this is found to be true, then such strains genomic studies can give virulence

factors that are lineage-specific (Reed et al., 2004). These factors then can be used for decisions related to infection control and patient care. Although, several characteristics of *M. tuberculosis* drug resistance, increase transmissibility and mortality (Warner et al., 2015) have been found to be strains specific but the results of such studies are inconsistent and hence are not immediately translated into clinical care (Madhukar et al., 2016). Additionally, the *M. tuberculosis* and host interactions are complex (Madhukar et al., 2016). Therefore, in the absence of host susceptibility determinants, studying virulence factors of *M. tuberculosis* can make synergistic interactions unclear (Madhukar et al., 2016. For example, a specific host- pathogen of East-Asian lineage may explain why they are highly pathogenic and infective in Asian population (Gagneux et al., 2006) but the same host-pathogen interaction may have normal epidemiological and clinical presentation in Switzerland (Fenner et al., 2012) or Canada (Albanna et al., 2011). On the other hand, strains that are unremarkable in East-Asia can cause an outbreak in other epidemiological and social settings (Lee et al., 2015).

2.3.3 LTBI

There are two board outcomes when human body is exposed to *M. tuberculosis* i.e. either it is eliminated or the pathogen persists. In case of elimination, innate immune or adaptive immune response is responsible. For innate immune response, interferon-γ (IFNγ) release assays (IGRAs) or tuberculin skin tests (TSTs) might be negative. In adaptive immune response, IGRAs and TSTs might be negative or positive depending on priming of memory T cell responses (Barry et al., 2009 & Esmail et al., 2014) FIG 1. LTBI therapy would not be beneficial for such individuals irrespective of which process of elimination has undergone (Madhukar et al., 2016). It has been recognized that half of the TST results are found to be negative in individuals having

close contact with patients of TB (Morrison et al., 2008). Studies have shown that some individual have negative TST results despite regular exposure to *M. tuberculosis* which suggest there exists genetic predisposition which is the reason for natural resistant to TB in some people (Cobat et al., 2009).

However, if infection of *M. tuberculosis* is not eliminated, persistence of pathogen in a quiescent or latent state can occur and, typically, positive IGRA and TST results will develop with no symptoms. LTBI therapy would probably be beneficial for such individual. Unfortunately, a positive IGRA or TST results are not automatic implication to LTBI as those individuals who has successfully eliminated the infection might still be IGRA or TST positive because of the response of memory T cell (Barry et al., 2009 & Esmail et al., 2014). The low prognostic value of IGRAs and TSTs are partly explained by this finding (Rangaka et al., 2012).

2.3.4 Immunology

Little is known when it comes to early phases of the infection of *M. tuberculosis* in humans. However, studies in non-human primates and small mammal (such as rabbits, guinea pigs and mice) have considerably helped in identification of important early events during primary infection (Orme et al., 2015). Respiratory tract is the entry route of *M. tuberculosis* after it is inhaled into body. *M. tuberculosis* after inhalation reaches to lower respiratory tract where it come across alveolar macrophages, which is the major cell type infected in *M. tuberculosis* infections (FIG. 3). Through receptor-mediated phagocytosis, internalization of bacterial cells is done. This process has been thoroughly studied without considering the alveolus microenvironment. The fluid lining of the epithelium is abundant with surfactants and important role of it in initial host-pathogen interaction is suggested (Watford et al., 2001). For example,

internalization by alveolar macrophages of *M. tuberculosis* can be prevented by surfactant protein D (Ferguson et al., 1999). Once *M. tuberculosis* is phagocytosed, fusion of phagosome with lysosome is actively blocked by it to ensure its survival (Russell et al., 2011). After that, phagosomal membrane is disrupted by *M. tuberculosis* through ESX-1 secretion system activity to release mycobacterial DNA and other bacterial products into the cytosol of macrophage; cytosol might also contain few bacteria in the ensuing days (van et al., 2007 & Houben et al., 2012). Active studies are in progress to understand the advantages of the delivery of bacterial products into the cytosol (Simeone et al., 2016 & Russell et al., 2016). It is suggested that cytosolic surveillance pathway is activated by it, resulting in type I IFN response induction which promotes intracellular bacterial pathogens growth and *M. tuberculosis* is one of them (Manca et al., 2009, Mayer-Barber et al., 2014, Stanley et al., 2007, Pandey et al., 2009 & Manzanillo et al., 2012). Additionally, experimental studies have suggested that cell death type (necrosis versus apoptosis) that macrophage which is infected experiences is crucial for responses of both adaptive and innate immunity towards infection (Kaufmann et al., 2016, Schaible et al., 2013 & Behar et al., 2010). On response to *M. tuberculosis* infection, macrophage derived from bone marrow are recruited in lung and studies should be performed to investigate the importance of bone marrow-derived macrophages versus residential alveolar macrophages. *M. tuberculosis* enters into lung inter stitium, after infecting alveolar macrophages, where infection process evolves (Madhukar et al., 2016). However, the mechanism of accesses to the parenchyma by *M. tuberculosis* is unknown. Two possible mechanisms might be involved though: one involves the direct infection of epithelial cells by *M. tuberculosis* and the second involves the transmigration across the epithelium of infected macrophages (Madhukar et al., 2016) (FIG. 3). Irrespective of the route, parenchyma is accessed by *M. tuberculosis*, which is

followed by recruitment of large quantity of cells to infection site to generate response of host towards infection consisting of multiple cells called a granuloma. After establishment of primary infection, either inflammatory monocytes (Samstein et al., 2013) or infected dendritic cells (Wolf et al., 2009) transport *M. tuberculosis* for T cell priming to pulmonary lymph nodes. Active delay of T cell trafficking in to lungs as well as initial T cell priming has been shown by *M. tuberculosis* (Wolf et al., 2009 & Chackerian et al., 2002). The amount of T cells (CD4+) are reduced in HIV infection and it is, because of this, a risk factor of *M. tuberculosis* infection to active TB disease. Although, some studies show that at the early stage of HIV infection, during which CD4+ T cell number in normal, there is enhanced risk of active TB disease suggesting impairment of other T cell-independent immune response (Sonnenberg et al., 2005).

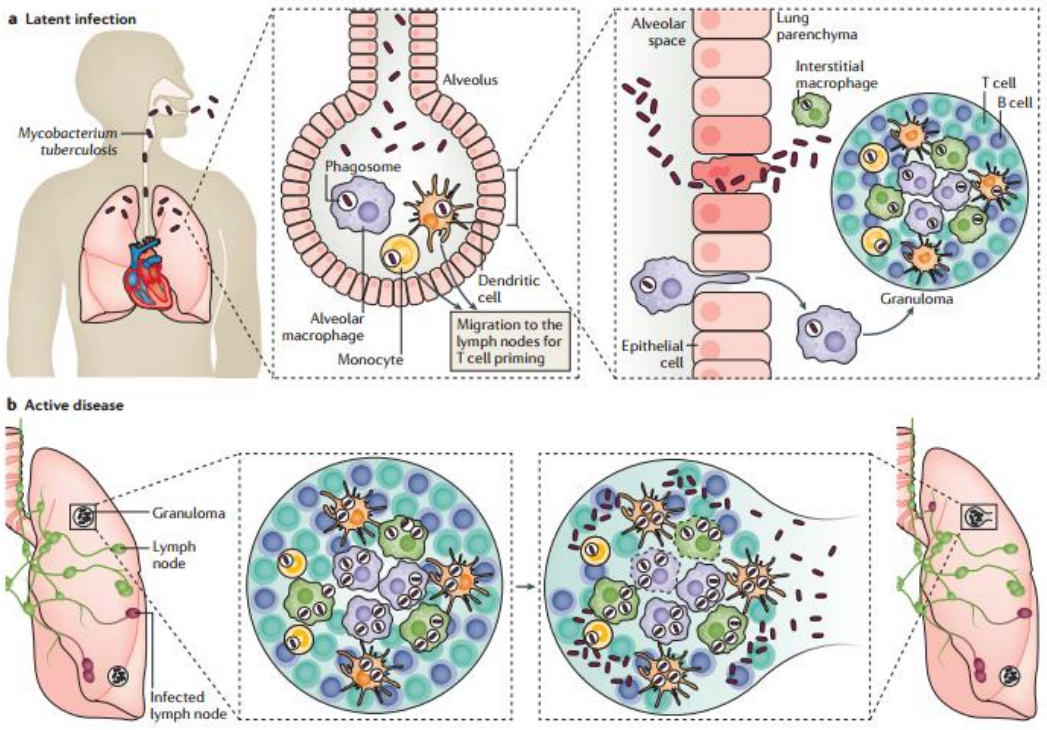


Figure 3| Infection of *Mycobacterium tuberculosis*.
a | Entry of M. tuberculosis into the lungs through inhalation is the starting point of the infection. Next checkpoint in the lungs is its entry into alveolar space where it come across resident alveolar macrophages. If the bacteria are not eliminated by the first line of defense, the intestinal tissue is invaded by the *M. tuberculosis*, either by direct alveolar epithelium infection or by the migration of infected alveolar macrophages to the parenchyma of lung. Subsequently, transportation of *M. tuberculosis* is done by either inflammatory monocytes or dendritic cells to T cell residing in the pulmonary lymph nodes for priming of T cell. This is followed by immune cells recruitment, including B cells and T cells, to the parenchyma of lung leading to the formation of granuloma. **b |** The bacteria increase its number within the growing granuloma. Failure in containing infection by the granuloma will occur if the bacterial number becomes too great and eventually spreading of bacteria to other organs will occur, including the brain. At this stage, entry of bacteria to the bloodstream or re-entry to the respiratory tract can occur — the infected host is suffering from active TB disease (Madhukar et al., 2016)

2.4.5 The granuloma

The granuloma demonstrates *M. tuberculosis* infection duality: from the perspective of the host, the granuloma is a 'prison' of the bacteria which potentially 'wall off' rest of the body from infection; however, from the perspective of the bacteria, it is a 'room' to grow and replicate in. For example, type I IFN response can be initiated by ESX-1 secretion system of *M. tuberculosis* which has been linked directly to recruitment of a unique myeloid population

(CD11b+F4/80+Gr1int) in nascent granuloma which has high permissiveness to *M. tuberculosis* infection (Antonelli et al., 2010). Interestingly, immune responses have been demonstrated in a study to segregate geographically around the granuloma, with the center of it containing components which are pro-inflammatory, while the neighboring tissue has anti-inflammatory components (Marakalala et al., 2016). It has also been suggested that granuloma might have a maximal carrying capacity and infection would only continue to progress after surpassing it (Lin et al., 2014). If the infection resides in granuloma and it hasn't induced significant pathology in tissue, then the person is said to have LTBI and is a candidate for preventive treatment (Madhukar et al., 2016).

2.3.5 Progression to active TB disease

Combination of T cells, dendritic cells and macrophages is enough for maintaining the infection in control and asymptomatic in majority of patients with LTBI. However, for some unclear reasons, the infection can move to clinical disease, in some hosts, in weeks and in others it may take decades (Madhukar et al., 2016).

Experiments have shown that tumor necrosis factor (TNF) might have a role in maintaining control of the *M. tuberculosis* infection (Tobin et al., 2012). In early 1990s, an experiment demonstrated that mice receiving anti-TNF treatments showed an increased active TB disease risk. Although, further experiments showed that mechanisms of TNF are complex as over activation of it makes the condition even more worse and less activation doesn't have the ability to contain the infection (Tobin et al., 2012 & Lalvani et al., 2012).

Birth defects in immune system can give us information of immune response mechanisms to TB (Bustamante et al., 2014). Every year vaccination of over 100 million infants are done with BCG and a portion of them develop dispersed BCG disease; thus, mapping mutations in genes which encodes proteins which are crucial for containment of Mycobacterium has been possible. Large number of such proteins involvement in IL-12-IFN γ axis has been found (Madhukar et al., 2016). Even though these mutations were identified originally in patients who have disease due to BCG vaccine or mycobacterium which is non-tuberculous, in some cases, linking of the identified mutations to active disease has also been found (Bustamante et al., 2014). Additionally, experimental TB in animal models have been linked to several other genes which were linked subsequently to human genetic studies (Abel et al., 2014). In conclusion, a likely explanation to why LTBI in some people progress to active TB disease can be given by genetic susceptibility; however, immunological pathways that have major role in controlling the mycobacterial infection needs to be unraveled thus requires additional investigation (Abel et al., 2014).

2.4. 6 Drug resistance Mechanisms

In 1948, drug resistance was for the first time associated with TB and it was discovered during first human trial for therapy of TB (Daniels et al., 1952). With every introduction of novel TB drug into clinical practice, new strains with resistance to the drug has emerged, typically within ten years (Madhukar et al., 2016). The process through which *M. tuberculosis* acquire resistance is genetic mutation. Target modification in genes are the two main mechanisms of resistance to drugs; for example, mutation in RNA polymerase of *M. tuberculosis*

leads to resistance to rifampicin as the drug no longer can act on RNA polymerase or a mutation in the enzyme which normally activates pro-drug into active state and this is described to happen in case of resistance to isoniazid (Nebenzahl et al., 2014).

2.4 Diagnosis

Different diagnostic tool exists for TB and selection of which one to use from them depends on whether the purpose is to detect LTBI or active TB or drug resistance (Madukar et al., 2016).

2.5.1 Diagnosis of LTBI

LTBI can be identified using two tests: TST and IGRA. The IGRA can also be used to distinguish between *M. tuberculosis* infection-induced and BCG-induced positive TST responses (Pai et al., 2014). Mantoux technique is used to perform TST in which 5 tuberculin units (5 TU) of purified protein derivative (PPD) S or 2 TU of PPD RT are intradermal injected (WHO,2014). Delayed-type hypersensitivity reaction will occur in people who have cell-mediated immunity to these antigens within 48-72 hours. Pre-test probability of the infection, the size of induration and the active TB disease developing risk are all taken into consideration in the interpretation of the TST (Menzies et al., 2008). Even though TST has several advantages including limited laboratory & skill requirements, low equipment & reagent costs, it has couple of limitation, firstly, late or repeated BCG vaccination (booster) and exposure to mycobacteria which non-tuberculous compromises its specificity (Farhat et al., 2006). Secondly, its predictive value

which is limited (Pai et al., 2014). Efforts are being made currently to replace PPD based TST with new skin tests (Pai et al., 2016).

IGRAs were introduced in early 2000s with the hope to substitute TSTs (Pai et al., 2004). IGRAs, an in vitro blood test, measure the release of IFNγ from T cell following RD1-encoded antigens stimulation (Abdallah et al., 2007 & Sorensen, A. L. et al., 1995). Specificity of RD1 antigens to *M. tuberculosis* is more than PPD as BCG vaccine strain or most of other non-tuberculosis mycobacteria does not encode them (Andersen et al., 2000). IGRAs do give solution to first limitation of TSTs but IGRAs predictive value is poor just like TSTs (Pai et al., 2014 & Rangaka et al., 2012). Research studies had made this clear that both the IGRA and the TST are imperfect but acceptable for LTBI (Pai et al., 2014 & Pai et al., 2016).

2.5.2 Diagnosis of Active TB disease

Four major technologies are available for active TB disease detection which includes microscopy (sputum smears), imaging techniques (PET-CT and chest X-rays), molecular tests and culture-base methods (Madhukar et al., 2016). Imaging tests are mainly used for screening, microbiological diagnosis are required for active TB. Technologies of Diagnostic that WHO have reviewed and recommended are overviewed in TABLE 1.

2.5.3 Diagnosis of Drug resistance TB

Phenotypic, culture based and molecular- based tests are used for the drug resistance detection of *M. tuberculosis* (Madhukar et al., 2016). In cultural based tests, bacteria are

checked for its ability to grow in media in which anti-TB drugs are added. In molecular based tests, genetic mutations which confers resistance to *M. tuberculosis* are detected (Madukar et al. 2016). Both of these methods are been overviewed in TABLE 1.

Table 1| Various diagnostic techniques for diagnosing TB.

