Infection, Genetics and Evolution 12 (2012) 1312–1318

Contents lists available at SciVerse ScienceDirect

Infection, Genetics and Evolution

journal homepage: www.elsevier.com/locate/meegid

Short communication

In silico identification of potential antigenic proteins and promiscuous CTL epitopes in *Mycobacterium tuberculosis*

Jagadish Chandrabose Sundaramurthi, S. Brindha, S.R. Shobitha, A. Swathi, P. Ramanandan, Luke Elizabeth Hanna *

ICMR-Biomedical Informatics Centre, Department of Clinical Research, National Institute for Research in Tuberculosis (Formerly Tuberculosis Research Centre), Indian Council of Medical Research, Chennai 600 031, Tamil Nadu, India

ARTICLE INFO

Article history:
Received 11 November 2011
Received in revised form 23 March 2012
Accepted 28 March 2012
Available online 5 April 2012

Keywords:
Mycobacterium tuberculosis
Bioinformatics
Antigenic proteins
Dormancy
Epitopes
Vaccine

ABSTRACT

Cell-mediated immunity is critical for the control of *Mycobacterium tuberculosis* infection. We hypothesized that those proteins of *M. tuberculosis* (MTB) that do not have homologs in humans as well as human gut flora, would mount a good antigenic response in man, and employed a bioinformatics approach to identify MTB antigens capable of inducing a robust cell-mediated immune response in humans. In the first step we identified 624 MTB proteins that had no homologs in humans. Comparison of this set of proteins with the proteome of 77 different microbes that comprise the human gut flora narrowed down the list to 180 proteins unique to MTB. Twenty nine of the 180 proteins are known to be associated with dormancy. Since dormancy associated proteins are known to harbor CTL epitopes, we selected four representative unique proteins and subjected them to epitope analysis using ProPred1. Nineteen novel promiscuous epitopes were identified in the four proteins. Population coverage for 7 of the 19 shortlisted epitopes including Rv3852 (58-KPAEAPVSL, 112-VPLIVAVTL, 118-VTLSLLALL and 123-LALLLIRQL), Rv2706c (66-RPLSGVSFL) Rv3466 (8- RIVEVFDAL and 38-RSLERLECL) was >74%. These novel promiscuous epitopes are conserved in other virulent MTB strains, and can therefore be further investigated for their immunological relevance and usefulness as vaccine candidates.

1. Introduction

Tuberculosis kills about 2 million persons annually and is the second leading cause of death due to an infectious disease (WHO, 2010). The intensity of the problem due to tuberculosis is compounded by the emergence of drug resistance and co-infection with HIV (Raviglione and Smith, 2007; Shah et al., 2007). Bacillus Calmette-Guerin (BCG) vaccine, which has been used for nearly 100 years, provides highly variable protection (0–80%) against pulmonary tuberculosis (Lienhardt and Zumla, 2005). Hence, a novel and effective vaccine is needed to combat tuberculosis.

Cellular immune responses play a major role in the control of *Mycobacterium tuberculosis* (MTB) infection (Stenger and Modlin, 1999). Identification of MTB peptides capable of inducing such a response is critical. The availability of the whole genome sequence of *M. tuberculosis* (Cole et al., 1998) has facilitated the design of new experiments in these lines. Zvi et al. (2008) identified 45 candidate proteins as antigens using data mining and bioinformatics analyses. Many tools are now available for the identification of promiscuous epitopes from antigenic proteins (Flower, 2003; Lin et al., 2008; Salimi et al., 2010). These tools have been successfully used for the identification of T-cell epitopes in *M. tuberculosis* (De Groot et al., 2005; McMurry et al., 2005; Mustafa and Shaban, 2006; Kumar et al., 2010a,b; Deenadayalan et al., 2010), as well as for other conditions including cancer, autoimmunity and allergy (De Groot, 2006). By integrating comparative proteomics and transcriptomics data we identified antigenic proteins unique to MTB and predicted promiscuous epitopes from four representative proteins using bioinformatics tools.

