Journal of
Bioengineering & Biomedical Science

Badapanda et al., J Bioengineer & Biomedical Sci 2016, 6:3
DOI: 10.4172/2155-9538.1000196

Research Article Open Access

Functional Annotation and Epitope Prediction of Hypothetical Proteins of *Mycobacterium tuberculosis* H37Rv: An Immunoinformatics Approach

Badapanda C[1,2*], Sahoo GC[2], Middha A[3], Majhi MC[4] and Nayak R[5]

[1]*Xcelris Labs Ltd., Ahmedabad, India*
[2]*Department of Biotechnology, OPJS University, Churu, Rajasthan, India*
[3]*Department of Pharmacy, OPJS University, Churu, Rajasthan, India*
[4]*Institute of Genomics & Integrated Biology, Environmental Biotechnology Division, New Delhi, India*
[5]*National Institute of Science Education and Research, School of Biological Sciences, Institute of Physics Campus, Bhubaneswar, Odisha, India*

Abstract

High-throughput genome sequencing technologies are revolutionizing bacterial genomics, resulting in the accumulation of 'unknown' or 'hypothetical' or 'conserved hypothetical' genes. However, approximately 40-50% of genes within a genome are often labeled as 'hypothetical' or 'conserved hypothetical' or 'unknown' whose function has not yet been established, inviting the functional annotation of these 'unknown genes'. *Mycobacterium tuberculosis* H37Rv has 3,924 protein coding genes, of which 606 proteins are classified as 'unknown proteins'. We here predict reliable functional annotation by integrating several bioinformatics annotation tools, sequential BLAST homology searches, InterProScan searches, Gene Ontology (GO) mapping, and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis, and established the putative function of 522 proteins with at least some functional annotations. The identified pathways from 'unknown proteins' are mapped to well-known pathways, and would provide many putative targets for the rational design of more effective anti-mycobacterial agents. In this work, we have computationally defined T cell epitopes of proteins of *M. tuberculosis* H37Rv to help in the design of a vaccine with haplotype specificity for a target population. The peptides of *M. tuberculosis* H37Rv which are predicted to bind different HLAs class-I (Human Leukocyte Antigens), do not show similarity with peptides of human proteome. Some of the nonameric peptides are promiscuous in their association with multiple alleles, and are considered for vaccine design because of their relevance in the wider coverage of human population. Altogether, functional annotations performed by integrative bioinformatics approaches should considerably enhance the interpretation of the unknown proteins of this medically important organism.

Keywords: Unknown proteins; *Mycobacterium tuberculosis* H37Rv; Functional annotation; Immunoinformatics; Vaccine; Epitope; Tuberculosis

Introduction

Tuberculosis continues to pose a major challenge in health care and ranks as the second largest killer infectious disease, after the human immune-deficiency virus (HIV) that causes the acquired immunodeficiency syndrome (AIDS) (WHO report, 2012). One in every three people on earth is believed to be infected with *Mycobacterium tuberculosis* [1], leading to almost 9 million cases of active tuberculosis (TB) and approximately 1.4 million TB deaths annually, including almost 990 000 deaths among HIV-negative people and 430 000 deaths among people who were HIV-positive (WHO report, 2012). Patients infected with the human immunodeficiency virus (HIV) are more susceptible to *M. tuberculosis* [2]. The situation is going to get worse by the emergence of multidrug-resistant (MDR), extensively drug-resistant (XDR), and totally drug-resistant (TDR) *M. tuberculosis* strains, which are virtually untreatable [3]. The situation is also exacerbated by the catastrophic nexus between AIDS and TB [4]. It is estimated that 35 million will die of TB between 2000 and 2020 if control and preventive measures are not strengthened (World Health Organization Annual Report, 2000). BCG is the only currently available vaccine for prevention of tuberculosis (TB), which has exhibited considerable variations in efficacy in clinical trials in a geographically distinct population. Although BCG prevents disseminated tuberculosis in new born, it fails to protect against the pulmonary tuberculosis in adults [5]. The failure of BCG is due to the following reasons, firstly BCG lacks important antigen, secondly BCG is not inducing important T-cell response, thirdly environmental Mycobacteria interact with BCG, and lastly BCG efficacy wanes over time [6]. Therefore, it is crucial to improve BCG or to develop a more effective vaccine than *M. bovis* BCG.

Next Generation Sequencing (NGS) methods have been resulted in accumulation of large number of sequences derived from different bacterial strains [7]. However, approximately 40-50% of genes within a genome are labeled as 'unknown', 'hypothetical' or 'conserved hypothetical' or 'orphan' genes, limiting our understanding of virulence and pathogenicity of bacteria species [8]. The genome of *M. tuberculosis* H37Rv strain has been sequenced in the year 1998 and predicted to have 3,924 protein coding sequences and four years after this first submission eighty-two more sequences have been added to it [9]. Nevertheless, 606 proteins (15% of the proteome) were classified as 'unknown proteins' due to lack of any information available to these sequences. Thus, one of the major tasks in the post-genomic era is the genome annotation, assigning functions to these unknown genes. So the first objective of our work is to annotate the 606 unknown proteins of *Mycobacterium tuberculosis* H37Rv to assign their putative functions, and the second objective of our work is to design a rational subunit vaccine with a much higher efficacy than the present available BCG vaccine. A subunit vaccine, consisting of a few key molecules of *M. tuberculosis* that are capable of inducing protective immunity, could have substantial advantages over BCG or other whole bacterial vaccines and also safer.