Test	Assay principle	Use	Sensitivity (%)	Specificity (%)	TAT*	Target setting‡	Year endorsed	Refs
Imaging techniques								
Chest X-ray	Imaging of the lungs	Active TB disease screening	87 (using TB abnormality as a threshold)	89 (using TB abnormality as a threshold)	Same day	Secondary and tertiary centres	Included in the WHO guidelines for many years	217
Microscopy								
Conventional sputum smear microscopy	Direct visualization of mycobacteria using light microscopy	Active TB disease diagnosis	32–94	50–99	Same day	Peripheral and reference laboratories	Included in the WHO guidelines for many years	218
LED fluorescence smear microscopy§	Direct visualization of mycobacteria using fluorescence microscopy	Active TB disease diagnosis	52–97	94–100	Same day	Peripheral and reference laboratories	2011	218
Culture-based techniques								
Liquid culture with DST	Mycobacterial culture on liquid media	• Active TB disease diagnosis • Drug resistance	• 89 (among smear-positive and culture-positive) • 73 (among smear-negative and culture-positive)	>99	10–21 days	Reference laboratory	2007	219
Antigen detection techniques								
LAM lateral flow assay§	Antigen detection	Active TB disease diagnosis in HIV-positive individuals	• 44 (all) • 54 (in HIV-positive individuals)	• 92 (all) • 90 (in HIV-positive individuals)	Same day	Peripheral laboratory	2015 (conditional recommendations in selected groups)	112
Molecular techniques (nucleic acid amplification tests)								
Xpert MTB/RIF§‖	NAAT (qPCR)	• Active TB disease diagnosis • Drug resistance (rifampicin)	• 98 (smear-positive and culture-positive) • 67 (smear-negative and culture-positive) • 95 (rifampicin resistance)	• 99 (smear-negative and culture-negative) • 98 (rifampicin resistance)	Same day	District or sub-district laboratory	2010	105
First-line LPA (GenoType MTBDRplus¶ and NIPRO#)	NAAT (LPA)	• Active TB disease diagnosis • Drug resistance (isoniazid and rifampicin)	• 98 (rifampicin resistance) • 84 (isoniazid resistance)	• 99 (rifampicin resistance) • >99 (isoniazid resistance)	1–2 days	Reference laboratory	2008	220
Second-line LPA (GenoType MTBDRsl¶)	NAAT (LPA)	Drug resistance (fluoroquinolones and second-line injectable drugs)	• 86 (fluoroquinolone resistance) • 87 (second-line injectable drugs)	• 98 (fluoroquinolone resistace) • 99 (second-line injectable drugs)	1–2 days	Reference laboratory	2016	121
Loopamp *Mycobacterium tuberculosis* complex assay§**	NAAT (LAMP)	Active TB disease diagnosis	76–80	97–98	Same day	Peripheral laboratory	2016	120

DST, drug susceptibility testing; LAM, lipoarabinomannan; LAMP, loop-mediated isothermal amplification; LED, light-emitting diode; LPA, line probe assay; NAAT, nucleic acid amplification test; qPCR: quantitative PCR; TAT, turnaround time; TB, tuberculosis. *May require longer TAT owing to batching of specimens. ‡Peripheral laboratories (basic microscopy centres) are typically located at the primary-care level. District-level laboratories are the next level of referral and have better infrastructure. The tertiary hospital or reference laboratory that offers the most sophisticated infrastructure are the highest and final level of referral. §Amenable to rapid 'test and treat'. ‖Newer versions of GeneXpert (Cepheid Inc., Sunnyvale, California, USA) instrument (OMNI) and cartridge (Xpert Ultra MTB/RIF) are currently under development and yet to be reviewed by the WHO. ¶Hain Lifescience GmbH, Nehren, Germany. #NIPRO Corporation, Osaka, Japan. **Eiken Chemical, Tokyo, Japan.

2.6 Conventional Vaccinology for *M. tuberculosis*

Currently, the only licensed vaccine available to avoid TB disease is Bacillus Calmette–Guérin (BCG) and vaccination of more than 90% of infants globally is done with it (Global routine vaccination coverage, 2014. & Zwerling et al., 2011). Practices and policies of BCG are available online at BCG Wolrd Atlas (http://www.bcgatlas.org) (Zwerling et al., 2011). In 1921, BCG vaccine on humans was used for the first time and since then it has been assessed in various observational studies and interventional trials. In adults, the BCG vaccine efficacy in clinical trials in adults against pulmonary TB has been stated to be 0-80 % (Mangtani et al., 2014 & Roy et al., 2014). This is unclear that why such variability exists in efficacy of BCG vaccine (Mangtani et al. 2014). The efficacy, however, of the BCG in infants and <5 years of old children have been reported in case-control studies to be 50-80 % (Trunz et al., 2006).

Morbidity and mortality as a result of TB can be high <5 years old children, so BCG is invaluable for this age group in preventing active TB disease. However, its efficacy is not certain in adults and adolescent which suffer more from pulmonary and transmissible Tuberculosis disease than infants and childerns (Barreto et al., 2011 & Tuberculosis Research Centre, 1999). Additionally, pediatric BCG vaccine efficacy meta-analysis has shown that there is generally up to 10-year protection period (Abubakar et al., 2013). BCG vaccine in most countries is administered once, at birth, and its efficacy suggest that it is unlikely to extend its protection consistently into adolescence thus it is doubtful that present form of BCG vaccine might significantly contribute in controlling TB epidemic globally (Zwerling et al., 2011). BCG vaccines cannot be administered to children and infants suffering from HIV as the live attenuated bacteria causes other disease in them (Madhukar et al., 2016).

2.7 Rationale for study

Even though efficacy of BCG is variable but still its induction of protective immunity against TB has proven, regardless the mechanism of protection is not well illustrated. Tb infection is not contained by 10 % of infected population and preventing disease development in these individuals, is indeed, the current vaccination research main goal. For assessing the ability of a vaccine to reach these goals, trial designs which are normal can be used (Ellis et al., 2015). Most population are at risk from TB disease which is active and transmissible thus it must be prevented to maximize vaccination efficacy on mortality and morbidity. Adults and adolescents mostly spread *M. tuberculosis* infection with active pulmonary TB disease, therefore most of the newly developing vaccines are designed focusing on these age groups. However, due to partial efficacy of BCG vaccines in newborns and its HIV-exposed newborns non-recommendation makes it desirable for development of new improved vaccine for infants (Madhukar et al., 2016). Modeling has illustrated that, in first 20 years, 30 million cases of active TB disease would be averted if a vaccine having efficacy of 60 % is administered to only 20 % of adults and adolescents and if also delivered to 90 % infants than, in total, 35 million active TB disease cases could be averted (AERAS, 2014). Another modeling study concludes that effect of vaccine targeting adult and adolescents on the global TB burden in time period of 2024-2050 is much greater than vaccines which targets infants (Knight et al., 2014).

2.8 Reverse vaccinology

Study has shown that *M. tuberculosis* human T cell epitopes are hyper conserved evolutionary (Inaki et al., 2010). 21 strains of *M. tuberculosis* were taken, sequenced and their comparison showed that *M. tuberculosis* non-essential genes conservation is less than that of essential genes and variation in the sequence is few in the most 491 human T cell epitopes which are experimentally confirmed (Inaki et al., 2010).

To date, only one reverse vaccinology study has been performed on *M. tuberculosis*, in which, proteome of strain *M. tuberculosis H37Rv* was taken and analyzed for promiscuous T cell epitopes through prediction software NERVE (New Enhanced Reverse Vaccinology Environment) (Gloria et al., 2015). Analysis of core genome provide us tools that are new and give us vaccine targets that are better which can help in overcoming evasion mechanisms of microorganism (Moriel et al., 2008). No reverse vaccinology study has been yet performed on core genome of all completely sequenced genome strains of *M. tuberculosis* to obtain T cell epitopes which can be effective for all the strains of *M. tuberculosis*.

To do analysis of Pan-genome of all available strains of *M. tuberculosis* on NCBI was the first aim of our study. Secondly, we wanted to assess how much alarming the antibiotic resistance phenomenon is in *M. tuberculosis* thus we analyzed the core genome of the pathogen for antibiotic resistance genes. Thirdly, we wanted to do identify candidate for vaccine which are conserved utilizing approaches of *in silico*.

The pipeline followed for identification of conserved vaccine proteins for prioritized vaccine candidates was of Hassan et al., whom used it for identification of vaccine candidate for *Acinetobacter baumannii* (Hassan et al., 2016). The flow chart of the pipeline is shown in Fig 4.

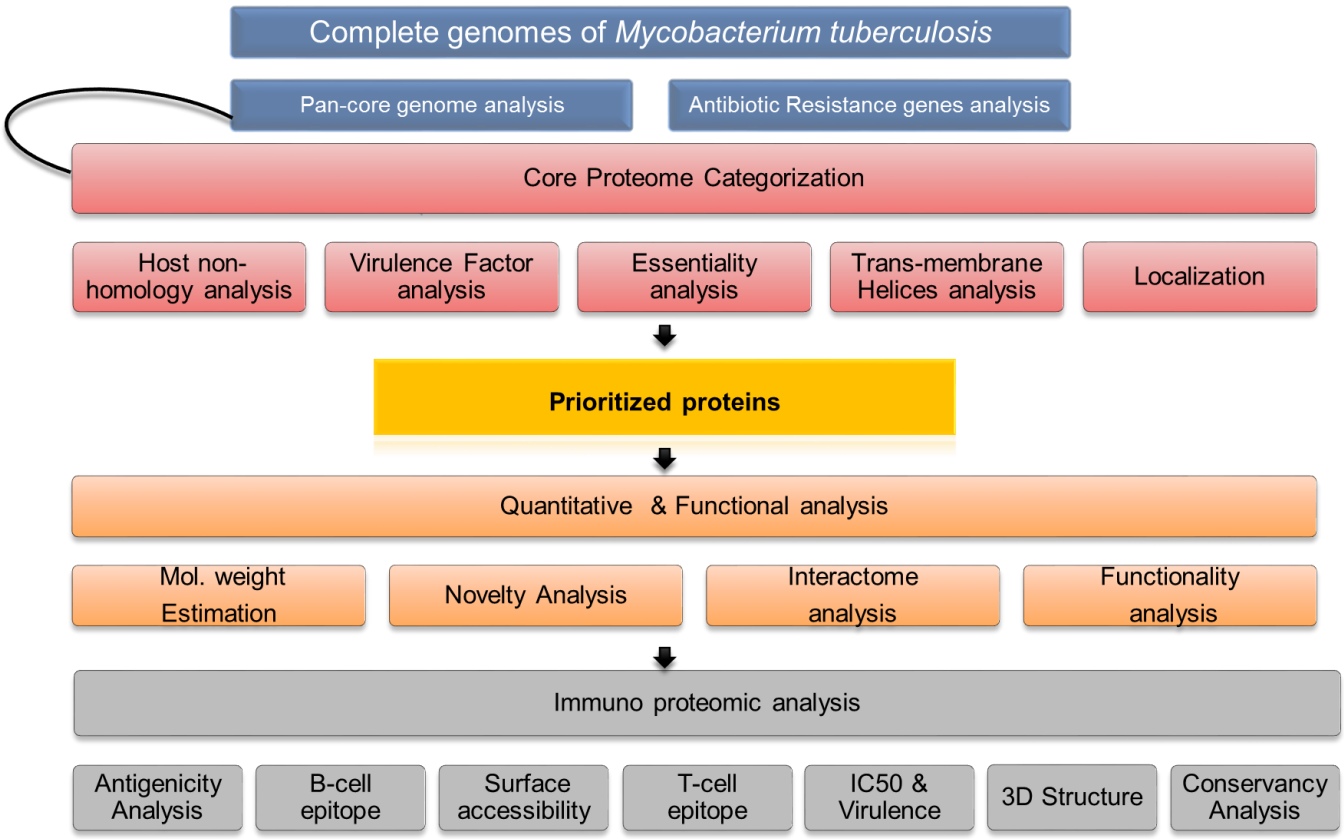


Figure 4: Flow chart of pipeline used.

CHAPTER THREE: METHODOLOGY

3.1 Genome selection and gene prediction

For a comprehensive analysis of the species pan genome, all available completely sequenced genomes of *Mycobacterium tuberculosis* strains (47 at the time of writing this manuscript, May 2016) were included in the study. The data is retrieved from NCBI Genbank (http://www.ncbi.nlm.nih.gov/genome). Chromosomal DNA sequences were retrieved in FASTA format for each genome. For prediction of genes and proteins Prodigal (Prokaryotic dynamic programming gene finding algorithm) was used for its increased specificity and sensitivity in prediction of gene with reduced false positive predictions and improved translation initiation site recognition (Hyatt et al. 2010). The genomic data of all *M. tuberculosis* strains, including genome size, number of genes, GC % and genome accession number as given by NCBI is shown in TABLE 2.

Table 2| Genome Statistics of all genomes of *Mycobacterium tuberculosis*.

Organism Name	Replicon	Size(MB)	GC%	Genes	Proteins	Pan genes	Core genes	Accessory genes	Unique genes
Mycobacterium tuberculosis CCDC5079	chromosome: CP001641.1	4.39881	65.6	3696	3647	3657	3657	474	12
Mycobacterium tuberculosis RGTB327	chromosome: CP003233.1	4.38012	65.6	3739	3691	3929	3546	534	113
Mycobacterium tuberculosis RGTB423	chromosome: CP003234.1	4.40659	65.6	3670	3622	4150	3484	539	136
Mycobacterium tuberculosis str. Haarlem/NITR202	chromosome: CP004886.1	4.40479	65.6	3730	3681	4514	3374	508	281
Mycobacterium tuberculosis CAS/NITR204	chromosome: CP005386.1	4.39288	65.6	4008	3960	4746	3307	539	196
Mycobacterium tuberculosis F1	chromosome: CP010329.1	4.42862	65.6	4366	4318	4802	3296	529	19
Mycobacterium tuberculosis 2242	chromosome: CP010335.1	4.41984	65.6	4512	4464	4853	3285	551	17
Mycobacterium tuberculosis 2279	chromosome: CP010336.1	4.40503	65.6	4647	4599	4910	3272	544	49
Mycobacterium tuberculosis Beijing-like	chromosome: CP010873.1	4.41122	65.6	4170	3829	4926	3265	510	10
Mycobacterium tuberculosis H37Rv	chromosome: NC_000962.3/AL123456.3	4.41153	65.6	4008	3906	4928	3261	470	0
Mycobacterium tuberculosis CDC1551	chromosome: NC_002755.2/AE000516.2	4.40384	65.6	4113	3964	4938	3251	463	6
Mycobacterium tuberculosis H37Ra; ATCC 25177	chromosome: NC_009525.1/CP000611.1	4.41998	65.6	4153	4069	4942	3249	472	2
Mycobacterium tuberculosis F11	chromosome: NC_009565.1/CP000717.1	4.42443	65.6	4139	4043	4943	3246	466	0
Mycobacterium tuberculosis KZN 1435	chromosome: NC_012943.1/CP001658.1	4.39825	65.6	4118	4014	4948	3245	463	1
Mycobacterium tuberculosis KZN 4207	chromosome: NC_016768.1/CP001662.1	4.39499	65.6	4115	4028	4948	3245	462	0
Mycobacterium tuberculosis CCDC5180	chromosome: NC_017522.1/CP001642.1	4.40598	65.6	4131	3982	4953	3244	467	4
Mycobacterium tuberculosis CTRI-2	chromosome: NC_017524.1/CP002992.1	4.39853	65.6	4122	4028	4955	3244	460	2
Mycobacterium tuberculosis KZN 605	chromosome: NC_018078.1/CP001976.1	4.39912	65.6	4122	4016	4956	3244	461	1
Mycobacterium tuberculosis H37Rv	chromosome: NC_018143.2/CP003248.2	4.41171	65.6	4132	4058	4956	3244	468	0
Mycobacterium tuberculosis 7199-99	chromosome: NC_020089.1/HE663067.1	4.4212	65.6	4121	4026	4956	3238	470	0
Mycobacterium tuberculosis str. Erdman (ATCC35801)	chromosome: NC_020559.1/AP012340.1	4.39235	65.6	4129	4010	4960	3235	475	3
Mycobacterium tuberculosis str. Beijing/NITR203	chromosome: NC_021054.1/CP005082.1	4.41113	65.6	4141	3937	4967	3235	490	4
Mycobacterium tuberculosis EAI5/NITR206	chromosome: NC_021194.1/CP005387.1	4.39031	65.6	4105	3869	4983	3230	478	14
Mycobacterium tuberculosis CCDC5079	chromosome: NC_021251.1/CP002884.1	4.41432	65.6	4136	4033	4985	3230	464	2
Mycobacterium tuberculosis EAI5	chromosome: NC_021740.1/CP006578.1	4.39117	65.6	4100	4002	4986	3230	468	0

Mycobacterium tuberculosis str. Haarlem	chromosome: NC_022350.1/CP001664.1	4.40822	65.6	4112	4015	4986	3230	462	0
Mycobacterium tuberculosis str. Kurono	chromosome: NZ_AP014573.1/AP014573.1	4.41508	65.6	4139	4054	4986	3230	471	0
Mycobacterium tuberculosis HKBS1	chromosome: NZ_CP002871.1/CP002871.1	4.40793	65.6	4126	4023	4989	3228	462	3
Mycobacterium tuberculosis BT2	chromosome: NZ_CP002882.1/CP002882.1	4.4019	65.6	4122	4019	4992	3222	444	3
Mycobacterium tuberculosis BT1	chromosome: NZ_CP002883.1/CP002883.1	4.3994	65.6	4125	4009	4993	3219	456	0
Mycobacterium tuberculosis CCDC5180	chromosome: NZ_CP002885.1/CP002885.1	4.41435	65.6	4134	4032	4995	3218	459	0
Mycobacterium tuberculosis H37RvSiena	chromosome: NZ_CP007027.1/CP007027.1	4.41091	65.6	4133	4056	4995	3218	467	0
Mycobacterium tuberculosis K	chromosome: NZ_CP007803.1/CP007803.1	4.38552	65.6	4102	3991	4997	3215	460	1
Mycobacterium tuberculosis KIT87190	chromosome: NZ_CP007809.1/CP007809.1	4.41079	65.6	4125	3976	5000	3212	469	1
Mycobacterium tuberculosis ZMC13-264	chromosome: NZ_CP009100.1/CP009100.1	4.41151	65.6	4125	3997	5001	3209	474	0
Mycobacterium tuberculosis ZMC13-88	chromosome: NZ_CP009101.1/CP009101.1	4.41151	65.6	4127	4005	5003	3209	474	2
Mycobacterium tuberculosis 96075	chromosome: NZ_CP009426.1/CP009426.1	4.37938	65.6	4114	3994	5005	3207	459	0
Mycobacterium tuberculosis 96121	chromosome: NZ_CP009427.1/CP009427.1	4.41094	65.6	4138	4007	5014	3198	476	9
Mycobacterium tuberculosis H37Rv; TMC 102	chromosome: NZ_CP009480.1/CP009480.1	4.39612	65.6	4135	4023	5020	3192	485	5
Mycobacterium tuberculosis F28	chromosome: NZ_CP010330.1/CP010330.1	4.4219	65.6	4153	4012	5022	3191	495	1
Mycobacterium tuberculosis 22115	chromosome: NZ_CP010337.1/CP010337.1	4.40183	65.6	4136	3907	5031	3188	514	7
Mycobacterium tuberculosis 37004	chromosome: NZ_CP010338.1/CP010338.1	4.41709	65.6	4151	3892	5039	3182	517	8
Mycobacterium tuberculosis 22103	chromosome: NZ_CP010339.1/CP010339.1	4.39942	65.6	4115	3901	5054	3179	512	14
Mycobacterium tuberculosis 26105	chromosome: NZ_CP010340.1/CP010340.1	4.42649	65.6	4150	3898	5064	3173	499	10
Mycobacterium tuberculosis W-148	chromosome: NZ_CP012090.1/CP012090.1	4.41855	65.6	4133	4025	5067	3172	462	3
Mycobacterium tuberculosis SCAID 187.0	chromosome: NZ_CP012506.2/CP012506.2	4.41183	65.6	4135	4027	5069	3171	455	2
Mycobacterium tuberculosis 49-02	chromosome I:NZ_HG813240.1/HG813240.1	4.41238	65.6	4126	4029	5069	3171	0	0