2. Materials and methods

2.1. Identification of antigenic proteins

The amino acid sequences of all 3988 proteins of *M. tuberculosis* H37Rv (NC_000962) were downloaded from GenBank (http://www.ncbi.nlm.nih.gov/Genbank/), and compared with that of human proteins (Refseq, ftp://ftp.ncbi.nih.gov/) using BLASTP (version 2.2.17) (Altschul et al., 1997) at an *E*-value of 10. A relaxed *E*-value was chosen with the intent to identify proteins that are

* Corresponding author. Address: ICMR-Biomedical Informatics Centre, Department of Clinical Research, National Institute for Research in Tuberculosis (Formerly Tuberculosis Research Centre), Indian Council of Medical Research, 1, Sathiyamoorthy Road, Chetpet, Chennai 600 031, Tamil Nadu, India. Tel.: +91 44 28369500; fax: +91 44 28362528.
E-mail address: hannanirt@gmail.com (L.E. Hanna).

1567-1348/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.meegid.2012.03.023

very distinct from human proteins and highly specific to MTB. Proteins that had no similarity with any of the human proteins were further compared with the proteome of 77 microbes that comprise the normal human gut flora at the same *E*-value. These proteins were further analyzed in comparison with available literature to narrow down the list further to a set of highly significant proteins which could be evaluated *in vitro* for their potential as diagnostic markers or vaccine candidates.

2.2. Prediction of epitopes and population coverage

ProPred1 (Singh and Raghava, 2003) was used to predict potential promiscuous epitopes (available at http://www.imtech.res.in/raghava/propred1/). The threshold value was set at 3, since the sensitivity and specificity of epitope prediction at this value lies in the range of 66–78% and 80–81%, respectively. Thirty-nine HLA alleles were included for this analysis. Those epitopes which were predicted as binders to 10 or more than 10 HLA alleles were identified as promiscuous epitopes. For the short-listed promiscuous epitopes, population coverage was calculated using IEDB source (Bui et al., 2006) available at http://tools.immuneepitope.org/tools/population/iedb_input. The conservation of epitopes in other virulent MTB isolates was determined using two databases, TB Database (http://tbdb.org/) and *M. tuberculosis* Comparative Database (http://www.broadinstitute.org/annotation/genome/mycobacterium_tuberculosis_spp/MultiHome.html).

3. Results

3.1. Potential antigenic proteins

MTB proteins that did not have any similar human protein at an *E*-value of 10 were short-listed, resulting in a set of 624 unique proteins (Supplementary Table 1). These proteins were further compared with the proteome of 77 microbes comprising the human gut flora (Supplementary Table 2), obtained from a report by Raman et al. (2008). One hundred and eighty of the 624 above proteins were found to have no homologs in human gut flora as well (Supplementary Table 3), and hence referred to as unique to MTB.

To add further value to the above findings, we analyzed the set of unique proteins with respect to transcriptomics and proteomics data available in the form of published literature and bioinformatics databases such as MTBreg (http://www.doe-mbi.ucla.edu/Services/MTBreg/) which houses a large amount of transcriptomics and proteomics data specifically for MTB. Comparison of the 180 unique MTB proteins with MTBreg database revealed that 29 of these proteins were reported to be upregulated in *M. tuberculosis* during dormancy-associated conditions (Table 1). Further, the shortlisted proteins were compared with previous reports of potentially antigenic proteins. Thirteen of the 180 proteins were found in the list of 249 MTB proteins (Table 2) reported as antigenic by Li et al. (2010). The protein Rv1039c was found to be one among the top 20 proteins in Li's list arranged according to the order of significance. Sixteen of the 180 unique proteins from

Table 1
Unique MTB proteins also reported to be highly expressed in dormancy-associated conditions.