***Corresponding author:** Chandan Badapanda, Xcelris Labs Ltd., Ahmedabad, Gujrat, India, Tel: +91-7966197777; E-mail:chandan.bioinfo@gmail.com

Received: June 08, 2016; **Accepted:** July 15, 2016; **Published**: July 25, 2016

Citation: Badapanda C, Sahoo GC, Middha A, Majhi MC, Nayak R (2016) Functional Annotation and Epitope Prediction of Hypothetical Proteins of *Mycobacterium tuberculosis* H37Rv: An Immunoinformatics Approach. J Bioengineer & Biomedical Sci 6: 196. doi:10.4172/2155-9538.1000196

Thus, the objective of our second work is to design rational vaccine for tuberculosis following reductionist and integrative approach taking into account the functions of immune system and immunological memory [10,11]. Combined with functional genomics approach as well as immune informatics based peptide candidate identification tools, the successful acceleration of the process of assigning putative functions to these unknown genes have been established. The present work also compliments our previous work performed on PE-PPE protein family and secretory proteins of *M. tuberculosis* H37Rv [12-14]. The present method can also be applied to annotate the unknown proteins from other prokaryotic and eukaryotic organisms [15]. We believe that these unknown proteins, which are mostly restricted to the *Mycobacterium tuberculosis* H37Rv genome, could significantly impact TB control and vaccine development strategies.

Material and Methods

Complete protein sequences of *M. tuberculosis* H37Rv strain were deduced from Sanger database. Firstly, we analyzed 606 proteins belonging to the functional class VI of 'unknown proteins' from *M. tuberculosis* H37Rv. The second objective is to define the MHC class I binding motifs of Mycobacterium proteins for presentation to T cell.

Blast homology searches and sequence annotation

BLAST searches were performed on a local server by using Blastall program at National Center for Biotechnology Information (NCBI). Homology searches (BLASTP) of sequences and functional annotation by Gene Ontology terms, InterPro terms (InterProScan, EBI), enzyme classification codes (EC), and metabolic pathways (KEGG, Kyoto Encyclopedia of Genes and Genomes) were determined using the BLAST2GO software suite v2.3.1 [16]. Sequences were searched against the NCBI non-redundant (nr) protein database using an E-value cut-off of 10-5. The GO data presented represent the level 2 or level 3 or level 5 analysis illustrating general functional categories. Enzyme classification codes, and KEGG metabolic pathway annotations, were generated from the direct mapping of GO terms to their enzyme code equivalents.

Immunoinformatics based T cell epitopes identification

The selection of candidate peptides of *M. tuberculosis* H37Rv can be achieved by application of immunoinformatics tool known as BIMAS (Bioinformatics and Molecular analysis section). All possible overlapping nonamers (9-mer peptides) were generated computationally from each of 606 proteins and analyzed for their potential to bind against 33 different alleles of HLA class I and seven MHC class I alleles from mouse, using a prediction algorithm HLA-BIND. Since the 9- amino acid length peptides were considered to be optimum for HLA class I binding [17], the algorithm was set to generate all possible nonamers. The binding was measured in terms of halftime ($T_{1/2}$) of β_2 microglobulin dissociation rate [18]. A default conservative cutoff value of $T_{1/2} \geq 100$ minutes was chosen in order to obtain high-affinity peptides as well as to pick up the antigenic peptides which are synthesized in tiny amount by the bacteria. The algorithm estimates the binding against 33 HLA class I alleles (which include 9 HLA-A alleles, 20 HLA-B alleles, 4 HLA-C alleles) and also 7 MHC class I alleles of mouse. Only those peptides, which were predicted to bind with any of the 33 alleles, were chosen for further analysis. The predicted peptides to bind to HLA were analyzed against the human proteome in order to screen only the non-self-peptides, as well as to identify self and partially self-peptides which should be excluded from vaccine construct (Figure 1).

Results

For annotation and in silico functional comparisons, all the sequences were subjected to BLAST homology searches followed by