3.2 *Mycobacterium tuberculosis* Pan and Core genome estimation

Pan genome, broadly classified as core and dispensable genomes, represent the entire gene repertoire (Medini et al. 2005). Core genome are essential for bacterial growth and are present in each genome of different strains of a specific species, while dispensable genome is not necessarily present in all strains and can lead to strain specific functions like resistance, pathogenicity etc. (Medini et al. 2005). The conserved core genome of *Mycobacterium tuberculosis* was estimated on similarities of BLAST between the genomes on the basis of formerly made rule of 50/50 (Ali et al. 2012, Lukjancenko et al. 2012 & Ussery et al. 2009). When 50 % of alignment (amino acid) were identical, blast hit was considered significant with the condition that the alignment length was at least 50 % of the long gene in comparison. Genes were clustered together according to this criterion in gene families, in case they were 50 % identical in their amino acid sequences. Multiple genes can sort into a single family, if the same 50/50 rule is followed by them. Likewise, grouping of all genes into gene families was carried out. A separate unique gene family was made for the genes that did not fall into any of the gene families. The gene families that had at minimum single gene in mutual between them were collectively put into core genome. Those genes that did not fit into the criteria went into pan-genome of the species (Ali et al. 2013, Trost et al. 2012 & Snipen et al. 2009).

3.3 Phylogenetic estimation of *M. tuberculosis* strains

The evolutionary relationships of the included MTB genomes were estimated through Phylogenetic tree based on concatenated core gene alignments and binary (presence/absence)

pan-matrix using BPGA (Bacterial Pan Genome Analysis pipline) (Narendrakumar et al., 2016). Calculation of Gene matrix was made on the basis likeness or difference in contribution of genes to orthologous gene clusters. MUSCLE was used for the multiple sequence alignment (Narendrakumar et al., 2016). All alignments were concatenated and neighbor-joining phylogenetic tree was constructed (Narendrakumar et al., 2016).

3.4 Antibiotic resistance genes in core genome

Antibiotic resistance genes were determined in the core proteome of the strains. The estimation was dependent on SNP models; homology and the resistance genes bioinformatics catalogue, CARD (Comprehensive antibiotic resistance database, https://card.mcmaster.ca/) was used for visual representation (McArthur et al., 2013). CARD is a novel curated collection of transmutation sequences and resistance genes, which provides models for visualizations and software for detection (McArthur et al. 2013). The data is visualized by resistance mechanisms (McArthur et al. 2013). The presence of antibiotic resistance genes in the core genome means these genes are conserved in all strains which suggests the influential role of these genes in the growth and survival of *Mycobacterium tuberculosis.*

3.5 Core Proteome Categorization

3.5.1: Core Essential genes estimation

In order to estimate the existence of essential genes in *M. tuberculosis* core genome of all included strains (47 genomes), Database for Essential genes was utilized (BLAST parameter

against DEG were set at: E-value cut off 1e-10 and minimum bit score was set at 100) (Luo et al., 2014) Essential genes are crucial for survival of microorganisms and those of multi drug resistance bacteria are considered as effective therapeutic targets (Galperin et al., 1999). Disruption of essential genes can destroy the basic functions carried out by them leading to the elimination of the microorganism, therefore, those essential genes which are common through all strains of a particular microorganism can act as ultimate target for broad spectrum drug (Hassan et al., 2016.

3.5.2 Non-Host proteins estimation of core proteome

To eliminate the chances of autoimmunity, the core proteome of *M. tuberculosis* was aligned with human proteome to pool out the human homologs (Naz et al., 2015) The proteins that were having percentage identity less than 35 % and E-value less than 0.005 were considered as non-host bacterial proteins.

3.5.3 Virulence factors estimation among Core Proteome

Utilizing VfDB (Virulence Factor database) (Chen et al., 2012) and MvirDB (microbial virulence database), the virulent proteins were estimated within core proteome. Subsequent search of Blastp was executed against all protein related to virulence by using the following parameters: greater than100 bit score, percentage Identity>35 % and E-Value<1.0 e-5. The VFDB database has considerable information regarding virulence factors, pathogenicity islands, virulence associated genes, protein function and structure characteristics. Another wide-ranging

database for virulence factors, protein toxins and antibiotic resistance genes identification is MvirDB. It helps in swift characterization of sequences associated to pathogenesis and virulence and extract data from eight databases which are publicly accessible (Zhou et al., 2007).

3.5.4 Transmembrane Helices estimation among Core proteome

Transmembrane helical segments localization and transmembrane proteins topology was predicted using HMMTOP (Tusnady et al., 2001). Furthermore, analysis of proteins for the presence of more than two transmembrane helices was performed. Those proteins which have more than one transmembrane helix often cause failure of purification, recombination (Xiang et al., 2013) and are difficult to clone and express thus such proteins are often omitted and those with one or no transmembrane helices were selected as nominees for suitable vaccine target (Krogh et al. 2001).

3.5.5 Sub-cellular localization of prioritized proteins

To analyze the sub-cellular localization of proteins PsortB and CELLO2GO was used (Yu et al. 2014). It aids in sorting out the bacterial proteins as cytoplasmic, periplasmic, extracellular, outer and inner membrane. Usually the proteins that are situated in periplasmic, extracellular, cell wall and outer membranes are preferred as effective vaccine candidates (Zagursky et al., 2003) but *M. tuberculosis* doesn't have periplasmic region so only the protein of extracellular, cell membrane and cell wall was preferred as effective vaccine candidates.

3.5.6 Molecular weight estimation of prioritized proteins

Proteins having molecular weight greater than 110KD are difficult to purify and process of vaccine development of it is ineffective, therefore, they are often omitted and proteins with molecular weight of less than 110KDa are chosen as potential vaccine targets (Barh et al., 2013). For the estimation of molecular weight of prioritized proteins, Expasy PI/MW tool was used (Gasteiger et al., 2005).

3.6.6 Epitope mapping of prioritized proteins

In peptide vaccine development process, epitopes of both T-cell and B-cell, which have ability to generate required immune responses, are identified (Sette et al., 1994). Immuno-informatics computational approaches are used extensively in predicting B and T-cell epitopes (Brusic et al., 2005). Therefore, prioritized proteins obtained in our study were subjected to sequential epitope mapping steps to predict suitable epitopes. One of the most important step in the development of vaccine is the Selection of peptides from the prioritized proteins that binds to MHC I and II molecules (Provenzano et al., 2006). ABCpred was used (threshold value >0.6) for the recovery of B-cell epitopes from the prioritized proteins. The B-cell epitopes were predicted by ABCpred on the basis of artificial neural networks (Saha et al., 2007). The 20-amino acid B-cell epitopes obtained were consequently scrutinized for T-cell epitopes by using Proped1 and Proped servers for their interaction with MHC I and MHC II class molecules, respectively (Singh et al., 2003). An antigen ability to produce immune response is dependent upon its recognition by both classes of MHC and its binding with these (Chaplin et al., 2003). NetSurfP was used to estimate surface exposure of the T-cell epitopes (Petersen et al. 2009). The antigenicity of the

epitopes was checked by using Vaxijen v2.0 (threshold value >0.4) (Doytchinova et al.,2007). Epitopes with values more than 0.4 were considered potentially antigenic. MHCpred2.0 was used to estimate the binding of the identified peptides with DRB0*0101 allele and results were sorted out based on half maximal inhibition concentration (IC50) score. A classification scheme is followed to simplify estimation of binding peptides, which divides the peptides into binders with high affinity(<50nM), binders with medium affinity (50-500nM), binders with low affinity (>500nM) and non-binders (Sette et al. 1994). Those epitopes that binds with DRB*0101 allele are generally preferred in various in silico reverse vaccinology studies, as DRB0*0101 is a prevalent and common allele in human population worldwide (Hosseingholi et al., 2014, Naz et al., 2015 & Rakesh et al., 2009). IC50 values are calculated from a competitive binding assay and are basically binding affinity measures (Blythe et al., 2002). VirulentPred was used to estimates the virulence of candidate epitopes (Garg et al., 2008).

3.7 Estimation of Epitope conservation

To confirm the conservation of the epitopic region across all strains of *M. tuberculosis*, conservation analysis was performed using CLC work bench (http://www.clcbio.com/products/clcgenomicsworkbench). This graphical and user-friendly software performs multiple sequence alignment, and consensus sequence for the epitopes was obtained. Analysis of conservation of the epitope throughout all the 47 strains of the *M. tuberculosis* was done by aligning the particular protein sequence of all strains with each other as shown in FIG 9. The degree to which specific epitopic region exhibit similarity or variability

among genomes provides Significant information regarding immunological response, evolutionary correlation, structure and function of prioritized proteins of peptides.

3.8 Protein structure

3D structure visualization of protein help in understanding functional sites, sequence patterns, binding sites and candidate proteins interactions with other targets (Gabdoulline et al., 2003). PHYRE2 was used to explore 3D structures of prioritized proteins by comparative modeling (Kelley et al., 2015). In comparative modeling method, templates which are experimental protein structures are used to generate models for target proteins (Kelley et al., 2015). For each of the query protein, PHYRE2 was employed which generated various templates through BLAST. The template whose sequence was more similar to the candidate proteins and has high identity was selected. For instance, esterase has 100 % sequence similar to alpha/Beta-Hydrolases (PDB id-1r88) with 89 % coverage. Similarly, for query proteins Secreted antigen 85-C (85C) protein, has 100 % similarity with alpha/beta Hydrolases (PDB id-1dqz) with coverage of 82 %. Query protein ESX conserved component 5 exhibited 33 % similarity with esx-1 secretion system protein eccb1(PDB id-4KK7) and the sequence coverage was 76 %. Query protein Lysine-N-oxygenase has sequence similarity of 54 % with l-lys monooxygenase (PDB id-4D7E) and the coverage for it was 91 %. Query protein ESX-2 secretion system 2 has 35 % similarity with esx-1 secretion system protein eccb1 (PDB id-4KK7) and coverage for it was 77 %. Query protein thiol peroxidase has 100 % similarity with probable thiol peroxidase (PDB id-2YZH) having coverage of 100 %. Structures for query protein PPE family protein (PPE42) and exported repetitive partial protein was generated through the intensive modeling

mode of PHYRE2 as there was no single protein found having 100 % similarity with them so their structure was generated using multiple templates and ab intio techniques (Kelley et al., 2015).

3.9 Functional annotation and Biological Molecular pathways estimation of prioritized proteins

Functional analysis of protein helps in studying its molecular, biological and biochemical behavior. Here Prioritized proteins functional annotation was performed through Blast2GO (Conesa et al., 2005). Cluster of orthologous groups (COG) of prioritized proteins were found through STRING database. To have a better understanding of molecular interactions and role of the prioritized proteins in pathways, the KEGG id of the prioritized proteins were extracted from Uniport.

3.10 Interactome analysis of prioritized proteins

Search tool for the retrieval of interacting genes (STRING) was used for the analysis of interactions of prioritized proteins (Szklarczyk et al., 2011). The STRING database (v10) is mainly devoted on finding global functional links among proteins (Szklarczyk et al., 2014). Protein-protein interactions are the backbone of cellular function and studying such interactions helps in revealing molecular basis of diseases, in understanding biological processes and molecular mechanism and in identifying potential therapeutic targets (Bultinck et al., 2012). STRING helps to determine interactions between different candidate proteins and also gives information about proteins domain estimation and functional cataloguing (Samant et al., 2016). Additionally, this helps in evaluation of different pathways that are triggered due to host/pathogen interaction that leads to different disease symptoms (Samant et al., 2016)

CHAPTER FOUR: RESULTS

4.2 Pan-genome and Core-genome analysis

Comprehensive pan-genome analysis studies can assist in understanding the bacterial species functional adaptation. In our study, analysis of *M. tuberculosis* pan genome reveals 5,069 sequences are coding, of which 3,170 sequences are of core and the remaining 1,898 sequences establishes the dispensable genome (TABLE 2). The dispensable genome makes pan genome relatively large because it contains various strains extensive pool of dispensable genes. The pan-core plot (Fig 4) illustrates, by adding new genome, core genome subsequently decreases, signified by the red power line going downward, while the pan-genome increase with it, the blue power line going up ward represent this. With addition of a new genome, it is observed that core genome stabilizes and at last a set of conserved 3170 genes are obtained which represents core genome of all the selected 47 strains (Fig 4) of the *M. tuberculosis*. The core sequences are highly conserved in nature and this is backed up by the statistic that core genome is about 77 % of the average genome of all the 47 strains. Bacteria acquire new genes via horizontal gene transfer under the selective antibiotic and environmental pressure to survive in a specific condition (Martínez et al., 2007) but the exposure of *M. tuberculosis* to wide range of antibiotics over the years has led this bacterium up regulate its innate resistance mechanisms mainly through acquiring new mutations as resistance due to acquiring new genes from environment is not reported (Madhukar et al., 2016).

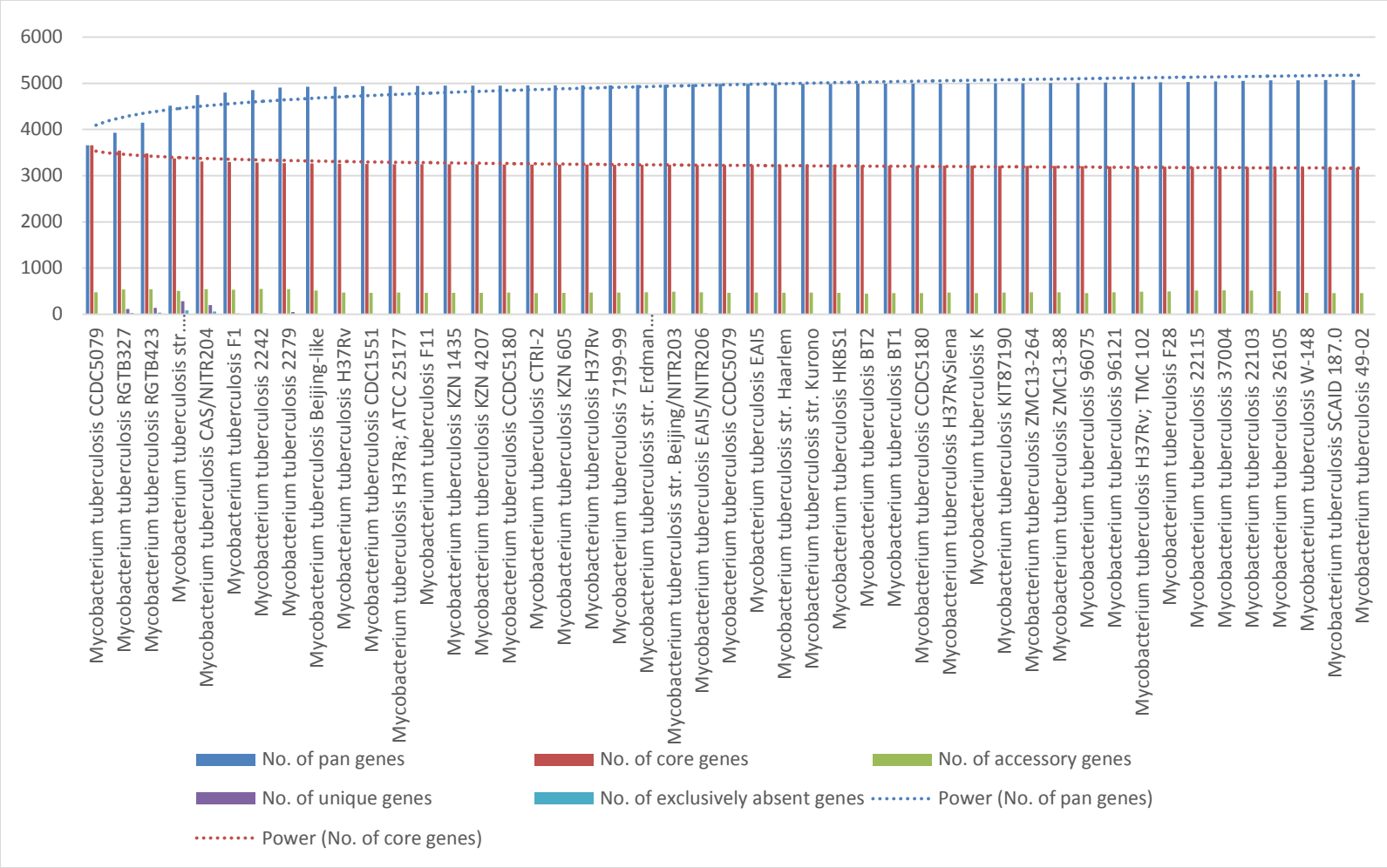


Figure 5| Pan-core genome plot of completely sequenced strains of *M. tuberculosis*.
M. tuberculosis strains are shown on the X-axis in the graphs while the number of genes present in different strains are represented by the Y-axis. The blue power line in the graph represents the expansion of pan-genome and the red power line indicates the conservation/development of the core genome with each new addition genome. Accessory genes number in each strain is represented by the green bar in the graph. The purple bars indicated the number of Unique genes, whereas, the light blue bar represent the number of exclusively absent genes in each strain.