S. No.	Locus tag (Rv number)	Protein name	Reference	Expression condition(s)
1	Rv2551c	Hypothetical protein Rv2551c	Betts et al. (2002)	Nutrition starvation
2	Rv3654c	Hypothetical protein Rv3654c	Stewart et al. (2002)	Heat shock
3	Rv3852	Histone-like protein HNS	(1) Stewart et al. (2002) (2) Betts et al. (2002)	(1) Heat shock (2) Nutrition
4	Rv2706c	Hypothetical protein Rv2706c	(1) Stewart et al. (2002) (2) Muttucumaru et al. (2004) (3) Schnappinger et al. (2003)	(1) Heat shock (2) Anaerobic conditions ((3) Presence of H_2O_2
5	Rv1195	PE family protein	Muttucumaru et al. (2004)	Anaerobic condition
6	Rv0742	PE-PGRS family protein	Bacon et al. (2004)	Reduced oxygen
7	Rv0834c[b]	PE-PGRS family protein	Voskuil et al. (2004)	Low-oxygen
8	Rv1806	PE family protein	Voskuil et al. (2004)	Low-oxygen
9	Rv1807	PPE family protein	Voskuil et al. (2004)	Low-oxygen
10	Rv3021c	PPE family protein	Voskuil et al. (2004)	Low-oxygen
11	Rv3022c	PPE family protein	Voskuil et al. (2004)	Low-oxygen
12	Rv2661c	Hypothetical protein Rv2661c	(1) Muttucumaru et al. (2004) (2) Betts et al. (2002)	(1) Anaerobic conditions (2) Nutrition starvation
13	Rv2929	Hypothetical protein Rv2929	Muttucumaru et al. (2004)	Anaerobic conditions
14	Rv3135	PPE family protein	Muttucumaru et al. (2004)	Anaerobic conditions
15	Rv1808	PPE family protein	Betts et al. (2002)	Nutrition starvation
16	Rv0285	PE family protein	(1) Betts et al. (2002) (2) Schnappinger et al. (2003)	(1) Nutrition starvation (2) Presence of H_2O_2
17	Rv0476[a]	Transmembrane protein	Betts et al. (2002)	Nutrition starvation
18	Rv3477	PE family protein	Betts et al. (2002)	Nutrition starvation
19	Rv0882	Hypothetical protein Rv0882	Betts et al. (2002)	Nutrition starvation
20	Rv2160c	Hypothetical protein Rv2160c	Betts et al. (2002)	Nutrition starvation
21	Rv3760[a]	Hypothetical protein Rv3760	Betts et al. (2002)	Nutrition starvation
22	Rv0426c	Hypothetical protein Rv0426c	Betts et al. (2002)	Nutrition starvation
23	Rv1588c	REP13E12 repeat-containing protein	(1) Betts et al. (2002) (2) Schnappinger et al. (2003)	(1) Nutrition starvation (2) presence of H_2O_2
24	Rv2665	Hypothetical protein Rv2665	Betts et al. (2002)	Nutrition starvation
25	Rv3466	Hypothetical protein Rv3466	(1) Betts et al. (2002) (2) Schnappinger et al. (2003)	(1) Nutrition starvation (2) Presence of H_2O_2
26	Rv0872c	PE-PGRS family protein	Betts et al. (2002)	Nutrition starvation
27	Rv3508	PE-PGRS family protein	Stewart et al. (2002)	Heat shock
28	Rv1791	PE family protein	Stewart et al. (2002)	Heat shock
29	Rv2081c	Transmembrane protein	Stewart et al. (2002)	Heat shock

[a] Present in Table 1 and Table 2.
[b] Present in Table 1 and Table 3.

Table 2
Unique MTB proteins also reported to be antigenic by Li et al. (2010).

S. No.	Locus tag (Rv number)	Protein name
1	Rv0109	PE-PGRS family protein
2	Rv0124	PE-PGRS family protein
3	Rv0476[a]	Transmembrane protein
4	Rv0532[c]	PE-PGRS family protein
5	Rv0608	Hypothetical protein Rv0608
6	Rv0661c	Hypothetical protein Rv0661c
7	Rv1039c	PPE family protein
8	Rv1468c	PE-PGRS family protein
9	Rv1517	Hypothetical protein Rv1517
10	Rv1803c	PE-PGRS family protein
11	Rv2571c	Transmembrane alanine and valine and leucine rich protein
12	Rv2741	PE-PGRS family protein
13	Rv3760[a]	Hypothetical protein Rv3760

[a] Present in Table 1 and Table 2.
[c] Present in Table 2 and Table 3.

our list were also found in the list of 248 MTB proteins reported by Tang et al. (2011) to contain peptide epitopes capable of activating polyfunctional CD8 + T cells (Table 3).

From the above analysis, it was found that among the 180 unique MTB proteins, 29 were previously reported to be highly expressed in dormancy-associated conditions, 13 were reported by Li et al. (2010) as reactive antigens, and 16 were identified by Tang et al. (2011) to contain epitopes capable of activating CD8 + T cells. Thus, the above set of 54 proteins (29 + 13 + 16 minus four proteins which are present in more than one list) can be considered as highly potential antigens. Two proteins, Rv0476 and Rv3760, were found to be highly expressed in dormancy and reported to possess antigenic properties (Li et al., 2010). One protein, Rv0834c, was found to be highly expressed in dormancy and has also been reported to contain immunogenic epitopes by Tang et al. (2011). Protein, Rv0532, was independently reported as antigenic both by Li et al. (2010) and Tang et al. (2011).