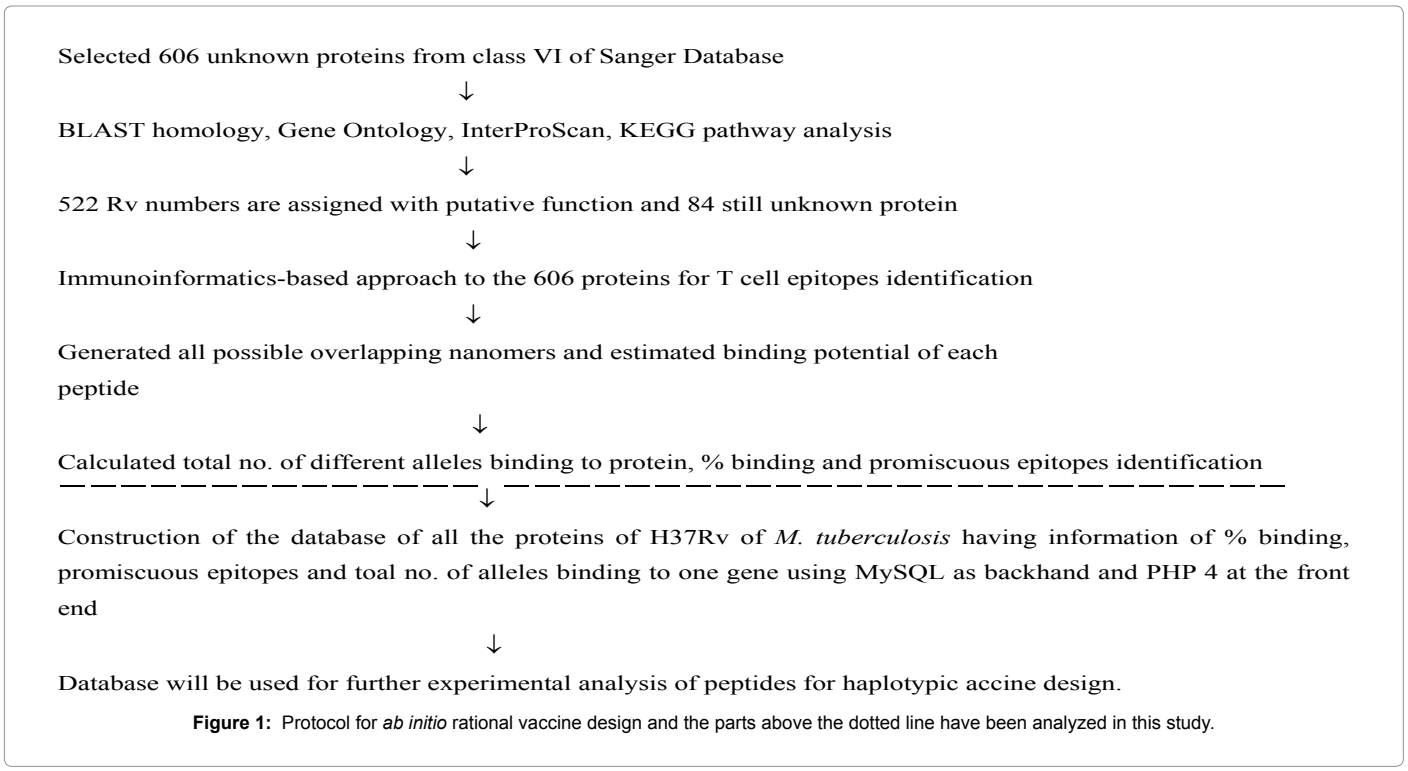

Figure 1: Protocol for *ab initio* rational vaccine design and the parts above the dotted line have been analyzed in this study.

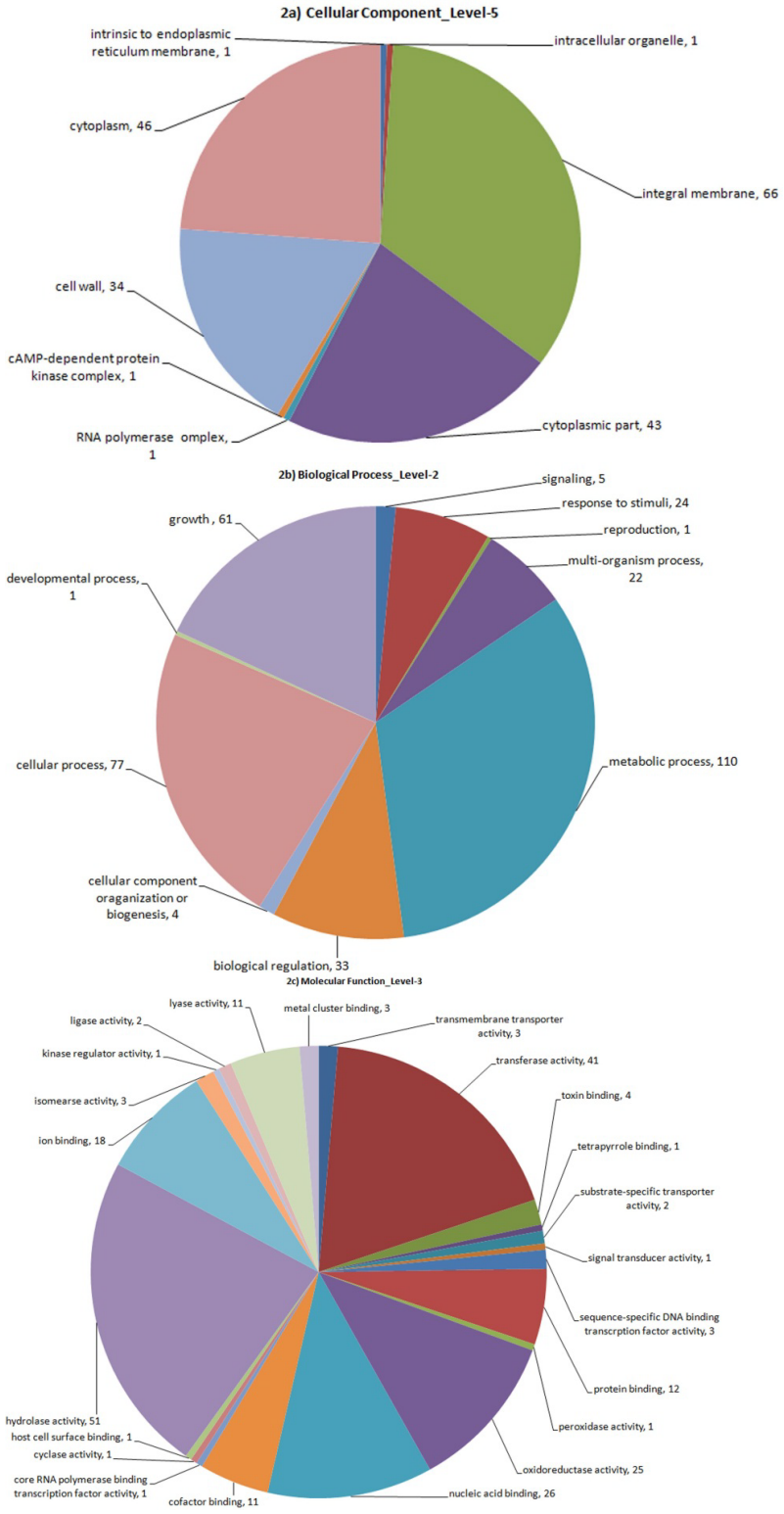

Figure 2: Gene Ontology (GO) analysis of the unknown proteins of Mycobacterium tuberculosis H37Rv for their involvement in 'Cellular Components', 'Biological Processes', and 'Molecular Functions'. (a) Data is represented at level 5 for 'Cellular Components'. (b) Level 2 for 'Biological Processes', and (c) Level 3 for 'Molecular Function' category.