4.3 Sub-cellular localization of Core-genome

The sub-cellular localization estimation of Core proteins reveals that 1639 proteins are cytoplasmic, 872 located on cytoplasmic membrane, no protein was of outer membrane and periplasmic, 45 proteins are extracellular, 11 proteins are of cell wall and 603 of them of were unknown origin (Additional file) (Gardy et al., 2005). Protein sub-cellular localization knowledge can improve identification of therapeutic target significantly (Yu et al., 2006). For example, drug molecules can easily access secreted and plasma membrane proteins as they are localized in the extracellular space or on the cell surface. In case of vaccine, such proteins are easily accessible to the immune system so already activated immune system towards such protein can destroy the pathogen by binding to these protein present on the pathogen. Thus, secreted and cell surface proteins of bacteria are of interest for their potential as diagnostic targets and as a vaccine candidate (Shanmugham et al., 2013) (Fig 5).

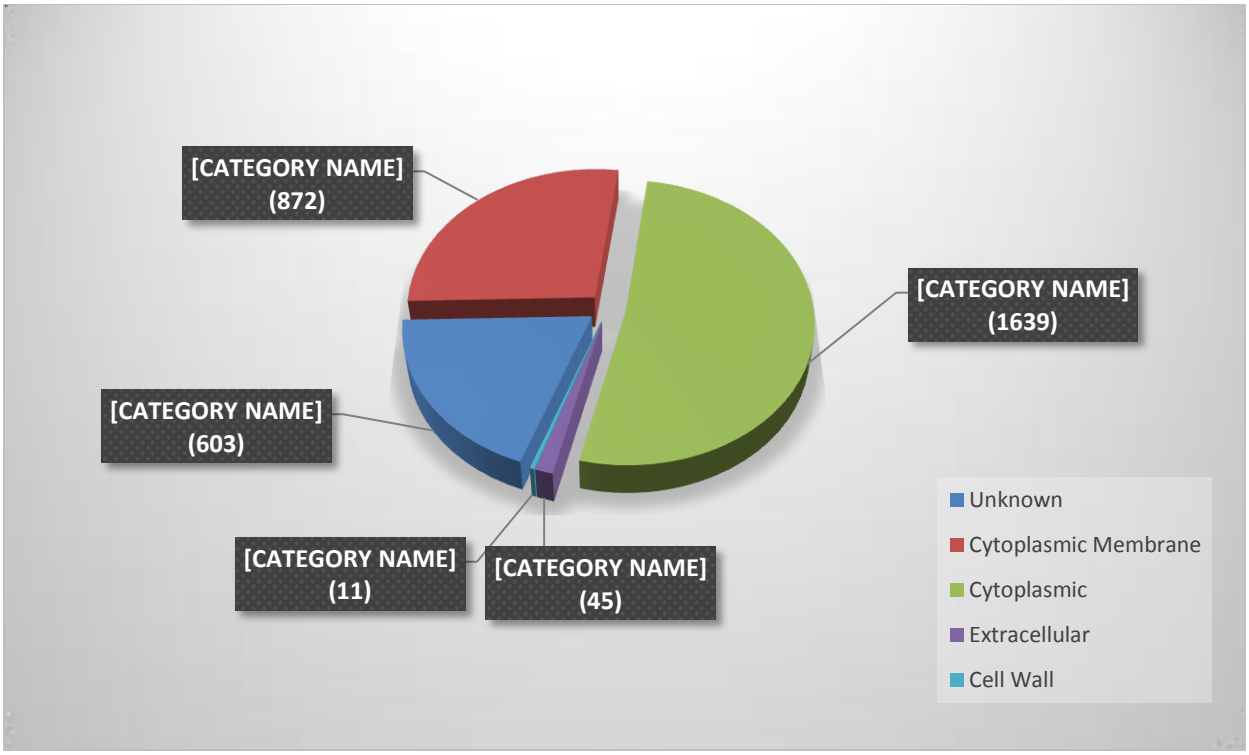


Figure 6| Core proteome sub-cellular localization for vaccine candidate prioritization.
PsortB was used to analyze the Core proteins (3170) sub-cellular localization and it reveals that 1639 (51.7 %) proteins are cytoplasmic, 872 (27.5%) located on cytoplasmic membrane, 45 (1.4%) proteins are extracellular, 11 (0.34%) proteins are of cell wall and 603 (19.2%) of them of were unknown origin.

4.4 Resistome analysis of Core-genome

The core resistance of *M. tuberculosis* was identified by utilizing CARD. The results were obtained by comparing the FASTA files of *M. tuberculosis* core with the resistance genes in the CARD database. Seventeen resistance genes were found to be present in the core genome. Three genes, mutant kasA mutant embC and gyrA, has G269S, V981L and S95T Single Nucleotide Polymorphism (SNP) respectively. drrA, drrB, drrC, mtrA and efpA are involved in efflux pump conferring antibiotic resistance. mfpA, gyrA and mfd are fluroquinolone resistance genes. abcA and tap genes are involved in resistance to tetracycline. abcA also confers resistance to fluroquinolone. ksA mutant, AAC(2')-Ic, alaS, ileS, murA and mutant embC are

isoniazid, aminoglycoside, aminocoumarin, mupirocin, fosfomycin and ethambutol resistance genes respectively. Whereas, Erm(37) confer resistance to lincosamide , streptogramin and macrolide. The results of antibiotic resistance are illustrated in Fig 6.

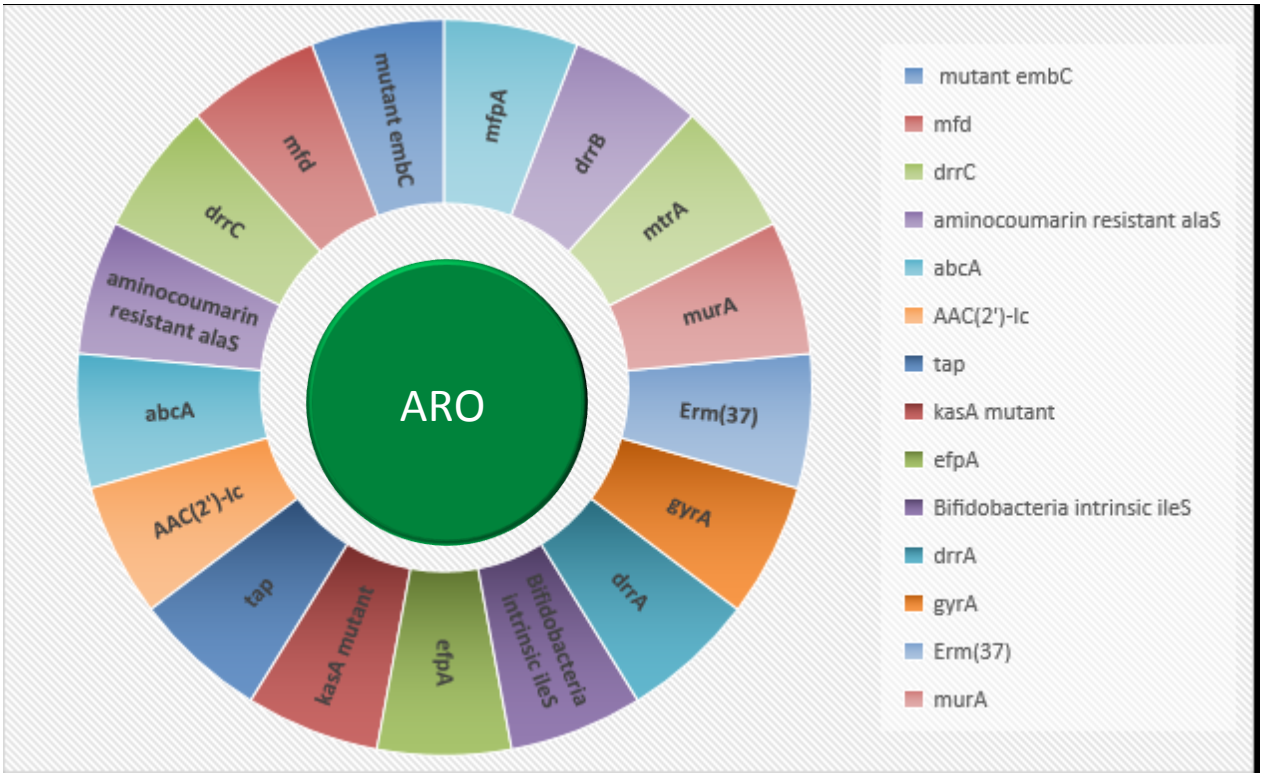


Figure 7| visual representation of antimicrobial genes found in the core genome of *M. tuberculosis* by resistance identifier database.
drrA, drrB, drrC, mtrA and efpA are involved in efflux pump conferring antibiotic resistance. mfpA, gyrA and mfd are fluroquinolone resistance genes. abcA and tap genes are involved in resistance to tetracycline. abcA also confers resistance to fluroquinolone. ksA mutant, AAC(2')-Ic, alaS, ileS, murA and mutant embC are isoniazid, aminoglycoside, aminocoumarin, mupirocin, fosfomycin and ethambutol resistance genes respectively. Whereas, Erm(37) confer resistance to lincosamide , streptogramin and macrolide.

4.5 Phylogenetic analysis

Phylogenetic analysis based on core genome helped to understand the pattern of evolution among the strains under study. Mainly, the phylogenetic tree resulted in two main clusters. The small cluster (YELLOW in Fig 7) included the strains namely *Mycobacterium tuberculosis* H37Rv, *Mycobacterium tuberculosis* H37Ra; ATCC 25177, *Mycobacterium tuberculosis* H37Rv, *Mycobacterium tuberculosis str. Kurono, Mycobacterium tuberculosis* H37RvSiena. Interestingly, based on core genes strains from USA and Japan fall into a distinct cluster. In the bigger cluster, strains from China (BROWN in Fig 7) (*Mycobacterium tuberculosis CCDC5180, Mycobacterium tuberculosis ZMC13-264, and Mycobacterium tuberculosis ZMC13-88*) make a separate clad, showing their uniqueness being belonging to a different geographical location. Similarly, strains from south-east Asia fall into a separate clad (BLUE in Fig 7) (*Mycobacterium tuberculosis RGTB423, Mycobacterium tuberculosis RGTB327, Mycobacterium tuberculosis 22103, Mycobacterium tuberculosis EAI5, Mycobacterium tuberculosis EAI5/NITR206*). Phylogenetic estimation based on core genomes clearly suggests that selective environmental pressures have led to evolution of strains in specific geographical and ecological niches.

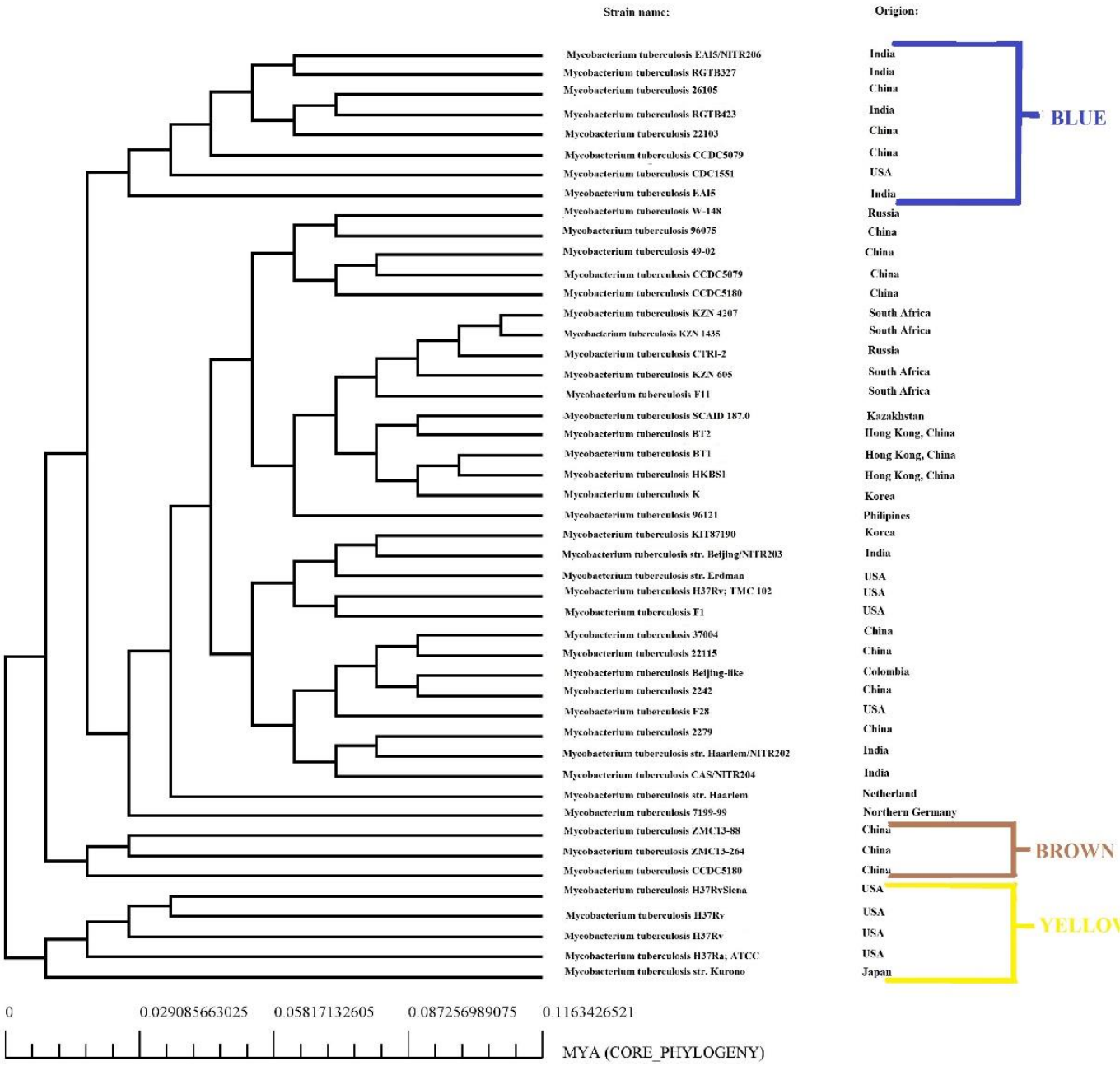


Figure 8| Phylogenetic tree based on core genes of *Mycobacterium tuberculosis.*
Mainly, the phylogenetic tree resulted in two main clusters. The small cluster YELLOW included the strains namely *Mycobacterium tuberculosis* H37Rv, *Mycobacterium tuberculosis* H37Ra; ATCC 25177, *Mycobacterium tuberculosis* H37Rv, *Mycobacterium tuberculosis* str. Kurono *Mycobacterium tuberculosis* H37RvSiena. Interestingly, based on core genes strains from USA and Japan fall into a distinct cluster. In the bigger cluster, strains from China BROWN (*Mycobacterium tuberculosis* CCDC5180, *Mycobacterium tuberculosis* ZMC13-264, and *Mycobacterium tuberculosis* ZMC13-88) make a separate clad, showing their uniqueness being belonging to a different geographical location. Similarly, strains from south-east Asia fall into a separate clad BLUE (*Mycobacterium tuberculosis* RGTB423, *Mycobacterium tuberculosis* RGTB327, *Mycobacterium tuberculosis* 22103, *Mycobacterium tuberculosis* EAI5, *Mycobacterium tuberculosis* EAI5/NITR206). The Phylogenetic estimation based on core genomes clearly suggests that selective environmental pressures have led to evolution of strains in specific geographical and ecological niches.