3.2. Potential promiscuous epitopes

Four proteins (Rv3852, Rv2706c, Rv2661c and Rv3466) were selected from among the set of proteins short-listed by us (Table 1) as representative candidates and subjected to epitope analysis.

Table 3
Unique MTB proteins also reported to have antigenic epitopes by Tang et al. (2011).

S. No.	Locus tag (Rv number)	Protein name
1	Rv0532[c]	PE-PGRS family protein
2	Rv0833	PE-PGRS family protein
3	Rv0834c[b]	PE-PGRS family protein
4	Rv1158c	Hypothetical protein Rv1158c
5	Rv3763	19 kDa lipoprotein antigen precursor LPQH
6	Rv1196	PPE family protein
7	Rv0440	Chaperonin GroEL
8	Rv0667	DNA-directed RNA polymerase subunit beta
9	Rv1073	Hypothetical protein Rv1073
10	Rv3846	Superoxide dismutase
11	Rv0570	Ribonucleoside-diphosphate reductase large subunit NrdZ
12	Rv1308	F0F1 ATP synthase subunit alpha
13	Rv1966	MCE-family protein MCE3A
14	Rv2006	Trehalose-6-phosphate phosphatase OtsB1
15	Rv3036c	Hypothetical protein Rv3036c
16	Rv3285	Bifunctional acetyl-/propionyl-coenzyme A carboxylase subunit alpha

[b] Present in Table 1 and Table 3.
[c] Present in Table 2 and Table 3.

Epitopes that bind to multiple HLA alleles could serve as potential candidates for vaccine design since a large number of people in the population could be covered; such epitopes are referred to as promiscuous epitopes. For our analysis we considered nanomers that were recognized by at least one fourth (10 or more) of the 39 HLA alleles listed in ProPred1, as promiscuous epitopes. A total of 290 nanomers were predicted to be epitopes from the selected 4 proteins. Of these only 19 (6.55%) nanomers were classified as promiscuous, based on their capacity to bind to 10 different HLA alleles. The list of the promiscuous epitopes and the HLA alleles they recognized is provided in Table 4 and details of the population coverage for individual proteins are provided in Table 5.

Rv3852 contained the maximum number of promiscuous epitopes (8) while Rv2706c, Rv2661c and Rv3466 had 3, 4 and 4 promiscuous epitopes, respectively. For each of the promiscuous epitopes, population coverage analysis was performed using the HLA profile generated with ProPred1. Thirteen of the 19 epitopes had a population coverage of >50%; While nine epitopes had a population coverage >65%, four had a population coverage >75% (Table 4). Population coverage was also calculated for each of the four proteins using their promiscuous epitopes and their respective HLA alleles. Rv3852, and Rv3466 had a population coverage >90% and Rv2706c and Rv2661c had a of population coverage of 85.19% and 71.09%, respectively (Table 5).

We also examined whether the shortlisted epitopes are conserved across different *M. tuberculosis* strains (CDC1551, F11, KZN DS, KZN MDR, KZN XDR and Harleem strains) by comparative analysis. Nine of the 19 promiscuous epitopes were found to be conserved in all the six strains. Four epitopes were found to be conserved in KZN DS, KZN MDR, KZN XDR, F11 and MTB Harleem strains, and one epitope from Rv3466 (13:RIVEVFDAL) was found to be conserved in KZN DS, KZN MDR, KZN XDR, F11 and CDC1551. In addition, two epitopes (49:RLPAVGHAL, and 50 LPAVGHALI) from Rv3466 were found to be conserved in four of the strains studied.