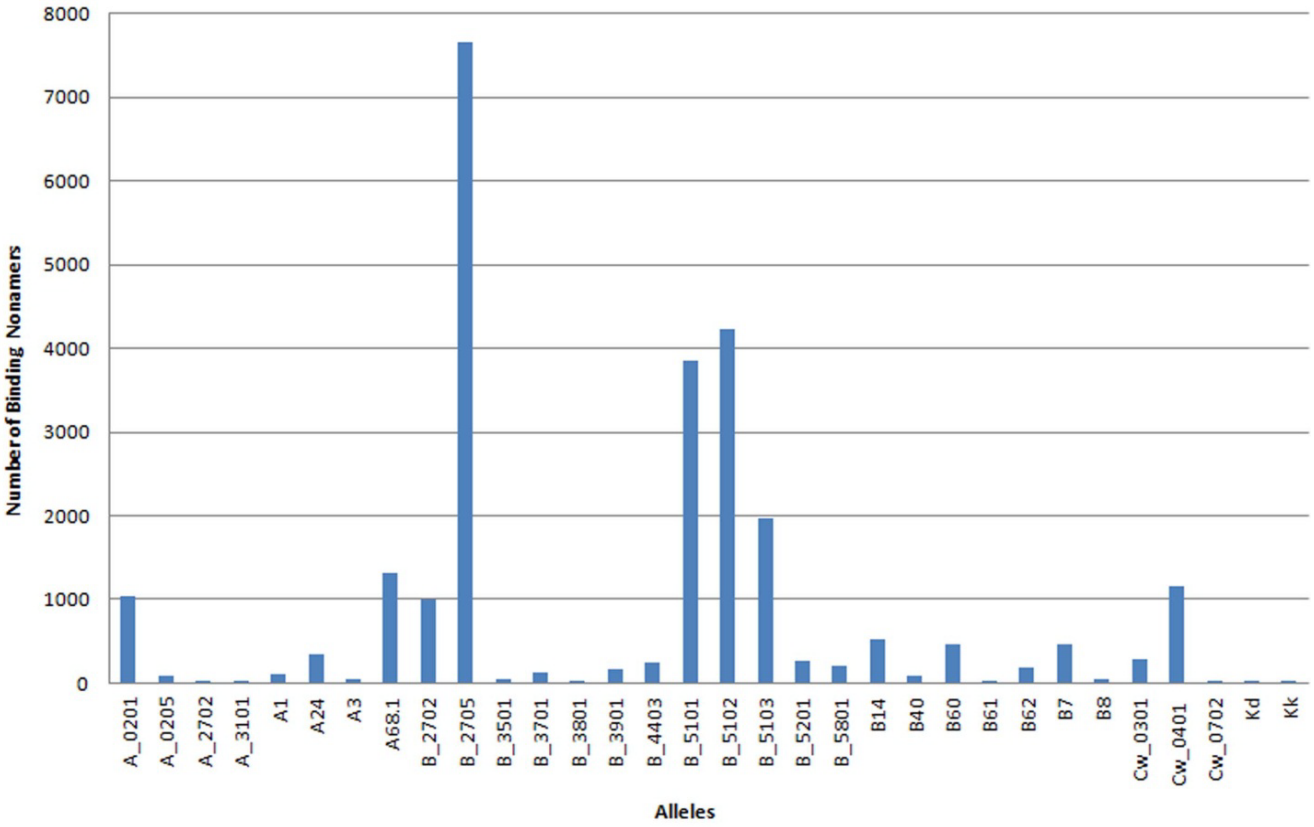

Figure 3: Observed variation in the binding of predicted peptides to the 33 different HLA class I and mouse alleles studied. The figure depicts the proportion of binding peptides to individual alleles. Alleles from B locus are shown to bind large number of peptides with high affinity as compared with A and C loci. HLA B_2705 shows the highest number of peptides binding with high affinity i.e. around 7657 peptides. Some alleles like A_1101, A_3302, A_3902, Cw_0601, do not bind to any peptides.

GO (Gene Ontolgy) analysis, InterProScan search, KEGG metabolic pathways analysis, and lastly, immunoinformatics based peptide candidate identification by BIMAS tool. 606 unknown proteins falling under the class VI of Sanger database classification system, were classified into either 'Cellular Components' or 'Biological Process' or 'Molecular Function' class 2 or 3 or 5 (Figure 2a-2c) in order to access the potential role of these proteins in the pathogen. Out of 606 unknown proteins, 1203 hits had either similarities with 'Molecular Function' or 'Biological Process' or 'Cellular Components' based on GO classification system, and 380 proteins were found to have InterProScan domain. From the top BLAST hits, it was observed that more than 575 top hits were from *M. tuberculosis,* and rest top BLAST hits were from *M. bovis*, *M. canetti*, *M. africanum*. Thus, from the BLAST analysis it can be inferred that most of these proteins belonging to the unknown class can act as good elicitor of T-cell immune response while designing the vaccine. The details of the functional annotation of 'unknown genes' from *M. tuberculosis* H37Rv is provided in the supplementary Table S1.