4.5 Core proteome analysis for prioritized proteins

One of the crucial step in designing therapeutics against bacterial infection is the identification of essential genes for the fact that most of the vaccines and antibiotic targets the essential cellular processes (Ali et al., 2015). Among the 3170 core proteins, 1286 (40.5 % of core proteome) are predicted as essential genes (Zhang et al., 2004) (shown in Venn diagram). Essential genes are comprised of minimal set of gene without which the cellular life cannot be supported. Essential have a broader potential for being therapeutic vaccine targets and targets for those drugs that aims to kill bacteria (Judson et al., 2000). *M. tuberculosis* essential genes has a role in major biological processes like Oxidation-reduction process, Glycolipid biosynthetic process, Lipid transport, Cell redox homeostasis, Cellular oxidant detoxification, Transport, Siderophore biosynthetic process from catechol, Response to oxidative stress, Pathogenesis, Response to antibiotic, Response to nitrosative stress, Evasion or tolerance by symbiont of host-produced nitric oxide, Growth, Mycolate cell wall layer assembly, growth of symbiont in host cell and Host tissue attachment. In order to avoid auto-immunity, the vaccine target shouldn't be similar to host proteome, for this reason, core sequences are aligned with proteome of human to filter out those proteins which have similarity with human proteome. Among the total 3170 core proteins, 2838 proteins (shown in Venn diagram) were given to have value below the threshold

due to which they were considered as bacterial proteins and thus can be considered for this study (Butt et al., 2012). Core conserved genome was analyzed for virulent genes for exploring virulence potential of *M. tuberculosis.* The analysis showed that the bacteria have a significant number of virulent genes (341) in its core genome which facilitates in probably survival in adverse conditions and pathogenesis of the organism.

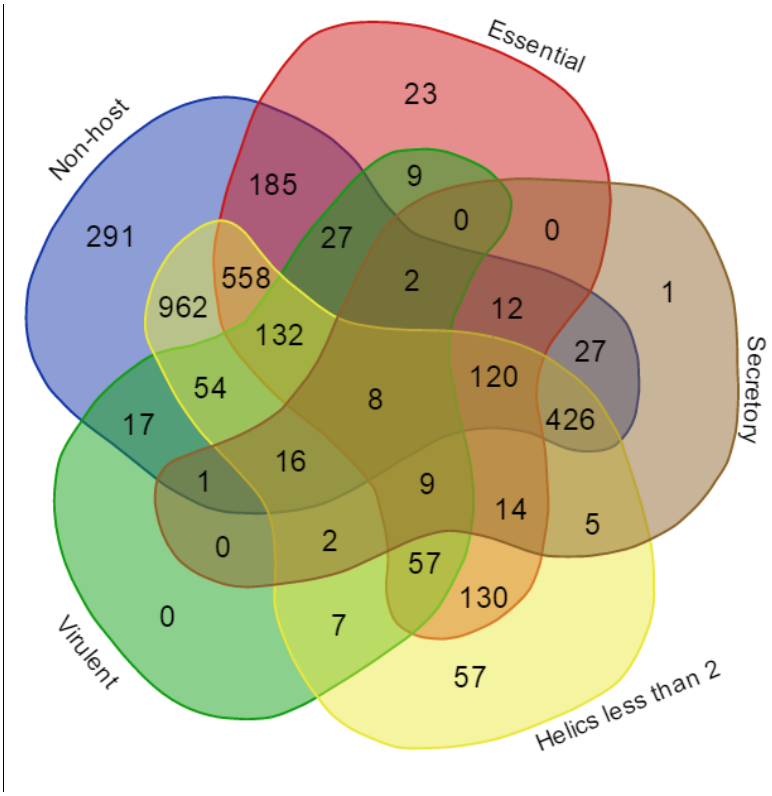


Figure 9|Visual representation of the vaccine candidate identification through the use of a Venn diagram. Five parameters were used to filter out candidate protein namely being part of the core genome (3170) non-host homology (2838), essential proteins (1286), virulent (341) and sub cellular localization (643). Eight proteins that full-filled the four criteria for prioritization (essential, secreted, non-host, virulent) were further subjected to epitope mapping approaches. In the above Venn diagram, the red color scheme represents Essential proteins, blue represents Non-host, green represents Virulent proteins, the yellow represents the proteins having helices less than 2 and brown color scheme represents proteins part of secretome/extraproteom.

4.6: Selection of Vaccine candidates

Evolution follows a conservative nature and core genes presence in the genome is a proof of it (Lapierre et al., 2009). It characterizes an ideal dataset for the investigation of suitable vaccine candidates against *M. tuberculosis*. After passing the core proteins through sequential steps of checking whether they are secreted or are at surface, their essentiality, non-host homology and virulence, eight proteins were selected as a prioritized protein including Esterase, Secreted antigen 85-C (85C), PPE family protein, ESX conserved component 5, lysine-N-oxygenase, ESX-2 secretion system 2, Exported repetitive partial and thiol peroxidase. These proteins are then further explored for their comparative homology modeling, protein-protein interactions, epitopes, surface topology and extensive functional analysis. The methodology in detail is described in the following sections.

4.7 Prioritized Proteins epitope mapping

Vaccine and drug development is a tiresome, expensive and lengthy process and filtered and prioritized proteins can help in optimization of the process and make it less laborious, expensive and less time consuming. After filtering in core proteins that are essential, virulent, non-host, had < 2 transmembrane helix and desired sub-cellular localization, 8 core proteins are found had all the above mentioned parameter and are considered as potential vaccine targets which includes Esterase, Secreted antigen 85-C (85C) (antigen 85 complex C)(Ag58C), ESX conserved component 5, lysine-N-oxygenase, , PPE family protein, ESX-2 secretion system 2, Exported repetitive partial and thiol peroxidase. These proteins molecular weight is estimated to be <110KDa and they fit in the criteria of transmembrane helices (less than 2) (Gasteiger et al., 2005). These 8 prioritized proteins are also found to be potentially antigenic after Antigenicity

analysis by using, Vaxijen v2.0 and exhibited score of more than 0.4 (Table 3). These proteins are then subjected to approaches of epitope mapping as they are considered as potential vaccine targets. B-cell epitopes for each of the protein was predicted (20-mer sequence) and are further analyzed on Propred and Propred1 for its binding capability with MHC I and MHC II molecules. Afterward, those epitopes of T-cell which IC50 values were optimal for DRB*0101 allele and binds to a maximum number of MHC I and II class molecules were selected. Three prioritized proteins included in our study are found to have high binding affinity having values of IC50 less than 50nM, these include Secreted antigen 85-C (85C) (antigen 85 complex C)(Ag58C), ESX conserved component 5 and exported repetitive partial. The remaining five prioritized proteins fall in the category of medium affinity binders as their IC50 value is between 50nM to 300nM. The 9-mer sequence of T-cell epitopes for each of the prioritized protein had maximum number of amino acid exposed (surface accessibility) and their vaxigen 2.0 score was more than the cut off value of 0.4. These 9-mer T-cell epitopes bind to different number of the MHC I and II alleles. The details of the surface accessibility, vaxigen 2.0 score, IC50 value and binding to different alleles of MHC I and II are given in the TABLE 3.

Table 3| Prioritized core vaccine candidates against *Mycobacterium tuberculosis*

Protein name	B cell epitope	T cell epitope	T cell epitope Location	MHC I allele count	MHC II allele count	Vexijen score (cut off = 0.4)	virulent pred (cut off = 0.5)	IC50	Surface Accessibility (cut off = 4)	COG ID	KEGG id	Molecular weight	pI
Esterase	YENLMVPSP SMGRDIPVAFL	MGRDIPVAF	46-54	12	9	0.894	1.075	58	5	COG0627	mtu: Rv3803c.	31088.89	6.13
Secreted antigen 85-C	GLTLRTNQ TFRDTYAADGGR	FRDTYAADG	284-292	8	1	1.477	1.061	3.3	5	COG0627	mtu: Rv0129c.	36771.27	5.92
PPE family protein (PPE42)	VMGGTDSL LPLPNIPLLEYA	LLPLPNIPL	279-287	20	21	1.939	1.061	109	4	COG5651	mtu: Rv2608.	59674.54	4.44
ESX conserved component 5	AGARFGVE DSKEARDALGLT	FGVEDSKEA	447-455	20	4	1.591	1.591	15	6	N.A	mtu: Rv1782.	53689.25	7.13
Lysine-N-oxygenase	HLRGRVA HAVGRQGQIRLT L	LRGRVAHAV	291-299	6	21	0.48	1.061	89	6	COG3486	mtu: Rv2378c.	46943.94	6.17
ESX-2 secretion system 2	ATLRALG LDPGAAVQAPWP L	LGLDPGAAV	442-450	18	8	0.823	1.066	163	6	N.A	mtu: Rv3895c.	51586.98	9.74
Exported repetitive partial	SQFGINIPP VPSLTGSGDAS	INIPPVPSL	71-79	28	3	0.552	1.06	36	7	N.A	mtu: Rv3810.	27668.37	4.34
Thiol peroxidase	QITLRGNAIN TVGELPAVGS	LRGNAINTV	6-14	5	30	0.61	1.056	177	7	COG2077	mtu: Rv1932.	16896.14	4.37

4.7 Analysis of Epitope Conservation

Conservation pattern of 100 % was shown by six epitopes, from the eight epitopes of prioritized proteins, throughout all the strains of *M. tuberculosis*. These epitopes were of the proteins Esterase, Secreted antigen 85-C (85C), PPE family protein, Lysine-N-oxygenase, ESX conserve component 5 and Thiol peroxidase. The epitopes alignment, the consensus sequences and the conservation of the selected epitopes (MGRDIPVAF, FRDTYAADG, LLPLPNIPL, LRGRVAHAV, FGVEDSKEA, LRGNAINTV) in all strain is shown in Fig. As they are conserved, these epitopes could efficiently induce immune response within the host against *M. tuberculosis*. considered. Two epitopes (LGLDPGAAV, INIPPVPSL for ESX-2 Secretion system 2 and Exported repetitive partial respectively) have shown conservation of 97.87%. The alignment, conservation and consensus sequences of all the epitopes are shown in Figure 8.

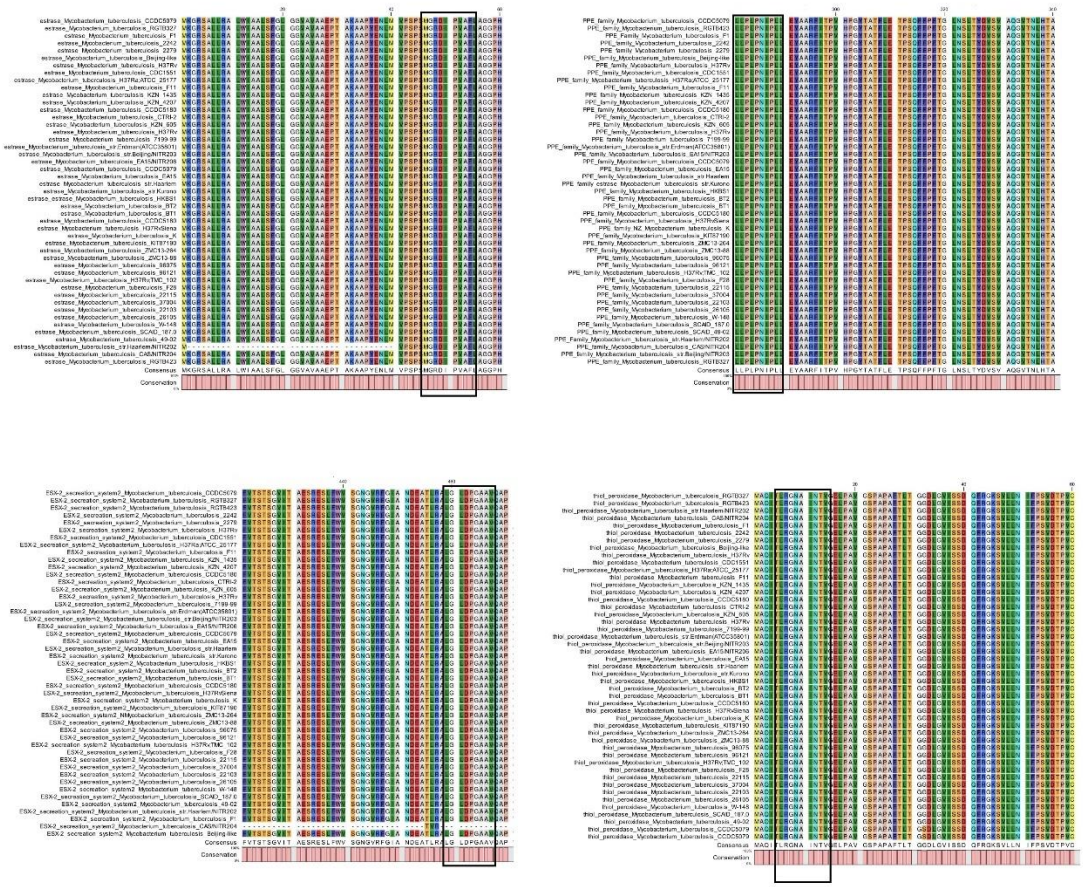


Figure 10| Analysis of epitope conservation.
The figure shows the alignment of prioritized protein sequence in all the strains, consensus sequences derived from the strains and the conservation of the epitopes. MGRDIPVAF, LLPLPNIPL LGLDPGAAV and LRGNAINTV of the proteins esterase, PPE family protein (PPE42), ESX-2 secretion system 2 and thiol peroxidase respectively in all of the strains.

4.8: Structural analysis of Prioritized proteins

The 3D structure of the prioritized proteins which were obtained from PHYRE 2 (Kelley et al., 2015) were analyzed and the selected 9-mer T-cell epitopes were visualized on the proteins in the form of red spheres as shown in the figure. For the recognition of epitopes by MHC molecules for inducing strong immunogenic response, the epitopes are required to be on the surface of the protein and the surface accessibility analysis and the figure below shows that the epitopes selected in our study exist on the surface of the protein (Dahlback et al., 2006).

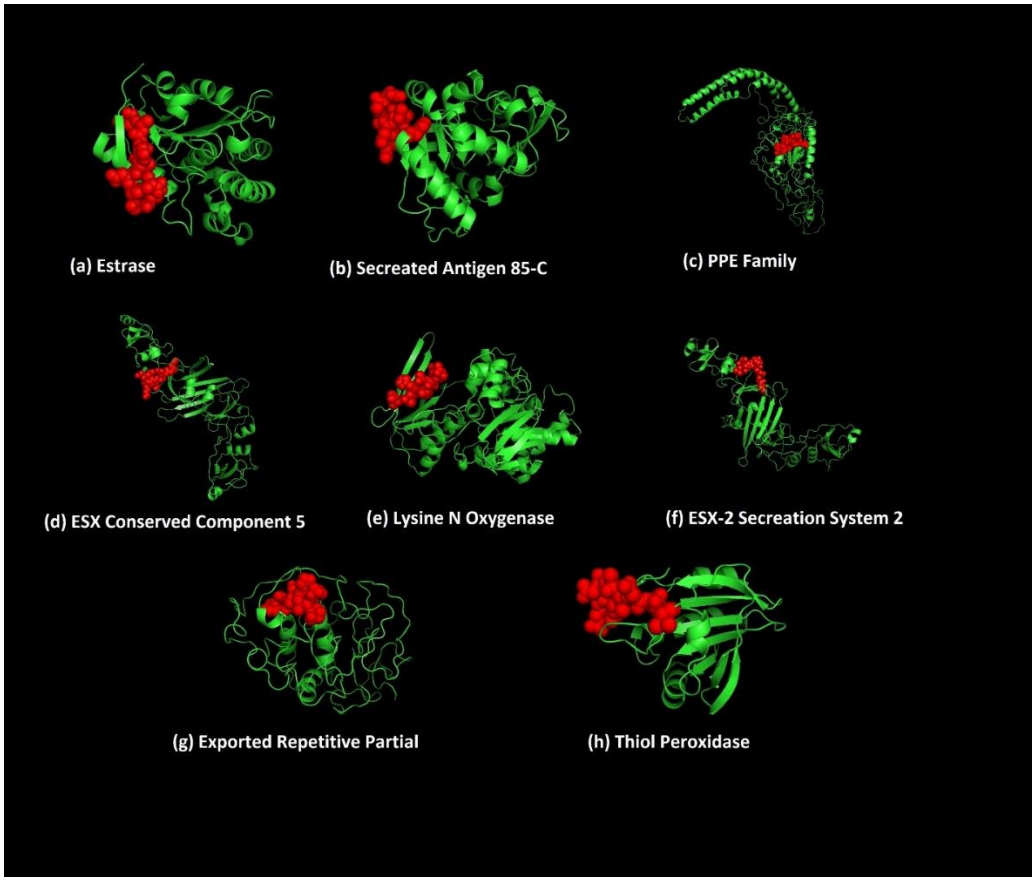


Figure 11| 3D structures of prioritized proteins with epitope.
The structures of protein were predicted using PHYRE2 and pymol was used to visualize the 9-mer T-cell epitopes on the proteins as red spheres. The red spheres illustrate the surface topology of the epitopes. For the recognition of epitopes by MHC molecules for inducing strong immunogenic response, the epitopes are required to be on the surface of the protein. All the predicted epitopes satisfy this criterion and they can be further explored for in vivo testing.

4.1 Genome Statistics and Organization

Mycobacterium tuberculosis has become a serious concern due to its drug resistance (Madhukar et al., 2016). Next generation sequencing technology has made available wide range of genome data on international databases (Hassan et al., 2016). *Mycobacterium tuberculosis's* 47 complete genome sequences are available to date on GenBank/NCBI. Our analysis involved all of these completely sequenced genome (Table 2). Whole proteome analysis of these genomes revealed total 380,976 genes and on average *M. tuberculosis* proteome/genome contain 4897 functional genes or proteins. GC content on average is observed around 65.6 %. Gene count on average is comprising 4111 genes, with lowest number of genes present in *Mycobacterium tuberculosis RGTB423* (3670 genes) and highest number of genes reported in *Mycobacterium tuberculosis 2279* (4647 genes) As a consequence of strong antibiotic selective pressure, the bacteria have attained virulence genes and significant antibiotic resistance for survival over time (Alonso et al., 2001). Single gene prediction program, Prodigal, is used to get consistency in the proteomic and genomic data of all the sequences (Hyatt et al., 2010).