4. Discussion

Ristori et al. (2000) reported that while epitopes that mimic human peptides are tolerated, peptides that do not resemble human peptides induce an immunogenic response in the human host. No or low level of similarity to the host proteome is considered as the common property that defines the immunological "nonself" nature of antigenic sequences in cancer, autoimmunity, infectious diseases and allergy (Kanduc, 2010) and guarantee highest specificity and lowest cross-reactivity in designing effective, safe and apparently infallible immunotherapeutical tools (Kanduc et al., 2007). Rolland et al. (2007) demonstrated that there exists an inverse relationship between the similarity of HIV antigens with human proteins and their recognition by immune cells. Further, it is also reported that the epitopes that share similarity with human peptides are responsible for auto-immune reactions (Amela et al., 2007), and immune evasion (Ludin et al., 2011). Thus, the immune system preferentially responds to antigenic sequences that are never or are only sporadically encountered in the repertoire of self-antigens, and therefore removing host-mimics from vaccine constructs could be crucial for designing efficacious vaccines. This principle of exclusion of host-mimics has been employed by several investigators while designing vaccine candidates against infectious pathogens including MTB (Vani et al., 2006; Wang et al., 2010), *Plasmodium falciparum* (Singh and Mishra, 2009), *Neisseria meningitidis* Serogroup B (Gupta et al., 2010), etc. However, it is also reported that the dissimilarity concept may not be used as the sole criterion for antigen identification, since an analysis of a set of known antigens and

Table 4
Potential CTL epitopes from proteins Rv3852, Rv2706c, Rv2661c and Rv3466.