'Cellular components' of *M. tuberculosis* H37Rv

The cell envelope of *M. tuberculosis* is complex and composed of peptidoglycans, mycolic acids, lipids, and carbohydrates. By proteomics study, a total of 528 proteins were identified associated with cell wall, out of which 35% were involved in small molecule metabolism and 25% of which were involved in macromolecular synthesis and degradation [19]. The above studies provide strong evidence that *M. tuberculosis* cell wall is actively engaged in mycobacterial survival and remodeling. In line with this we identified 193 hits related to 'Cellular Component' based on GO classification system, majority of them fell into integral to membrane (66), cytoplasmic part (43), cytoplasm (46), and cell wall (34), and are represented in Figure 2a. Many membrane-associated proteins represent potential drug targets, diagnostics probes and can be used as vaccine candidate. 100 intergral proteins with the presence of at least one predicted transmembrane alpha-helix were detected from *M. tuberculosis* H37Rv by experimental validation [20]. Thus, it is conceivable that 'Cellular Components' of *M. tuberculosis* H37Rv proteins belonging to the unknown class, could play vital role in *Mycobacterial* survival, immune modulation inside host, and can be used as potential drug targets.

'Biological process' components of *M. tuberculosis* H37Rv

We performed genes belonging to GO 'Biological Process', most of which belong to metabolic process (110); cellular processes (77) such as transcription, translation, transport as well as lipid metabolism; growth (61); biological regulation (33); response to stimuli (24); multi-organism process (22); signaling (5) followed by other processes, and are represented in Figure 2b. Out of all the genes falling under the category of 'Biological Process', Rv3903c which is an alanine rich genes predicted to be involved in pathogenesis process, Rv3542c was predicted to be involved in growth of symbionts inside the host, Rv2147c was predicted to be involved in cell division, and details of the analysis is provided in the supplementary Table S1.

'Molecular function' components of *M. tuberculosis* H37Rv

The major genes falling under 'Molecular Functions' in *Mycobacterium* unknown proteins were mainly divided into binding proteins such as toxin binding (4), protein binding (12), nucleic acid binding (26), metal cluster binding (3), ion binding (18), cofactor binding (11), host cell surface binding (1), tetrapyrrole binding (1); hydrolase activity (51); transferase activity (41); oxidoreductase activity (25) along with other functional genes, and are represented in Figure 2c. Few of the immunogenic genes encountered are as follows: Rv1926c was predicted as immunogenic protein, Rv1174c was predicted as T cell antigenic protein, Rv2687c was predicted to be involved in the efflux transmembrane transporter activity etc., and rest functional genes are represented in supplementary Table S1.

Toxin-antitoxin system

The toxin-antitoxin act in a cognate pair forming a tight protein-protein complex, due to which the toxin is neutralized by the activity of antitoxin. So far 88 toxin-antitoxin pairs has been established from the genome sequence of *M. tuberculosis H37Rv* and may act in the participation of dormancy, nutrient starvation, and stress regulation [21,22]. In line with this we identified two proteins (Rv0299, Rv2019) encoding for toxin-antitoxin system, and four toxin-binding proteins (Rv1103c, Rv2550c, Rv1955, Rv1960c) which belong to the class VI of unknown protein.

Metabolic pathway analysis

After mapping the 606 unknown proteins of *M. tuberculosis* H37Rv against the KEGG metabolic pathway database, we encountered well known metabolic pathways related to 65 proteins along with their enzyme code equivalent, suggesting that these proteins could play essential role in governing major cellular processes as well as contribute to the survival of *M. tuberculosis* H37Rv strain. Essential genes are defined as those genes which are required for optimal growth of the *Mycobacterium*, and many predicted to be involved in several central metabolic pathways. These pathways include those required for synthesis of amino acids, nucleic acids, co-factors, and cholesterol metabolism. The vast majority of genes that play roles in other central cellular processes, including replication, transcription, protein synthesis and cell division also produced growth attenuation when mutated [23,24]. In line with this we identified 31 unique proteins mapped to different putative metabolic pathways represented in the supplementary table S1, out of which few of them can be categorized under the class of essential genes.

Vitamin B6 biosynthesis is essential for survival, virulence, and their inhibition is expected to affect the metabolism of *M. tuberculosis* at multiple sites. The biosynthesis of essential cofactors is of particular interest as these pathways are absent in man, and thus, might represent a candidate pathway for the development of new antitubercular agents [25]. In line with this, we identified three genes (Rv1155, Rv2061c, Rv2074) which were mapped to Vitamin B6 metabolism pathway.

Riboflavin (vitamin B(2)) acts as the precursor for FMN and FAD in almost all organisms that utilize the redox-active isoalloxazine ring system as a coenzyme in enzymatic reactions. The role of flavin is not only limited to redox reaction, but also widely used as a signaling and sensing molecule in biological processes such as phototropism and nitrogen fixation [26]. As riboflavin metabolism is essential for the growth of human pathogen *M. tuberculosis* [24, 26] and its complete understanding could provide insights into new intervention strategies that can target riboflavin pathway. We identified Rv2577 and Rv3007c which were mapped to riboflavin metabolic pathway.

There is an urgent need for the development of new antibiotics that can kill bacteria via novel mechanisms and the fatty acid biosynthetic pathway is an essential metabolic process in bacteria and presents several novel targets for antibiotic development [27]. In line with this we identified Rv0241c, Rv0241c, Rv0636 which were mapped to fatty acid biosynthesis pathway.

Stress-related genes

During the host-pathogen interaction, *M. tuberculosis* must cope with a variety of antimicrobial reactive oxygen and nitrogen species (ROS/RNS), produced from the macrophage of host human. *M. tuberculosis* utilizes a variety of defense cascades against ROS/RNS. One of the defense mechanism is executed by its thick cell wall composed of lipoarabinomannan (LAM), cyclopropanated mycolic acids, as well as phenolic glycolipid I (PGL-1), which act as potent scavengers of oxygen radicals [28]. The second mechanism is carried by a battery of protective enzymes such as catalases, superoxide dismustases, peroxidases, peroxynitrite reducatse complex, thioredoxin complex, glutathione metabolic and sulphur assimilation pathway [29]. Not only this but also the pathogen has to use the redox stress manifested by the host for synchronizing the metabolic pathways and expression of virulence factors, is the key to its success as a pathogen [30]. In line with this we identified genes encoding for heat shock protein90, universal stress response protein, anti-oxidant enzymes which are listed in the supplementary Tables S1, and these genes could play vital role for Mycobacterium survival success inside the host human.