4.9: Prioritized proteins Functional annotation

Proteins functional annotation is important as it helps in understanding their biochemical activities, physiological behavior and biological processes (Gotz et al., 2008). Prioritized proteins were found to have significant role in various biological process such as: oxidation-reduction process, glycolipid biosynthetic process, lipid transport, cell redox homeostasis, cellular oxidant detoxification, transport, siderophore biosynthetic process from catechol, response to oxidative stress, pathogenesis, response to antibiotic, response to nitrosative stress, evasion or tolerance by symbiont of host-produced nitric oxide, growth, mycolate cell wall layer

assembly, cell growth of symbiont in host cell and host tissue attachment as shown in Fig 10. Apart from involvement in biological process, analysis of molecular function of the prioritized proteins shows that they carry out trehalose O-mycolyltransferase activity, thioredoxin peroxidase activity, disulfide oxidoreductase activity, L-lysine 6-monooxygenase (NADPH) activity, transferase activity, transferring acyl groups, protein binding, diacylglycerol O-acyltransferase activity, transferase activity/transferring acyl groups other than amino-acyl groups, ATP binding, hydrolase activity, peroxidase activity, peroxiredoxin activity, thioredoxin peroxidase activity and short-chain carboxylesterase activity as shown in Fig 10.

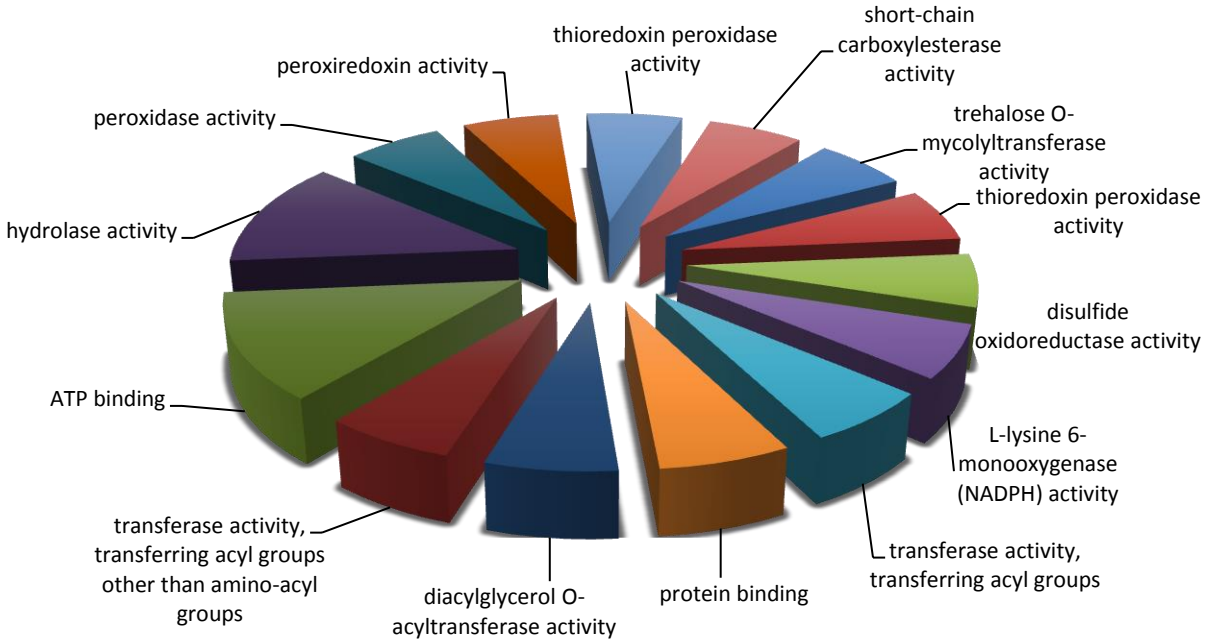


Figure 12|Prioritized protein functional annotation:
BlAST2GO was used for prediction and analysis of protein. (A) shows the different molecular function of the prioritized proteins that includes trehalose O-mycolyltransferase activity, thioredoxin peroxidase activity, disulfide oxidoreductase activity, L-lysine 6-monooxygenase (NADPH) activity, transferase activity, transferring acyl groups, protein binding, diacylglycerol O-acyltransferase activity, transferase activity/transferring acyl groups other than amino-acyl groups, ATP binding, hydrolase activity, peroxidase activity, peroxiredoxin activity, thioredoxin peroxidase activity and short-chain carboxylesterase activitythe.

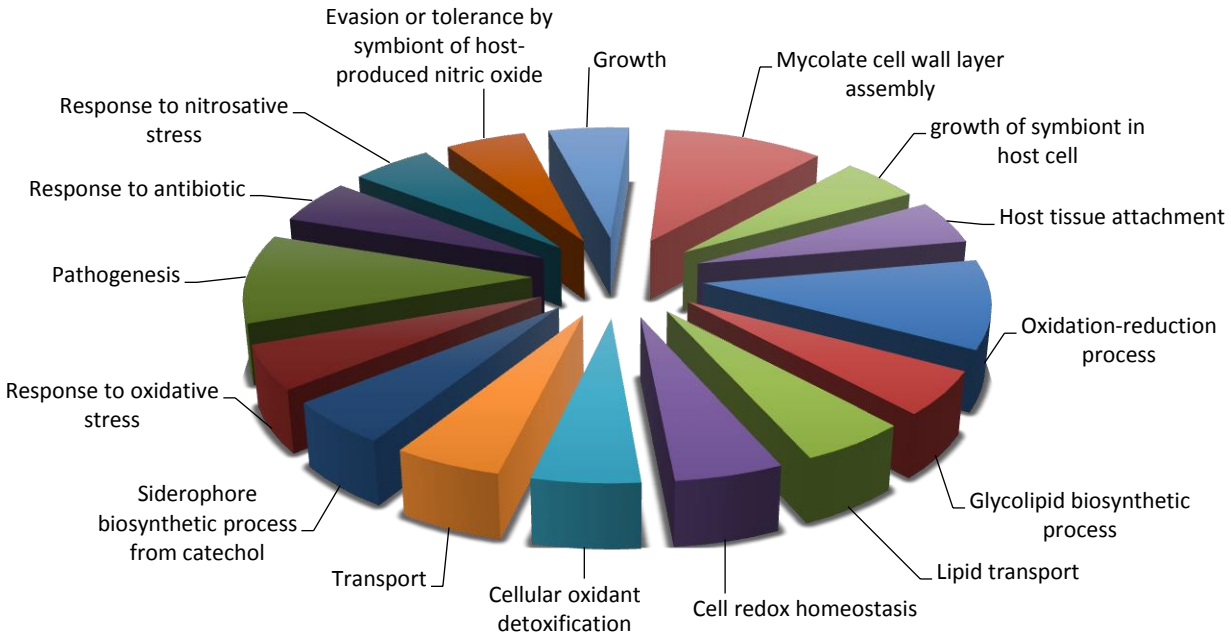


Figure 13: Prioritized protein functional annotation:

This figure illustrates the biological processes that prioritized proteins takes part in which includes Oxidation-reduction process, Glycolipid biosynthetic process, Lipid transport, Cell redox homeostasis, Cellular oxidant detoxification, Transport, Siderophore biosynthetic process from catechol, Response to oxidative stress, Pathogenesis, Response to antibiotic, Response to nitrosative stress, Evasion or tolerance by symbiont of host-produced nitric oxide, Growth, Mycolate cell wall layer assembly, growth of symbiont in host cell and Host tissue attachment.

4.10 Interactome analysis of Prioritized Proteins

Analysis of protein-protein interactions of prioritized proteins was performed to study the functional importance of them in a metabolic network (Pe'er et al., 2001). STRING database was used to get high confidence protein network of the prioritized proteins. Esterase, Secreted

antigen 85-C, ESX conserved component 5, Lysine-N-oxygenase, ESX-2 secretion system 2 and thiol peroxidases have more than 5 interactors as shown in figure. Protein with more interactors are usually considered to be metabolically important as they participate in many pathways and thus can act as a suitable drug candidate. PPE family protein (ppe42) and Exported repetitive partial had one and three interactors respectively. The proteins that thiol peroxidase interacts with mainly plays role in biological processes like growth, protein refolding, response to heat, response to hypoxia, adhesion of symbiont to host, cell redox homeostasis, response to oxidative stress, evasion from host-produced nitic oxide, cysteine biosynthetic process from serine, sulfate assimilation, glycerol ether metabolic process, response to oxidative stress, nucleotide organization, DNA protection, response to nitrosative stress glutamine metabolic process and glycine decarboxylation via glycine cleave system. Lysine-N-Oxygenase interacts with proteins that are involved in biological process like protective immune responses, in synthesis of mycobactin and metabolic processes. ESX-2 secretion system 2 interacts with proteins that are mainly involved in growth, transport, response to antibiotic, glycolipid biosynthetic process, mycolate cell wall layer assembly and fatty acid biosynthetic process. ESX conserved component 5 interacts with proteins that are involved mainly in growth, transport and fatty acid biosynthetic process. PPE family protein (PPE42) interact protein that is involved in pyridoxine biosynthetic process. Exported repetitive partial interacts with proteins that are mainly involved in cell wall biogenesis/degradation. Secreted antigen 85-C interacts with proteins that are mainly involved in metabolic process, response to acid chemical, growth of symbiont in host cell, transport, response to antibiotic, growth, cell wall organization and cell wall polysaccride biosynthesis. Esterase mainly interacts with proteins that are involved in processes like oxidation-reduction process, mycothiol-dependent detoxification, evasion or tolerance by

symbiont of host produced nitric oxide, growth of symbiont in host cell, response to acid chemical, growth, lipid catabolic process, transport, glycolipid biosynthesis process and aromatic compound catabolic process.

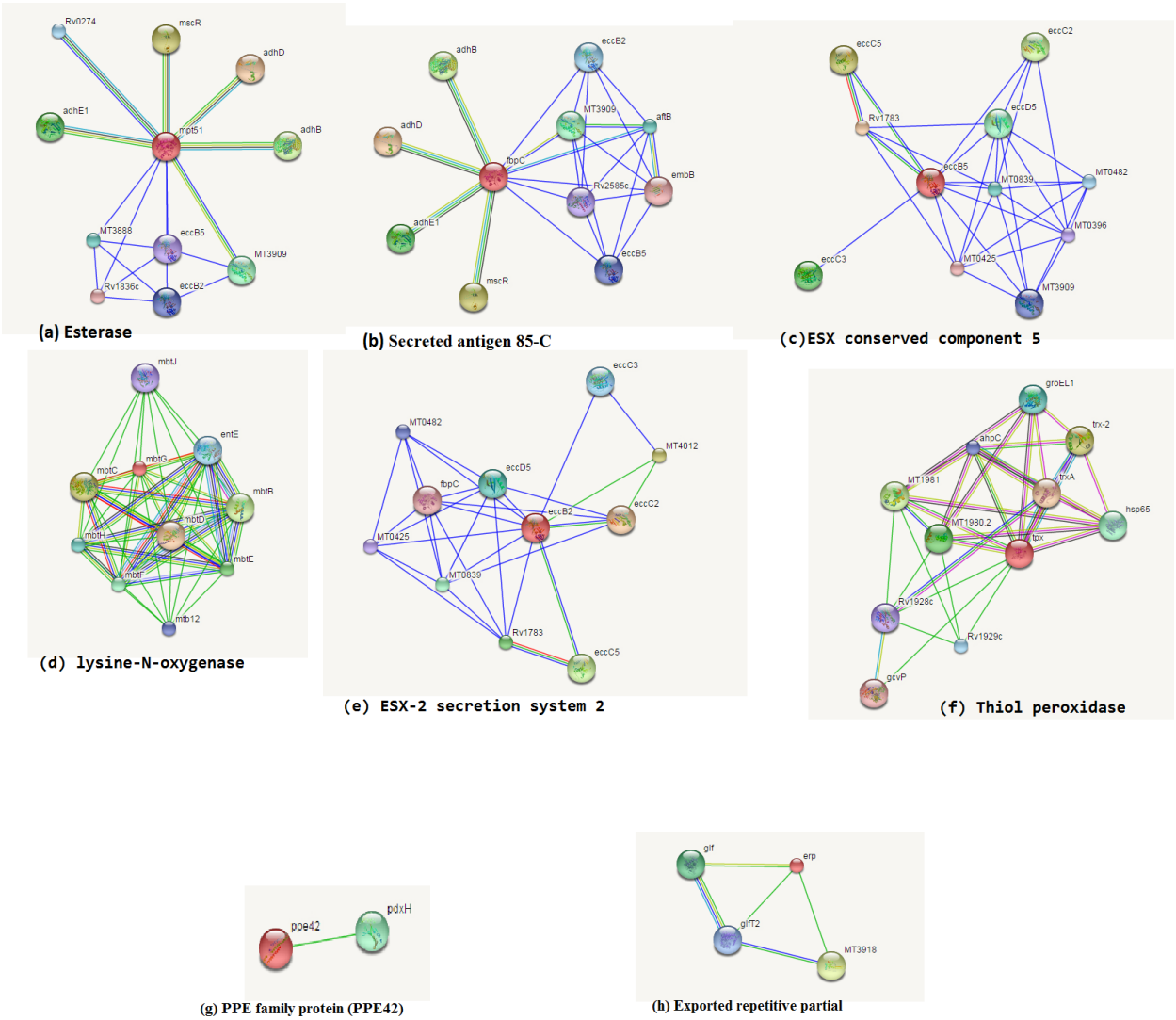


Figure 14| The protein-protein interactions were predicted by STRING database.
The figure shows protein-protein interactions of prioritized proteins; Esterase, secreted antigen 85-C, PPE family protein, ESX conserved component 5, lysine-N-oxygenase, ESX-2 secretions system 2, exported repetitive partial and thiol peroxidase.

CHAPTER FIVE: DISCUSSION & CONCLUSION

5.1 Discussion

Tuberculosis is a major cause of morbidity and mortality worldwide, especially in developing countries. Comparative pan-genomics approach can provide a better understanding of *M. tuberculsosis* genetic makeup(Vernikos et al., 2015). Next generation sequencing has opened the avenues for a thorough understanding. We performed an exhaustive analysis on all the completely sequenced strains of *M. tuberculosis* available on NCBI website. Pan-genome of all the 47 strains estimated to be 5069 genes. Our analysis showed that the genome size increases with addition of new genes, which is suggest that *M. tuberculosis* has an open pan-genome. This shows that bacteria are capable of frequent horizontal gene transfers, which will help it to survive in diverse environmental conditions, including selective antibiotic pressures. In another study conducted on 96 strains form *M. tuberculosis* complex (MTBC), MTBC is estimated to have an open pangenome (Periwal, et al., 2015).

Phylogenetic analysis is significant in understanding the relationships of the strains(Baum et al., 2013). Phylogenetic analysis based on core genome of strains describes the origin, spread and migration pattern of the *M. tuberculosis* strains included in the study. It is highly suggestive of specific geographical distribution of core genes. The strains belonging to South East Asia region fall in a same clad, showing the presence of specific genes in these strains which help these to persist and survive in these environmental pressures. The question that why certain areas have

high burden of tuberculosis (even multi, extremely, pan drug resistant *M. tuberculosis*) can be answered on understanding of these phylogenies and core genes annotations (Mehra, et al., 2013). The strains belonging to South East Asia region that fall in same clad split in numerous smaller clads, and this can be due to genomic diversity within closely related strains. A study conducted by Niemann et al in 2009, showed that *M. tuberculosis* isolates exhibiting identical DNA fingerprinting patterns can harbor substantial genomic diversity (Neimann et al. 2009). The smaller arm of the phylogenetic tree represents the strains mostly from USA, and they all make a separate clad. This signifies the uniqueness of these strains on basis of core genes. The bigger arm of the tree harbors strains from Southeast Asia region (China, India, Hong Kong etc). Further exploration of these geographically based core genes will facilitate in understand the antibiotic resistant and virulence genes, subsequent the survival of TB in these area(Espinal et al., 2001).

Core genomes obtained from 47 strains was analyzed for the presence of antibiotic resistance genes. Very interesting findings were observed. The genes namely drrA, drrB, drrC, mtrA and efpA are seen which have a role in multidrug efflux pump, leading to multi drug resistance strains. mfpA, gyrA and mfd are found within core genome which confer resistance to fluroquinolones. Now this is alarming finding, as florouquinolones are second line antibiotics against MTB. The presence of these genes in core genome explains the phenomenon of extreme drug resistance in MTB(4). Few more genes namely abcA and tap genes are involved in resistance to tetracycline. ksA mutant, AAC(2')-Ic, alaS, ileS, murA and mutant embC genes are found out which confer resistance to isoniazid, aminoglycoside, aminocoumarin, mupirocin, fosfomycin and ethambutol, respectively. Erm(37) confer resistance to lincosamide , streptogramin and macrolide. Isoniazid and ethambutol are first line of choices against MTB,

when MTB confers resistance to isoniazid it is considered as MDR. Fluroquinolones, aminoglycosides are reserved for such MDR cases(Blanchard et al., 1996). But presence of resistance genes against these both first and second line drugs in core genome is very alarming and unfortunate. As our core genome comprised genes from various geographical locations, which emphasize that these antibiotic resistance genes are prevalent in a large number.