S. No.	Name of the protein	Protein ID	Start position of potential epitope	Aminoacid sequence of the 9-mer	Number of HLA alleles recognizing the promiscuous epitopes	List of HLA alleles recognizing the epitope	Theoretical population coverage (%)
1	Possible histone-like protein	Rv3852	58	KPAEAPVSL	15	HLA-A24, HLA-B*2705, HLA-B*3501, HLA-B*3801, HLA-B40, HLA-B*5101, HLA-B*5102, HLA-B*5103, HLA-B*5301, HLA-B*5401, HLA-B*51, HLA-B60, HLA-B7, HLA-B*0702, HLA-Cw*0401.	74.47
2	Possible histone-like protein	Rv3852	104	VPAPSHSPV	11	HLA-B*3501, HLA-B*5101, HLA-B*5102, HLA-B*5103, HLA-B*5301, HLA-B*5401, HLA-B*51, HLA-B7, HLA-B*0702, HLA-Cw*0401, HLA-B61	46.41
3	Possible histone-like protein	Rv3852	106	APSHSPVPL	12	HLA-B*2705, HLA-B*3801, HLA-B40, HLA-B*5101, HLA-B*5102, HLA-B*5103, HLA-B*5301, HLA-B*5401, HLA-B60, HLA-B7, HLA-B*0702, HLA-Cw*0401.	55.74
4	Possible histone-like protein	Rv3852	110	SPVPLIVAV	13	HLA-B*3501, HLA-B*5101, HLA-B*5102, HLA-B*5103, HLA-B*5301, HLA-B*5401, HLA-B*51, HLA-B7, HLA-B*0702, HLA-Cw*0401, HLA-B61, HLA-B*5201, HLA-Cw*0602.	52.03
5	Possible histone-like protein	Rv3852	112	VPLIVAVTL	17	HLA-A24, HLA-B*3501, HLA-B*3801,HLA-B40, HLA-B*5101, HLA-B*5102, HLA-B*5103, HLA-B*5301, HLA-B*5401, HLA-B*51, HLA-B60, HLA-B7, HLA-B*0702, HLA-Cw*0401, HLA-B14, HLA-B*3902, HLA-Cw*0301.	75.69
6	Possible histone-like protein	Rv3852	118	VTLSLLALL	16	HLA-A24, HLA-B40, HLA-B*5301, HLA-B*51, HLA-B7, HLA-Cw*0401,HLA-Cw*0602,HLA-B14, HLA-B*3902, HLA-Cw*0301, HLA-A*0205, HLA-A2.1, HLA-B*3701, HLA-B*3901, HLA-B*5801, HLA-B60	78.29
7	Possible histone-like protein	Rv3852	119	TLSLLALLL	10	HLA-B*2705, HLA-B7, HLA-B*3902, HLA-A2, HLA-A*0201, HLA-A*0205, HLA-A3, HLA-A2.1, HLA-B*3701, HLA-B62.	49.58
8	Possible histone-like protein	Rv3852	123	LALLLIRQL	14	HLA-A24, HLA-B40, HLA-B*5101, HLA-B*5102, HLA-B*5103, HLA-B7, HLA-Cw*0401, HLA-Cw*0602, HLA-B14, HLA-B*3902, HLA-Cw*0301, HLA-B*3901, HLA-B*5801, HLA-B60.	74.44
9	Putative uncharacterized protein	Rv2706c	22	HPSCSATAV	10	HLA-B*3501, HLA-B*5101, HLA-B*5102, HLA-B*5103, HLA-B*5301, HLA-B*5401, HLA-B*51, HLA-B7, LA-B*0702,HLA-Cw*0401.	46.41
10	Putative uncharacterized protein	Rv2706c	47	SPFSGIFTF	14	HLA-B*3501, HLA-B*5101, HLA-B*5102, HLA-B*5301, HLA-B*5401, HLA-B*51, HLA-B*0702, HLA-Cw*0401, HLA-B*2702, HLA-B*3801, HLA-B*5201, HLA-B*5801, HLA-B62, HLA-Cw*0702.	62.49
11	Putative uncharacterized protein	Rv2706c	66	RPLSGVSFL	21	HLA-B*3501, HLA-B*5101, HLA-B*5102, HLA-B*5103, HLA-B*5301, HLA-B*5401, HLA-B*51, HLA-B7, HLA-B*0702, HLA-Cw*0401, HLA-B*3801,HLA-Cw*0702, HLA-A24, HLA-A2.1, HLA-B14, HLA-B*2705, HLA-B*3701, HLA-B*3901, HLA-B40, HLA-B60, HLA-Cw*0301.	88.40
12	Putative uncharacterized protein	Rv2661c	38	NPQARPREL	12	HLA-A24, HLA-B14, HLA-B*3501, HLA-B*3801, HLA-B*5101, HLA-B*5102, HLA-B*5103, HLA-B*5301, HLA-B7, HLA-B*0702, HLA-B8, HLA-Cw*0401.	67.87
13	Putative uncharacterized protein	Rv2661c	46	LPVLCGWPV	11	HLA-B*3501, HLA-B*5101, HLA-B*5102, HLA-B*5103, HLA-B*5301, HLA-B7, HLA-B*0702, HLA-Cw*0401, HLA-B*5401, HLA-B*51, HLA-B61.	46.50
14	Putative uncharacterized protein	Rv2661c	52	WPVVRVEPV	13	HLA-B14, HLA-B*3501, HLA-B*5101, HLA-B*5102, HLA-B*5103, HLA-B*5301, HLA-B7, HLA-B*0702, HLA-B8, HLA-Cw*0401, HLA-B*5401, HLA-B*51, HLA-B61.	53.17
15	Putative uncharacterized protein	Rv2661c	66	EPVCGQAEV	10	HLA-B*3501, HLA-B*5101, HLA-B*5102, HLA-B*5103, HLA-B*5301, HLA-B7, HLA-Cw*0401, HLA-B*5401, HLA-B*51,HLA-B61.	46.50
16	Uncharacterized protein	Rv3466	8	RIVEVFDAL	12	HLA-A2, HLA-A*0201, HLA-A*0205, HLA-A24, HLA-B*2705, HLA-B*3501, HLA-B*3801, HLA-B*3901, HLA-B*3902, HLA-B7, HLA-B*0702, HLA-Cw*0301.	74.90
17	Uncharacterized protein	Rv3466	38	RSLERLECL	14	HLA-A24, HLA-B*2705, HLA-B*3501, HLA-B*3801, HLA-B*3901, HLA-B7, HLA-B*0702, HLA-Cw*0301, HLA-B14, HLA-B*5801, HLA-B8, HLA-Cw*0401, HLA-B40, HLA-B60.	77.08
18	Uncharacterized protein	Rv3466	49	RLPAVGHAL	11	HLA-A2, HLA-A*0201, HLA-A*0205, HLA-A24, HLA-B*2705, HLA-B*3902, HLA-B7, HLA-B*0702, HLA-A2.1, HLA-B*2702, HLA-B62.	71.02
19	Uncharacterized protein	Rv3466	50	LPAVGHALI	12	HLA-B*3501, HLA-B7, HLA-B*0702, HLA-Cw*0401, HLA-B*3701, HLA-B*5101, HLA-B*5102, HLA-B*5103, HLA-B*5201, HLA-B*5301, HLA-B*5401, HLA-B*51.	48.85

non-antigens gave comparable levels of similarity with the human proteome (Ramakrishnan and Flower, 2010). On the other hand, an inverse relationship has been reported to exist between epitopes' sequence similarity with human proteome and their immunogenicity in the case of MTB (Lucchese et al., 2010). We therefore undertook an integrated approach to identify a set of probable antigenic proteins from the whole genome of the MTB using the principle of exclusion of host mimics, and subjected them through further screening procedures to select the most potent candidates for use in vaccine formulation.