Immunoinformatics based T cell epitopes prediction

Lastly, 606 unknown proteins were predicted for haplotype specific vaccine design against tuberculosis. All overlapping nonamer peptides were generated from the dataset and were screened for their potential to bind against HLA using BIMAS algorithm. The BIMAS algorithm has been a well-accepted method in a number of cases to predict T cell epitopes. About 25,964 nonamers bind to one or more HLA alleles with a cutoff value $T_{1/2} \geq 100$ minutes. The percentage of HLA-binding peptides generated from different individual proteins varies from different 0.48% to 28.57 %. The % of binding is calculated as below:

% of Binding = (Peptides Binding / Peptides Generated) × 100

The alleles binding to the generated 9-mer peptides from different individual proteins varies from 1-23 alleles. The number of binding peptides was independent of size of protein; i.e. larger protein did not always yield proportionally larger number of binding peptides. The details of analysis of the above results are given in the supplementary Table S1. An analysis of the binding profiles exhibited by the unknown proteins identified with various features for vaccine design are described below. The profiles were studied at the peptide level for individual epitope, then at the protein level that reflects cumulative binding properties of these peptides and finally from an allele perspectives.

Specificity and promiscuity in peptide-HLA binding

Most of the peptides were found to exhibit mono-allele specificity, meaning that they bind to a single allele. Some of them, however, appeared to bind multiple alleles; the highest number of alleles a given nonamer could bind being 5 out of the 33 alleles tested. This observation was not restricted to one locus but was the same in all 3 loci: A, B, C, thus providing a clear idea of specificity exhibited by individual

peptides. Infact, out of the 25,964 HLA-binding nonamers, 3 peptides bind to 5 alleles, 77 peptides bind to 4 alleles, 2097 bind to 3 alleles, 3717 bind to 2 alleles, 11,916 peptides show mono-allelic binding with a cutoff value of $T_{1/2} \geq 100$ minutes.

Identification of peptides binding to larger number of alleles

Identifying proteins with many epitopes that can bind to many HLA class I molecules is important, considering the polymorphic nature of HLA and its diversity in populations of different geographical regions. Therefore, a good T cell antigen should have peptides recognized by many HLA alleles. The analysis revealed that 606 proteins, leads to 25,964 potential peptides, since they bind to HLA varying from 0-23 alleles for different Rv numbers.

Few alleles bind larger number of peptides than others, for short-listing potential vaccine candidates, it is important to analyze the binding profiles from an allele perspective. Of the 33 alleles studied, the largest number of nonamers was found to be recognized by allele B_2705 (7657) followed by B_5102 (4226), B_5101 (3855), B_5103 (1979), A68.1 (1312), Cw_0401 (1148), A_0201 (1034), B_2702 (998), A14 (523) and B60 (457), as illustrated in the histogram in Figure 3. It is worth nothing that several alleles such as A_1101, A_3302, A_3902, Cw_0601, did not bind to any peptides, and alleles Cw_0702 bind to only two peptides, A_3101 bind to only one peptide, at the chosen cutoff $T_{1/2} \geq 100$. These alleles may not have high affinity for peptides, it is important to study the prevalence of these alleles in different populations and consider variations in immune response during vaccine design. Systematic studies such as this can also facilitate analysis of correlation of vaccine failures with absence of these alleles.

Allelic variation in binding affinity

The $T_{1/2}$ with which a peptides binds to HLA ranges from 100 to 20,000 minutes. 25,964 peptides from unknown protein are high affinity binder, which bind with $T_{1/2} \geq 100$. In general, the binding affinity of peptides to B locus alleles (especially B_2705, B_5101) is higher as compared to alleles of A and C loci.

Self or partially self-peptides among binding nonamers

The total of 25,964 nonamer binding peptides generated from 606 unknown proteins were analyzed for similarities with each of 45,513 proteins of the human [31] and peptides had similarities in 5 or 6 or 7 or 8 out of 9 positions, might act as partial self- peptides, or complete self-peptides (9 out of 9 positions), were eliminated for further analysis. Only peptides were found to be completely non-self (matching 4 or below 4 out 9 positions) generated from proteins can be taken for future experimental approach [32] and a complete analysis of these peptides with human proteomes were performed.

Discussion

The design and development of new generation vaccines have been focused on individual proteins or DNA coding for limited number of proteins. Peptides binding to HLA are essential for generation and maintenance of an immune response. Therefore, it becomes imperative that the proteins which are identified as possible vaccine candidates, must generate peptides that are recognized by MHC class-I for cytotoxic T cell response. This becomes important, since humans carry only a limited number of co-dominant HLA alleles in their genome, out of which hundreds of polymorphic alleles that are present in the population. Therefore a candidate vaccine must generate peptides that can bind to a wide range of HLA molecules to provide good population coverage. Without this, the protein even if generates good primary response and memory response, will work only in a limited number of individuals. To address this issue of selection of appropriate proteins as candidate vaccine, we have carried out a systematic analysis of *Mycobacterium* peptides derived from unknown class VI of Sanger database, determination of their HLA binding specificity to narrow down the proteins containing sufficient peptides for total population coverage.

Several computational methods are now available for analysis of binding of peptides to MHC [33]. We have used BIMAS binding tool available in public domain to study the binding of peptides to thirty-three HLA class-I alleles and seven mouse alleles. We use BIMAS for ranking of potential peptides as this tool has been developed based on experimental data of half time of dissociation of β_2 microglobulin from the HLA-peptide complex [18]. A $T_{1/2}$ of 100 minute was chosen as a cutoff point in order to select relatively higher binding peptides. We also cannot exclude the possibility of missing some good T cell epitopes which bind with HLA with relatively less affinity, but may still bind to TCR with high affinity. It is the TCR-HLA-peptide complex, which is also crucial in determining the T-cell response.