In order to tackle *M. tuberculosis*, vaccination is considered to be more reliable due to widespread antibiotic resistance. Currently the vaccine used worldwide against TB is the Bacillus Calmette–Guérin (BCG) which is found to have effectiveness in infants and young children but is not found to be effective in controlling TB epidemic globally (Andersen et al., 2005). Thus, developing a new vaccine with improved efficacy is desirable. The conventional approaches for the vaccine candidate identification are discouraged due to being cumbersome and costly procedure. In contrast, developing vaccine through *in silico* approaches of reverse vaccinology (RV) is alternative feasible approach, as the sub unit vaccine made through this process are efficient in generating (Sette et al., 2010). These approaches are sequence based and their use ensures the vaccine candidates reliability as results are obtained after passing them through various filters and search tools. Combination of various approaches were followed for the identification of vaccine candidates in this study again *M. tuberculosis*. Core proteome of the 47 strains included in the study was obtain to identify the proteins which are conserved among all of these strains. We have then selected those proteins from the core proteins which are of extracellular in origin or are found on the surface of the bacteria. Then these proteins were passed through parameters such as virulence, non-host homologs, trans-membrane helices and essential extracting 8 core proteins which can be considered as *M. tuberculosis* potential vaccine candidates. These include Esterase, Secreted antigen 85-C (85C), PPE family protein, ESX

conserved component 5, lysine-N-oxygenase, ESX-2 secretion system 2, Exported repetitive partial and thiol peroxidase. In another study conducted by Zvi et al, a different methodology was adapted to search T-cell antigens against *M. tuberculosis*. They have done comprehensive *in silico* and literature-based analysis which led to the identification of 45 top-hits antigens, out of the 3989 ORF products of the whole genome. Their list contains three antigens that are also found out in our study, namely Esterase, PPE family protein and ESX-2 secretion system 2 (Zvi et al., 2008).

Prioritized proteins were found to have significant role in various biological process including Oxidation-reduction process, Glycolipid biosynthetic process, Lipid transport, Cell redox homeostasis, Cellular oxidant detoxification, Transport, Response to oxidative stress, pathogenesis, growth, evasion or tolerance by symbiont of host-produced nitric oxide, response to antibiotics, mycolate cell wall layer assembly, host tissue attachment, growth of symbiont in host cell and response to nitrosative stress. Apart from involvement in biological process, analysis of molecular function of the prioritized proteins shows that they carry out trehalose O-mycolyltransferase activity, thioredoxin peroxidase activity, disulfide oxidoreductase activity, L-lysine 6-monooxygenase (NADPH) activity, transferase activity, transferring acyl groups, protein binding, diacylglycerol O-acyltransferase activity, transferase activity/transferring acyl groups other than amino-acyl groups, ATP binding, hydrolase activity, peroxidase activity, peroxiredoxin activity, thioredoxin peroxidase activity and short-chain carboxylesterase activity.

In development of vaccine effective against tuberculosis, recently considerable progress have been made. Still challenges exist in methodology of developing effective and safe vaccine against *M. tuberculosis.* Comprehensive analysis like ours which includes analysis such as

resistome, immunoproteomic, reverse vaccinology, comparative genomics and pan-genomics would contribute to realistic design of vaccine against *M. tuberculosis.*

5.2 Conclusion

Potential candidates for vaccine against *M. tuberculosis* were found in this study following novel approaches of reverse vaccinology, proteomics, core genomics and pan genomics. Eight putative antigens were revealed after extensive analysis of the all available completely sequenced genomes of *M. tuberculosis*. These antigens are named as Esterase, Secreted antigen 85-C (85C), PPE family protein, ESX conserved component 5, lysine-N-oxygenase, ESX-2 secretion system 2, Exported repetitive partial and thiol peroxidase. Besides, core genome is found to carry significant antibiotic resistance gene in it, which is an alarming phenomenon. These candidate proteins need to be further characterized and its immunology need to be studied in animal models.

REFERENCES

Barry, C. E., Boshoff, H. I., Dartois, V., Dick, T., Ehrt, S., Flynn, J., ... & Young, D. (2009). The spectrum of latent tuberculosis: rethinking the biology and intervention strategies. Nature Reviews Microbiology, 7(12), 845-855.

Esmail, H., Barry, C. E., Young, D. B., & Wilkinson, R. J. (2014). The ongoing challenge of latent tuberculosis. Phil. Trans. R. Soc. B, 369(1645), 20130437.

Comas, I., Coscolla, M., Luo, T., Borrell, S., Holt, K. E., Kato-Maeda, M., ... & Yeboah-Manu, D. (2013). Out-of-Africa migration and Neolithic coexpansion of Mycobacterium tuberculosis with modern humans. Nature genetics, 45(10), 1176-1182.

Warner, D. F., Koch, A., & Mizrahi, V. (2015). Diversity and disease pathogenesis in Mycobacterium tuberculosis. Trends in microbiology, 23(1), 14-21.

Reed, M. B., Domenech, P., Manca, C., Su, H., Barczak, A. K., Kreiswirth, B. N., ... & Barry, C. E. (2004). A glycolipid of hypervirulent tuberculosis strains that inhibits the innate immune response. Nature, 431(7004), 84-87.

Gagneux, S., DeRiemer, K., Van, T., Kato-Maeda, M., De Jong, B. C., Narayanan, S., ... & Hilty, M. (2006). Variable host–pathogen compatibility in Mycobacterium

tuberculosis. Proceedings of the National academy of Sciences of the United States of America, 103(8), 2869-2873.

Albanna, A. S., Reed, M. B., Kotar, K. V., Fallow, A., McIntosh, F. A., Behr, M. A., & Menzies, D. (2011). Reduced transmissibility of East African Indian strains of Mycobacterium tuberculosis. PloS one, 6(9), e25075.

Fenner, L., Gagneux, S., Helbling, P., Battegay, M., Rieder, H. L., Pfyffer, G. E., ... & Dolina, M. (2012). Mycobacterium tuberculosis transmission in a country with low tuberculosis incidence: role of immigration and HIV infection. Journal of clinical microbiology, 50(2), 388-395.

Lee, R. S., Radomski, N., Proulx, J. F., Levade, I., Shapiro, B. J., McIntosh, F., ... & Behr, M. A. (2015). Population genomics of Mycobacterium tuberculosis in the Inuit. Proceedings of the National Academy of Sciences, 112(44), 13609-13614.

Morrison, J., Pai, M., & Hopewell, P. C. (2008). Tuberculosis and latent tuberculosis infection in close contacts of people with pulmonary tuberculosis in low-income and middle-income countries: a systematic review and meta-analysis. The Lancet infectious diseases, 8(6), 359-368.

Cobat, A., Gallant, C. J., Simkin, L., Black, G. F., Stanley, K., Hughes, J., ... & Boland-Auge, A. (2009). Two loci control tuberculin skin test reactivity in an area hyperendemic for tuberculosis. The Journal of experimental medicine, 206(12), 2583-2591.

Rangaka, M. X., Wilkinson, K. A., Glynn, J. R., Ling, D., Menzies, D., Mwansa-Kambafwile, J. & Pai, M. (2012). Predictive value of interferon-γ release assays for incident active

tuberculosis: a systematic review and meta-analysis. The Lancet infectious diseases, 12(1), 45-55.

Orme, I. M., Robinson, R. T., & Cooper, A. M. (2015). The balance between protective and pathogenic immune responses in the TB-infected lung. Nature immunology, 16(1), 57-63.

Watford, W. T., Wright, J. R., Hester, C. G., Jiang, H., & Frank, M. M. (2001). Surfactant protein A regulates complement activation. The Journal of Immunology, 167(11), 6593-6600.

Ferguson, J. S., Voelker, D. R., McCormack, F. X., & Schlesinger, L. S. (1999). Surfactant protein D binds to Mycobacterium tuberculosis Bacilli and Lipoarabinomannan via carbohydrate-lectin interactions resulting in reduced phagocytosis of the bacteria by macrophages1. The Journal of immunology, 163(1), 312-321.

Russell, D. G. (2011). Mycobacterium tuberculosis and the intimate discourse of a chronic infection. Immunological reviews, 240(1), 252-268.

Houben, D., Demangel, C., Van Ingen, J., Perez, J., Baldeón, L., Abdallah, A. M., ... & Van Der Laan, T. (2012). ESX-1-mediated translocation to the cytosol controls virulence of mycobacteria. Cellular microbiology, 14(8), 1287-1298.

van der Wel, N., Hava, D., Houben, D., Fluitsma, D., van Zon, M., Pierson, J., ... & Peters, P. J. (2007). M. tuberculosis and M. leprae translocate from the phagolysosome to the cytosol in myeloid cells. Cell, 129(7), 1287-1298.

Simeone, R., Majlessi, L., Enninga, J., & Brosch, R. (2016). Perspectives on mycobacterial vacuole-to-cytosol translocation: the importance of cytosolic access. Cellular microbiology.

Russell, D. G. (2016). The ins and outs of the Mycobacterium tuberculosis-containing vacuole. Cellular microbiology.

Manca, C., Tsenova, L., Bergtold, A., Freeman, S., Tovey, M., Musser, J. M., ... & Kaplan, G. (2001). Virulence of a Mycobacterium tuberculosis clinical isolate in mice is determined by failure to induce Th1 type immunity and is associated with induction of IFN-α/β. Proceedings of the National Academy of Sciences, 98(10), 5752-5757.

Mayer-Barber, K. D., Andrade, B. B., Oland, S. D., Amaral, E. P., Barber, D. L., Gonzales, J., ... & Yuan, X. (2014). Host-directed therapy of tuberculosis based on interleukin-1 and type I interferon crosstalk. Nature, 511(7507), 99-103.

Stanley, S. A., Johndrow, J. E., Manzanillo, P., & Cox, J. S. (2007). The Type I IFN response to infection with Mycobacterium tuberculosis requires ESX-1-mediated secretion and contributes to pathogenesis. The Journal of Immunology, 178(5), 3143-3152.

Pandey, A. K., Yang, Y., Jiang, Z., Fortune, S. M., Coulombe, F., Behr, M. A., ... & Kelliher, M. A. (2009). NOD2, RIP2 and IRF5 play a critical role in the type I interferon response to Mycobacterium tuberculosis. PLoS Pathog, 5(7), e1000500.

Manzanillo, P. S., Shiloh, M. U., Portnoy, D. A., & Cox, J. S. (2012). Mycobacterium tuberculosis activates the DNA-dependent cytosolic surveillance pathway within macrophages. Cell host & microbe, 11(5), 469-480.

Kaufmann, S. H., & Dorhoi, A. (2016). Molecular Determinants in Phagocyte-Bacteria Interactions. Immunity, 44(3), 476-491.

Schaible, U. E., Winau, F., Sieling, P. A., Fischer, K., Collins, H. L., Hagens, K., ... & Kaufmann, S. H. (2003). Apoptosis facilitates antigen presentation to T lymphocytes through MHC-I and CD1 in tuberculosis. Nature medicine, 9(8), 1039-1046.

Behar, S. M., Divangahi, M., & Remold, H. G. (2010). Evasion of innate immunity by Mycobacterium tuberculosis: is death an exit strategy?. Nature Reviews Microbiology, 8(9), 668-674.

Pai, M., Behr, Marcel A. Behr, Dowdy, D. , Keertan Dheda, Maziar Divangahi, Catharina C. Boehme, Ann Ginsberg, Soumya Swaminathan, Melvin Spigelman, Haileyesus Getahun, Dick Menzies & Mario Raviglione (2016). Tuberculosis, Nature Reviews Disease Primers 2, Article number: 16076 (2016) doi:10.1038/nrdp.2016.76

Wolf, A. J., Desvignes, L., Linas, B., Banaiee, N., Tamura, T., Takatsu, K., & Ernst, J. D. (2008). Initiation of the adaptive immune response to Mycobacterium tuberculosis depends on antigen production in the local lymph node, not the lungs. The Journal of experimental medicine, 205(1), 105-115.

Samstein, M., Schreiber, H. A., Leiner, I. M., Sušac, B., Glickman, M. S., & Pamer, E. G. (2013). Essential yet limited role for CCR2+ inflammatory monocytes during Mycobacterium tuberculosis-specific T cell priming. Elife, 2, e01086.

Chackerian, A. A., Alt, J. M., Perera, T. V., Dascher, C. C., & Behar, S. M. (2002). Dissemination of Mycobacterium tuberculosis is influenced by host factors and precedes the initiation of T-cell immunity. Infection and immunity, 70(8), 4501-4509.

Sonnenberg, P., Glynn, J. R., Fielding, K., Murray, J., Godfrey-Faussett, P., & Shearer, S. (2005). How soon after infection with HIV does the risk of tuberculosis start to increase? A retrospective cohort study in South African gold miners. Journal of Infectious Diseases, 191(2), 150-158.

Antonelli, L. R., Rothfuchs, A. G., Gonçalves, R., Roffê, E., Cheever, A. W., Bafica, A., ... & Sher, A. (2010). Intranasal Poly-IC treatment exacerbates tuberculosis in mice through the pulmonary recruitment of a pathogen-permissive monocyte/macrophage population. The Journal of clinical investigation, 120(5), 1674-1682.

Lin, P. L., Ford, C. B., Coleman, M. T., Myers, A. J., Gawande, R., Ioerger, T., ... & Flynn, J. L. (2014). Sterilization of granulomas is common in active and latent tuberculosis despite within-host variability in bacterial killing. Nature medicine, 20(1), 75-79.

Marakalala, M. J., Raju, R. M., Sharma, K., Zhang, Y. J., Eugenin, E. A., Prideaux, B., ... & Eum, S. Y. (2016). Inflammatory signaling in human tuberculosis granulomas is spatially organized. Nature medicine.

Tobin, D. M., Roca, F. J., Oh, S. F., McFarland, R., Vickery, T. W., Ray, J. P., ... & Vary, J. C. (2012). Host genotype-specific therapies can optimize the inflammatory response to mycobacterial infections. Cell, 148(3), 434-446.

Lalvani, A., Behr, M. A., & Sridhar, S. (2012). Innate immunity to TB: a druggable balancing act. Cell, 148(3), 389-391.

Bustamante, J., Boisson-Dupuis, S., Abel, L., & Casanova, J. L. (2014, December). Mendelian susceptibility to mycobacterial disease: genetic, immunological, and clinical features of

inborn errors of IFN-γ immunity. In Seminars in immunology (Vol. 26, No. 6, pp. 454-470). Academic Press.

Abel, L., El-Baghdadi, J., Bousfiha, A. A., Casanova, J. L., & Schurr, E. (2014). Human genetics of tuberculosis: a long and winding road. Phil. Trans. R. Soc. B, 369(1645), 20130428.

Daniels, M., & Hill, A. B. (1952). Chemotherapy of pulmonary tuberculosis in young adults. British Medical Journal, 1(4769), 1162.

Nebenzahl-Guimaraes, H., Jacobson, K. R., Farhat, M. R., & Murray, M. B. (2013). Systematic review of allelic exchange experiments aimed at identifying mutations that confer drug resistance in Mycobacterium tuberculosis. Journal of antimicrobial chemotherapy, dkt358.

Pai, M., Denkinger, C. M., Kik, S. V., Rangaka, M. X., Zwerling, A., Oxlade, O., ... & Banaei, N. (2014). Gamma interferon release assays for detection of Mycobacterium tuberculosis infection. Clinical microbiology reviews, 27(1), 3-20.

World Health Organization. (2014). Drug-resistant TB: surveillance and response: supplement to global tuberculosis report 2014.

Menzies, D., Gardiner, G., Farhat, M., Greenaway, C., & Pai, M. (2008). Thinking in three dimensions: a web-based algorithm to aid the interpretation of tuberculin skin test results. The International Journal of Tuberculosis and Lung Disease, 12(5), 498-505.

Farhat, M., Greenaway, C., Pai, M., & Menzies, D. (2006). False-positive tuberculin skin tests: what is the absolute effect of BCG and non-tuberculous mycobacteria?[Review Article]. The International Journal of Tuberculosis and Lung Disease, 10(11), 1192-1204.

Pai, M., & Sotgiu, G. (2016). Diagnostics for latent TB infection: incremental, not transformative progress. European Respiratory Journal, 47(3), 704-706.

Pai, M., Riley, L. W., & Colford, J. M. (2004). Interferon-γ assays in the immunodiagnosis of tuberculosis: a systematic review. The Lancet infectious diseases, 4(12), 761-776.

Sørensen, A. L., Nagai, S., Houen, G., Andersen, P., & Andersen, A. B. (1995). Purification and characterization of a low-molecular-mass T-cell antigen secreted by Mycobacterium tuberculosis. Infection and immunity, 63(5), 1710-1717.

Andersen, P., Munk, M. E., Pollock, J. M., & Doherty, T. M. (2000). Specific immune-based diagnosis of tuberculosis. The Lancet, 356(9235), 1099-1104.

Pai, M., Denkinger, C. M., Kik, S. V., Rangaka, M. X., Zwerling, A., Oxlade, O., ... & Banaei, N. (2014). Gamma interferon release assays for detection of Mycobacterium tuberculosis infection. Clinical microbiology reviews, 27(1), 3-20.

[No authors listed.] Global routine vaccination coverage, 2014. Wkly Epidemiol. Rec. 90, 617–623 (2015).