Table 5
Summary of the CTL epitope analysis.

S. No.	Protein name	Protein ID	Total number of epitopes predicited	Total number of HLA alleles predicted to bind with all epitopes	Total number of promiscous epitopes (bind to at least 10 HLA alleles) identified	Total number of HLA alleles predicted to bind with promiscuous epitopes	Theoretical population coverage for the promiscuous epitopes (%)
1	Possible histone like protein	Rv3852	57	38	8	31	92.59
2	Putative uncharacterized protein	Rv2706c	44	38	3	25	85.19
3	Putative uncharacterized protein	Rv2661c	65	36	4	15	71.09
4	Uncharacterized protein	Rv3466	124	39	4	29	91.68

Raman and co-workers (2008) compared MTB proteins with the human proteome and identified 378 MTB proteins to have human homologs. We adopted a similar strategy to exclude homologs proteins but employed a relaxed *E*-value in order to exclude even distantly related proteins. This resulted in the exclusion of 3364 MTB proteins, leaving us with a short list of only 624 MTB proteins for further analysis. Further comparison of these unique proteins with the proteome of 77 microbes that constitute the human gut flora (selected from the list of 95 organisms provided by Raman et al. (2008)), again using a relaxed *E*-value, gave us a set of 180 proteins unique to MTB.

Eighty six of the 180 unique MTB proteins were found to belong to the PE and PPE family (51 PE-PGRS, 19 PPE and 16 PE proteins) and 61 were hypothetical proteins. Cole et al. (1998) showed that about 10% of the *M. tuberculosis* genome codes for two multigene families, PE and PPE, and suggested that these genes could be immunologically significant. Chaitra et al. (2007a) reported that three PE/PPE proteins of *M. tuberculosis,* Rv3018c, Rv1818c and Rv3812, could be potential vaccine candidates. Further, these investigators demonstrated that each of these proteins elicited strong T cell immune responses in mice (Chaitra et al., 2007b, 2008). A recent review reports about 20 PE/PPE proteins, either in the form of whole recombinant proteins or as individual peptides, to have the capacity to elicit CD4 and/or CD8 responses, indicating that PE/PPE proteins are worthy of further evaluation as potentially protective antigens for inclusion in new TB vaccines (Sampson, 2010). Thus, the 86 PE/PPE family proteins in our list could be further investigated for their immunogenic potential. Interestingly, 24 of these 86 proteins have also been reported by others (listed in tables 1–3), highlighting their importance.

Proteins upregulated during dormancy are likely to be exposed to the host immune system for prolonged periods of time, and hence likely to provoke a consistent immunological response. Various biological conditions in which the dormancy-like status is induced include heat shock (Stewart et al., 2002), nutrient starvation (Betts et al., 2002), acidic environment (Fisher et al., 2002), presence of reactive nitrogen intermediates (Ohno et al., 2003), hypoxia (Muttucumaru et al., 2004; Voskuil et al., 2004; Bacon et al., 2004), and presence of H_2O_2 (Schnappinger et al., 2003). Transcriptomics and proteomics studies have provided us with a list of proteins that are highly expressed in *M. tuberculosis* under dormancy-like conditions. These proteins could serve not only as potential immunogens in vaccine formulations but also as diagnostic markers.

We compared the set of unique MTB proteins with a list of proteins already reported to be upregulated in dormant conditions. We found that 29 of the 180 unique proteins were reported in earlier studies to be associated with dormancy (Table 1). Among the 29 proteins, 14 were found to belong to the PE/PPE family (5 PE, 4 PE-PGRS and 5 PPE), 13 were hypothetical and 2 were transmembrane proteins. While five genes, Rv0285 (coding for PE family protein), Rv3852 (coding for histone-like protein HNS), Rv2661c and Rv3466 (hypothetical proteins), and Rv1588c (REP13E12 repeat-containing protein), have been reported to be upregulated in two different dormancy-associated conditions, one protein Rv2706c (hypothetical protein) was found to be upregulated in three different conditions, viz. heat-shock (Stewart et al., 2002), anaerobic environment (Muttucumaru et al., 2004), and presence of H_2O_2 (Schnappinger et al., 2003). These evidences suggest an increased value in investigating these six proteins for their antigenic and diagnostic potential.