The nonameric sequences of class VI of Sanger database were predicted to contain high percentage of binding peptides to human class-I HLA. Most of the peptides were predicted to bind to alleles from B-locus, and affinity with which the peptides are binding to this locus is far higher than alleles of A and C loci. This implies that distribution of B loci alleles in the population may play a role in the susceptibility of the population for certain infection.

Many of the peptides are monoallelic binders i.e. they bind to a single allele. The T cell epitopes that are recognized in context of more than one HLA and more than one T cell clones are called promiscuous epitopes. Promiscuous peptides are of prime interest for vaccine design because of their relevance to cover higher proportions of human population. This *in silico* approach would help to predict some of the HLA-binding motifs, which could act as promiscuous epitopes. There are larger numbers of binders from unknown proteins, for which either self or partially self-host peptides exist. These finding implicates the possible failure of certain epitopes as vaccine candidates. Partially self-peptides can mount an autoimmune response in the host upon immunization, whereas the inclusion of self-peptides has no obvious advantage. So, we tried to screen the peptides from this family which is having no sequence homology with human proteome, and can be considered as non-self-peptides in the host human to generate good immune response.

A large number of researchers have been addressing the identification of T-cell epitopes of individual *Mycobacterial* proteins [34]. Computational methods have been used to identify peptides which would bind to HLA [35-37]. In 2002 re-annotation of unknown proteins of *M. tuberculosis* H37Rv was performed, and 272 proteins were still reported to be unknown genes [9]. A structural genomics project has been initiated aimed at determining the structures of 270 unique proteins from *M. tuberculosis*, and out of which 30 structures have been determined by the Indian structural biology group [38]. For example, Rv2827c of *M. tuberculosis* is classified as 'unknown protein', and structural analysis revealed that it is a DNA-binding protein [39]. These results are encouraging as functional understanding can be deduced from experimental solved structures.

The present work differ from earlier investigations in that a genome-

based functional approach has been followed, and all overlapping peptides belonging to the unknown class VI of *M. tuberculosis* proteins have been analyzed to predict T cell reactive peptides. We here predict reliable functional annotation of 522 genes out of 606 'unknown genes' by integrating several bioinformatics anaylsis and annotation tools. Again the immunoinformatics method used here for characterization of genes at peptide level would complement the existing experimental technologies and provide a much quicker, less expensive way to molecular biologists towards the development of a better vaccine than the presently available BCG vaccine. All together, the integrative approach offers a powerful means to undertake the huge challenge to characterize the rapidly growing number of unknown proteins with each newly sequenced genome.

Conclusions

We have generated a large datasets on possible HLA class-I binding peptides, which have been predicted to bind different HLA. Most of these peptides binding to HLA are specific with few having high promiscuous binders. So, it is obvious that a large cocktail of proteins are required to achieve reasonable population coverage. This work also suggests the feasibility of designing of haplotype specific subunit vaccine, which can be given to individuals with known HLA haplotype. The haplotype specific vaccines can be combined to target a population where the distributions of HLA alleles are known. Besides the above facts, this work also indicates that the use of single or limited number of genes in a DNA vaccine may not be suitable to cover a given population. Thus, the initial computational filtration of probable candidate antigens carried out in the present work would be cost and time effective. Also it can be inferred from BLAST analysis that most of the unknown proteins are restricted to the *M. tuberculosis* H37Rv genome, and could play essential role for their survival, stress response, virulence, and adaptability inside the host. The conclusions drawn for Mycobacterium proteins are expected to be generally true for all pathogens. Of note, these antigenic proteins could provide valuable targets against the drug-resistance and drug-susceptible H37Rv and its related reference strains.