Zwerling, A., Behr, M. A., Verma, A., Brewer, T. F., Menzies, D., & Pai, M. (2011). The BCG World Atlas: a database of global BCG vaccination policies and practices. PLoS Med, 8(3), e1001012.

Mangtani, P., Abubakar, I., Ariti, C., Beynon, R., Pimpin, L., Fine, P. E., ... & Sterne, J. A. (2014). Protection by BCG vaccine against tuberculosis: a systematic review of randomized controlled trials. Clinical infectious diseases, 58(4), 470-480.

Roy, A., Eisenhut, M., Harris, R. J., Rodrigues, L. C., Sridhar, S., Habermann, S., ... & Abubakar, I. (2014). Effect of BCG vaccination against Mycobacterium tuberculosis infection in children: systematic review and meta-analysis.

Trunz, B. B., Fine, P. E. M., & Dye, C. (2006). Effect of BCG vaccination on childhood tuberculous meningitis and miliary tuberculosis worldwide: a meta-analysis and assessment of cost-effectiveness. The Lancet, 367(9517), 1173-1180.

Barreto, M. L., Pereira, S. M., Pilger, D., Cruz, A. A., Cunha, S. S., Sant'Anna, C., ... & Rodrigues, L. C. (2011). Evidence of an effect of BCG revaccination on incidence of tuberculosis in school-aged children in Brazil: second report of the BCG-REVAC cluster-randomised trial. Vaccine, 29(31), 4875-4877.

[No authors listed.] Fifteen year follow up of trial of BCG vaccines in south India for tuberculosis prevention. Tuberculosis Research Centre (ICMR), Chennai. Indian J. Med. Res. 110, 56–69 (1999).

Abubakar, I., Pimpin, L., Ariti, C., Beynon, R., Mangtani, P., Sterne, J. A. C., ... & Watson, J. M. (2013). Systematic review and meta-analysis of the current evidence on the duration of protection by bacillus Calmette–Guerin vaccination against tuberculosis.

Ellis, R. D., Hatherill, M., Tait, D., Snowden, M., Churchyard, G., Hanekom, W., ... & Ginsberg, A. M. (2015). Innovative clinical trial designs to rationalize TB vaccine development. Tuberculosis, 95(3), 352-357.

AERAS. TB vaccine research and development: a business case for investment. AERAS (2014)

Knight, G. M., Griffiths, U. K., Sumner, T., Laurence, Y. V., Gheorghe, A., Vassall, A., ... & White, R. G. (2014). Impact and cost-effectiveness of new tuberculosis vaccines in low- and middle-income countries. Proceedings of the National Academy of Sciences, 111(43), 15520-15525.

Hyatt, D., Chen, G. L., LoCascio, P. F., Land, M. L., Larimer, F. W., & Hauser, L. J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC bioinformatics, 11(1), 1.

Medini, D., Donati, C., Tettelin, H., Masignani, V., & Rappuoli, R. (2005). The microbial pan-genome. Current opinion in genetics & development, 15(6), 589-594.

Ali, A., Soares, S. C., Santos, A. R., Guimarães, L. C., Barbosa, E., Almeida, S. S., ... & Hassan, S. S. (2012). Campylobacter fetus subspecies: Comparative genomics and prediction of potential virulence targets. Gene, 508(2), 145-156.

Lukjancenko, O., Ussery, D. W., & Wassenaar, T. M. (2012). Comparative genomics of Bifidobacterium, Lactobacillus and related probiotic genera. Microbial ecology, 63(3), 651-673.

Ussery, D. W., Wassenaar, T. M., & Borini, S. (2009). Microbial communities: core and pan-genomics. In Computing for Comparative Microbial Genomics (pp. 213-228). Springer London.

Trost, E., Blom, J., de Castro Soares, S., Huang, I. H., Al-Dilaimi, A., Schröder, J., ... & Azevedo, V. (2012). Pangenomic study of Corynebacterium diphtheriae that provides insights into the genomic diversity of pathogenic isolates from cases of classical diphtheria, endocarditis, and pneumonia. Journal of bacteriology, 194(12), 3199-3215.

Ali, A., Soares, S. C., Barbosa, E., Santos, A. R., Barh, D., Bakhtiar, S. M., ... & Azevedo, V. (2013). Microbial comparative genomics: an overview of tools and insights into the genus Corynebacterium. Journal of Bacteriology & Parasitology, 2013.

Snipen, L., Almøy, T., & Ussery, D. W. (2009). Microbial comparative pan-genomics using binomial mixture models. BMC genomics, 10(1), 385.

Luo, H., Lin, Y., Gao, F., Zhang, C. T., & Zhang, R. (2014). DEG 10, an update of the database of essential genes that includes both protein-coding genes and noncoding genomic elements. Nucleic acids research, 42(D1), D574-D580.

Galperin, M. Y., & Koonin, E. V. (1999). Searching for drug targets in microbial genomes. Current opinion in biotechnology, 10(6), 571-578.

Naz, A., Awan, F. M., Obaid, A., Muhammad, S. A., Paracha, R. Z., Ahmad, J., & Ali, A. (2015). Identification of putative vaccine candidates against Helicobacter pylori. Helicobacter, 3, 1.

Chen, L., Xiong, Z., Sun, L., Yang, J., & Jin, Q. (2011). VFDB 2012 update: toward the genetic diversity and molecular evolution of bacterial virulence factors. Nucleic acids research, gkr989.

Zhou, C. E., Smith, J., Lam, M., Zemla, A., Dyer, M. D., & Slezak, T. (2007). MvirDB—a microbial database of protein toxins, virulence factors and antibiotic resistance genes for bio-defence applications. Nucleic acids research, 35(suppl 1), D391-D394.

Gardy, J. L., Laird, M. R., Chen, F., Rey, S., Walsh, C. J., Ester, M., & Brinkman, F. S. (2005). PSORTb v. 2.0: expanded prediction of bacterial protein subcellular localization and insights gained from comparative proteome analysis. Bioinformatics, 21(5), 617-623.

Zagursky, R. J., Olmsted, S. B., Russell, D. P., & Wooters, J. L. (2003). Bioinformatics: how it is being used to identify bacterial vaccine candidates. Expert review of vaccines, 2(3), 417-436.

Tusnady, G. E., & Simon, I. (2001). The HMMTOP transmembrane topology prediction server. Bioinformatics, 17(9), 849-850.

Krogh, A., Larsson, B., Von Heijne, G., & Sonnhammer, E. L. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. Journal of molecular biology, 305(3), 567-580.

Sette, A., Vitiello, A., Reherman, B., Fowler, P., Nayersina, R., Kast, W. M., ... & Sidney, J. (1994). The relationship between class I binding affinity and immunogenicity of potential cytotoxic T cell epitopes. The Journal of Immunology, 153(12), 5586-5592.

Brusic, V., & Petrovsky, N. (2005). Immunoinformatics and its relevance to understanding human immune disease. Expert review of clinical immunology, 1(1), 145-157.

Provenzano, M., Panelli, M. C., Mocellin, S., Bracci, L., Sais, G., Stroncek, D. F., ... & Marincola, F. M. (2006). MHC–peptide specificity and T-cell epitope mapping: where immunotherapy starts. Trends in molecular medicine, 12(10), 465-472.

Saha, S., & Raghava, G. P. (2007). Prediction methods for B-cell epitopes. Immunoinformatics: Predicting Immunogenicity In Silico, 387-394.

Singh, H., & Raghava, G. P. S. (2003). ProPred1: prediction of promiscuous MHC Class-I binding sites. Bioinformatics, 19(8), 1009-1014.

Chaplin, D. D. (2003). 1. Overview of the immune response. Journal of Allergy and Clinical Immunology, 111(2), S442-S459.

Petersen, B., Petersen, T. N., Andersen, P., Nielsen, M., & Lundegaard, C. (2009). A generic method for assignment of reliability scores applied to solvent accessibility predictions. BMC structural biology, 9(1), 1.

Guan, P., Doytchinova, I. A., Zygouri, C., & Flower, D. R. (2003). MHCPred: a server for quantitative prediction of peptide–MHC binding. Nucleic acids research, 31(13), 3621-3624.

Southwood, S., Sidney, J., Kondo, A., del Guercio, M. F., Appella, E., Hoffman, S., ... & Sette, A. (1998). Several common HLA-DR types share largely overlapping peptide binding repertoires. The Journal of Immunology, 160(7), 3363-3373.

Hosseingholi, E. Z., Rasooli, I., & Gargari, S. L. M. (2014). In silico analysis of Acinetobacter baumannii phospholipase D as a subunit vaccine candidate. Acta biotheoretica, 62(4), 455-478.

Rakesh, S., Pradhan, D., & Umamaheswari, A. (2009). In silico approach for future development of subunit vaccines against Leptospira interrogans serovar Lai. Int J Bioinformatics Res, 1, 85-92.

Blythe, M. J., Doytchinova, I. A., & Flower, D. R. (2002). JenPep: a database of quantitative functional peptide data for immunology. Bioinformatics, 18(3), 434-439.

Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S. E., Wilkins, M. R., Appel, R. D., & Bairoch, A. (2005). Protein identification and analysis tools on the ExPASy server (pp. 571-607). Humana Press.

Barh, D., Barve, N., Gupta, K., Chandra, S., Jain, N., Tiwari, S., ... & Almeida, S. (2013). Exoproteome and secretome derived broad spectrum novel drug and vaccine candidates in Vibrio cholerae targeted by Piper betel derived compounds. PloS one, 8(1), e52773.

Garg, A., & Gupta, D. (2008). VirulentPred: a SVM based prediction method for virulent proteins in bacterial pathogens. BMC bioinformatics, 9(1), 1.

Doytchinova, I. A., & Flower, D. R. (2007). VaxiJen: a server for prediction of protective antigens, tumour antigens and subunit vaccines. BMC bioinformatics, 8(1), 1.

Hassan, A., Naz, A., Obaid, A., Paracha, R. Z., Naz, K., Awan, F. M., ... & Ali, A. (2016). Pangenome and immuno-proteomics analysis of Acinetobacter baumannii strains revealed the core peptide vaccine targets. BMC genomics, 17(1), 732.

Workbench CG. v3. 6.(2010). Now new version can be available at http://www.clcbio.com/products/clcgenomicsworkbench

Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N., & Sternberg, M. J. (2015). The Phyre2 web portal for protein modeling, prediction and analysis. Nature protocols, 10(6), 845-858.

Gabdoulline, R. R., Hoffmann, R., Leitner, F., & Wade, R. C. (2003). ProSAT: functional annotation of protein 3D structures. Bioinformatics, 19(13), 1723-1725

Conesa, A., Götz, S., García-Gómez, J. M., Terol, J., Talón, M., & Robles, M. (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics, 21(18), 3674-3676

Szklarczyk, D., Franceschini, A., Kuhn, M., Simonovic, M., Roth, A., Minguez, P., ... & Jensen, L. J. (2011). The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. Nucleic acids research, 39(suppl 1), D561-D568.

Bultinck, J., Lievens, S., & Tavernier, J. (2012). Protein-protein interactions: network analysis and applications in drug discovery. Current pharmaceutical design, 18(30), 4619-4629.

Alonso, A., Sanchez, P., & Martinez, J. L. (2001). Environmental selection of antibiotic resistance genes. Environmental microbiology, 3(1), 1-9.

Yu, C. S., Cheng, C. W., Su, W. C., Chang, K. C., Huang, S. W., Hwang, J. K., & Lu, C. H. (2014). CELLO2GO: a web server for protein subCELlular LOcalization prediction with functional gene ontology annotation. PloS one, 9(6), e99368.

Yu, C. S., Lin, C. J., & Hwang, J. K. (2004). Predicting subcellular localization of proteins for Gram-negative bacteria by support vector machines based on n-peptide compositions. Protein Science, 13(5), 1402-1406.

Pan, A., Lahiri, C., Rajendiran, A., & Shanmugham, B. (2015). Computational analysis of protein interaction networks for infectious diseases. Briefings in bioinformatics, bbv059.

Ali, A., Naz, A., Soares, S. C., Bakhtiar, M., Tiwari, S., Hassan, S. S., ... & Figueiredo, H. C. P. (2015). Pan-genome analysis of human gastric pathogen H. pylori: comparative genomics

and pathogenomics approaches to identify regions associated with pathogenicity and prediction of potential core therapeutic targets. BioMed research international, 2015.

Zhang, R., Ou, H. Y., & Zhang, C. T. (2004). DEG: a database of essential genes. Nucleic acids research, 32(suppl 1), D271-D272.

Judson, N., & Mekalanos, J. J. (2000). Transposon-based approaches to identify essential bacterial genes. Trends in microbiology, 8(11), 521-526.

Butt, A. M., Nasrullah, I., Tahir, S., & Tong, Y. (2012). Comparative genomics analysis of Mycobacterium ulcerans for the identification of putative essential genes and therapeutic candidates. PloS one, 7(8), e43080.

Lapierre, P., & Gogarten, J. P. (2009). Estimating the size of the bacterial pan-genome. Trends in genetics, 25(3), 107-110.

Dahlbäck, M., Rask, T. S., Andersen, P. H., Nielsen, M. A., Ndam, N. T., Resende, M., ... & Pedersen, A. G. (2006). Epitope mapping and topographic analysis of VAR2CSA DBL3X involved in P. falciparum placental sequestration. PLoS Pathog, 2(11), e124.

Pe'er, D., Regev, A., Elidan, G., & Friedman, N. (2001). Inferring subnetworks from perturbed expression profiles. Bioinformatics, 17(suppl 1), S215-S224.

Gardy, J. L., Laird, M. R., Chen, F., Rey, S., Walsh, C. J., Ester, M., & Brinkman, F. S. (2005). PSORTb v. 2.0: expanded prediction of bacterial protein subcellular localization and insights gained from comparative proteome analysis. Bioinformatics, 21(5), 617-623.

Götz, S., García-Gómez, J. M., Terol, J., Williams, T. D., Nagaraj, S. H., Nueda, M. J., ... & Conesa, A. (2008). High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic acids research, 36(10), 3420-3435.

Fu, L. M., & Fu-Liu, C. S. (2002). Is Mycobacterium tuberculosis a closer relative to Gram-positive or Gram–negative bacterial pathogens?. Tuberculosis, 82(2), 85-90.

Tu, H. Z., Chang, S. H., Huaug, T. S., Huaug, W. K., Liu, Y. C., & Lee, S. S. J. (2003). Microscopic morphology in smears prepared from MGIT broth medium for rapid presumptive identification of Mycobacterium tuberculosis complex, Mycobacterium avium complex and Mycobacterium kansasii. Annals of Clinical & Laboratory Science, 33(2), 179-183.

Hui-Zin Tu1, Shu-Huei Chang1, Tsi-Shu Huaug1,3, Wen-Kuei Huaug1,3, Yung-Ching Liu1,2 and Susan Shin-Jung Lee

Attorri, S., Dunbar, S., & Clarridge, J. E. (2000). Assessment of Morphology for Rapid Presumptive Identification of Mycobacterium tuberculosis andMycobacterium kansasii. Journal of clinical microbiology, 38(4), 1426-1429.

Alva, A., Aquino, F., Gilman, R. H., Olivares, C., Requena, D., Gutiérrez, A. H., ... & Moore, D. A. (2013). Morphological characterization of Mycobacterium tuberculosis in a MODS culture for an automatic diagnostic through pattern recognition. PloS one, 8(12), e82809.

Brennan, P. J. (2003). Structure, function, and biogenesis of the cell wall of Mycobacterium tuberculosis. Tuberculosis, 83(1), 91-97.

Kaur, D., Guerin, M. E., Škovierová, H., Brennan, P. J., & Jackson, M. (2009). Biogenesis of the cell wall and other glycoconjugates of Mycobacterium tuberculosis. Advances in applied microbiology, 69, 23-78.

Vernikos G, Medini D, Riley DR, Tettelin H. Ten years of pan-genome analyses. Current opinion in microbiology. 2015;23:148-54.

Baum DA, Smith SD. Tree thinking: an introduction to phylogenetic biology: Roberts; 2013.

Smith NH, Hewinson RG, Kremer K, Brosch R, Gordon SV. Myths and misconceptions: the origin and evolution of Mycobacterium tuberculosis. Nature Reviews Microbiology. 2009;7(7):537-44.

Espinal MA, Laszlo A, Simonsen L, Boulahbal F, Kim SJ, Reniero A, et al. Global trends in resistance to antituberculosis drugs. New England Journal of Medicine. 2001;344(17):1294-303.

Blanchard JS. Molecular mechanisms of drug resistance in Mycobacterium tuberculosis. Annual review of biochemistry. 1996;65(1):215-39.

Andersen P, Doherty TM. The success and failure of BCG—implications for a novel tuberculosis vaccine. Nature Reviews Microbiology. 2005;3(8):656-62.

Sette A, Rappuoli R. Reverse vaccinology: developing vaccines in the era of genomics. Immunity. 2010;33(4):530-41.

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

Niemann, S., Köser, C. U., Gagneux, S., Plinke, C., Homolka, S., Bignell, H., ... & Kokko-Gonzales, P. (2009). Genomic diversity among drug sensitive and multidrug resistant

isolates of Mycobacterium tuberculosis with identical DNA fingerprints. PloS one, 4(10), e7407.

Periwal, V., Patowary, A., Vellarikkal, S. K., Gupta, A., Singh, M., Mittal, A., ... & Garg, P. (2015). Comparative whole-genome analysis of clinical isolates reveals characteristic architecture of Mycobacterium tuberculosis pangenome. PloS one, 10(4), e0122979.