References

1. Ginsberg AM (1998)The tuberculosis epidemic. Scientific challenges and opportunities. Public Health Rep 113: 128-36.
2. Harries AD (1990) Tuberculosis and human immunodeficiency virus infection in developing countries. Lancet 335: 387-90.
3. Ottenhoff TH, Kaufmann SH (2012) Vaccines against tuberculosis: where are we and where do we need to go? PLoS Pathog 8: e1002607.
4. Chaisson RE, Slutkin G (1989) Tuberculosis and human immunodeficiency virus infection. J Infect Dis 159:96–100.
5. Baily GV (1980) Tuberculosis prevention Trial, Madras. Indian J Med Res 72:1-74.
6. Agger EM, Andersen PA (2002) A novel TB vaccine; towards a strategy based on our understanding of BCG failure. Vaccine 21: 7-14.
7. Vallenet D, Belda E, Calteau A, Cruveiller S, Engelen S, et al. (2013) MicroScope--an integrated microbial resource for the curation and comparative analysis of genomic and metabolic data. Nucleic Acids Res 41: D636-47.
8. Mazandu GK, Mulder N.J. (2012) Functional Prediction and Analysis of Mycobacterium tuberculosis Hypothetical Proteins. Int J.Mol Sci 13: 7283-7302.
9. Camus JC, Pryor MJ, Médigue C, Cole ST (2002) Re-annotation of the genome sequence of Mycobacterium tuberculosis H37Rv. Microbiology 148: 2967-73.
10. Nayak R, Lal G, Shaila MS (2005) Perpetuation of immunological memory: role of serum antibodies and accessory cells. Microbes Infect 7:1276-83.
11. Nayak R, Mitra-Kaushik S, Shaila MS (2001) Perpetuation of immunological memory: a relay hypothesis. Immunology 102: 387-95.
12. Chaitra MG, Shaila MS, Nayak R (2008) Characterization of T-cell immunogenicity of two PE/PPE proteins of Mycobacterium tuberculosis. J Med Microbiol 57: 1079-86.
13. Vani J, Shaila MS, Chandra NR, Nayak R (2006) A combined immuno-informatics and structure-based modeling approach for prediction of T cell epitopes of secretory proteins of Mycobacterium tuberculosis. Microbes Infect 8: 738-46.
14. Chaitra MG, Hariharaputran S, Chandra NR, Shaila MS, Nayak R (2005) Defining putative T cell epitopes from PE and PPE families of proteins of Mycobacterium tuberculosis with vaccine potential. Vaccine 23:1265-72.
15. Badapanda C (2013) Suppression subtractive hybridization (SSH) combined with bioinformatics method: an integrated functional annotation approach for analysis of differentially expressed immune-genes in insects. Bioinformation 9:216-221.
16. Conesa A, Götz S (2008) Blast2GO: A comprehensive suite for functional analysis in plant genomics. Int J Plant Genomics 2008: 619832.
17. Schumacher TN, De Bruijn ML, Vernie LN, Kast WM, Melief CJ, et.al. (1991) Peptide selection by MHC class I molecules. Nature 350: 703-6.
18. Parker KC, Bednarek MA, Coligan JE (1994) Scheme for ranking potential HLA-A2 binding peptides based on independent binding of individual peptide side-chains. J Immunol 152:163-75.
19. Wolfe LM, Mahaffey SB, Kruh NA, Dobos KM (2010) Proteomic definition of the cell wall of Mycobacterium tuberculosis. J Proteome Res 9: 5816-26.
20. Xiong Y, Chalmers MJ, Gao FP, Cross TA, Marshall AG (2005) Identification of Mycobacterium tuberculosis H37Rv integral membrane proteins by one-dimensional gel electrophoresis and liquid chromatography electrospray ionization tandem mass spectrometry. J Proteome Res 4: 855-61.
21. Ramage HR, Connolly LE, Cox JS (2009) Comprehensive functional analysis of Mycobacterium tuberculosis toxin-antitoxin systems: implications for pathogenesis, stress responses, and evolution. PLoS Genet 5: e1000767.
22. Kumar P, Issac, B, Dodson E.J, Turkenburg, J.P, Mande, S.C. (2008) Crystal structure of Mycobacterium tuberculosis YefM antitoxin reveals that it is not an intrinsically unstructured protein. J Mol Biol 383: 482-93.
23. Griffin JE, Gawronski JD, Dejesus MA, Ioerger TR, Akerley BJ, et.al. (2011) High-resolution phenotypic profiling defines genes essential for mycobacterial growth and cholesterol catabolism. PLoS Pathog 7: e1002251.
24. Assetti CM, Boyd DH, Rubin EJ (2003) Genes required for mycobacterial growth defined by high density mutagenesis. Mol Microbiol 48:77-84.
25. Dick T, Manjunatha U, Kappes B, Gengenbacher M (2010) Vitamin B6 biosynthesis is essential for survival and virulence of Mycobacterium tuberculosis. Mol Microbiol 78: 980-8.
26. Macheroux P, Kappes B, Ealick, SE (2011) Flavogenomics--a genomic and structural view of flavin-dependent proteins. FEBS J 278: 2625-34.
27. Payne DJ (2004) The potential of bacterial fatty acid biosynthetic enzymes as a source of novel antibacterial agents. Drug News Perspect 17: 187-94.
28. Flynn JL, Chan J (2001) Immunology of tuberculosis. Annu Rev Immunol 19: 93-129.
29. Voskuil MI, Bartek IL, Visconti K, Schoolnik GK (2011) The response of mycobacterium tuberculosis to reactive oxygen and nitrogen species. Front Microbiol 2: 105.
30. Trivedi A, Singh N, Bhat SA, Gupta P, Kumar A (2012) Redox biology of tuberculosis pathogenesis. *Adv* Microb Physiol 60: 263-324.
31. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, et al. (2001) The sequence of the human genome. Science 291:1304-51.
32. Parida R, Shaila MS, Mukherjee S, Chandra NR, Nayak R (2007) Computational analysis of proteome of H5N1 avian influenza virus to define T cell epitopes with vaccine potential. Vaccine 25: 7530-7539.
33. Lundegaard C, Lund O, Nielsen M (2012) Predictions versus high-throughput experiments in T-cell epitope discovery: competition or synergy? Expert Rev Vaccines 11: 43-54.
34. Rueckert C, Guzmán CA (2012) Vaccines: from empirical development to rational design. PLoS Pathog 8.
35. Davila J, McNamara LA, Yang Z (2012) Comparison of the predicted population coverage of tuberculosis vaccine candidates Ag85B-ESAT-6, Ag85B-TB10.4,

and Mtb72f via a bioinformatics approach. PLoS One 7: e40882.

36. Zhang GL, Ansari HR, Bradley P, Cawley GC, Hertz T, et al. (2011) Machine learning competition in immunology - Prediction of HLA class I binding peptides. J Immunol Methods 374: 1-4.

37. Martin W, Sbai H, De Groot AS (2003) Bioinformatics tools for identifying class I-restricted epitopes. Methods 29: 289-98.

38. Arora A, Chandra NR, Das A, Gopal B, Mande SC, et al. (2011) Structural biology of *Mycobacterium tuberculosis* proteins: the Indian efforts. Tuberculosis (Edinb) 91: 456-68.

39. Janowski R, Panjikar S, Eddine AN, Kaufmann SH, Weiss MS (2009) Structural analysis reveals DNA binding properties of Rv2827c, a hypothetical protein from Mycobacterium tuberculosis. J Struct Funct Genomics 10:137-50.