Li et al. (2010) performed a proteomic scale analysis of differential response of MTB proteins to serum of TB patients and healthy individuals and identified a list of 249 antigenic proteins. Comparison of the 180 unique MTB proteins with Li's list revealed that 13 proteins were common to both lists (Table 2). Majority of these proteins belong to the PE/PPE/PE-PGRS family of proteins (6 PE-PGRS and 1 PPE), four were hypothetical proteins, one (Rv0476) was a transmembrane protein and one (Rv2571c) was a transmembrane alanine and valine and leucine rich protein. Rv0476 (transmembrane protein) and Rv3760 (hypothetical protein) were highly expressed under nutritional starvation (Betts et al., 2002). Rv1039c (uncharacterized PPE family protein, PPE15) was also identified by Li et al. (2010) as one of the top 20 antigenic proteins.

Tang et al. (2011) identified and reported a set of 248 MTB proteins that contain peptide antigens capable of activating polyfunctional CD8 + T cells, using *in silico* and *in vitro* studies. We compared our list of 180 unique proteins with Tang's list and found 16 proteins to be common between the two. Three of these 16 proteins belong to the PE-PGRS family, one to the PPE family, and three were hypothetical proteins. Of these, Rv0532 (PE-PGRS family protein) was also reported as a reactive antigenic protein by Li et al. (2010). Rv0834c (PE-PGRS family protein) reported to be antigenic by Tang et al. (2011) and also short listed by us, has previously been reported by Voskuil et al. (2004) to be upregulated under hypoxic conditions. Such proteins which have been repeatedly identified as potential antigens by multiple conditions or by multiple investigators can be considered as highly potential candidates, and could be taken up for further investigation as targets for cellular immune response or as diagnostic markers.

There are about 30 tools available for the prediction of CTL epitopes (Lin et al., 2008). ProPred1 is one such tool having sensitivity of 66–78% and specificity of 80–81% (Singh and Raghava, 2003). We employed ProPred1 for the prediction of CTL epitopes from the selected proteins. Nineteen epitopes were found to be promiscuous as they bind to at least 10 different HLA alleles. None of these are listed in either of the two epitope databases, IEDB (www.immune-epitope.org) and AntigenDB (http://www.imtech.res.in/raghava/

antigendb/), and is being reported by us for the first time. A significant amount of polymorphism occurs in the peptide-binding region of MHC, and thus MHC molecules tend to have variable binding specificities. Also the frequency of HLA polymorphism is widely unique across ethnicities (Bui et al., 2006). Hence, it is important to estimate the population coverage for the shortlisted promiscuous epitopes. To accomplish this goal, an online server for population coverage from IEDB (http://tools.immuneepitope.org/tools/population/iedb_input) was used. McMurry et al. (2005) and De Groot et al. (2005) demonstrated that promiscuous epitopes derived from multiple proteins of *M. tuberculosis* stimulate immune response in mice. In this line potential promiscuous epitopes with broader population coverage would be useful candidate epitopes for the design of an effective vaccine.

The protein Rv3852 (possible histone-like protein HNS) was found to have the maximum number of eight promiscuous epitopes binding to 10–17 different HLA alleles. Four epitopes (123:LALLLIRQL, 58:KPAEAPVSL, 118:VTLSLLALL, and 112:VPLIVAVTL) were found to bind to 14–17 of the 39 different HLA alleles included in the analysis, and gave a population coverage of 74.44–77.29%. The total population coverage for the combination of all 8 promiscuous epitopes in Rv3852 was 92.59%.

Protein Rv2706c (putative uncharacterized protein) was found to have three promiscuous epitopes. One epitope (66:RPLSGVSFL) was identified to bind to 21 different HLA alleles giving a population coverage of 88.40%. This epitope is identified as a promiscuous binder to the maximum number of HLAs, and having the highest population coverage in this analysis.

Protein Rv2661c (putative uncharacterized protein) was found to have 4 promiscous epitopes at positions 38, 46 and 52 and 66, binding to 12, 11, 13 and 10 different HLAs, respectively. One epitope (15:NPQARPREL) was predicted to have a population coverage of 67.87%, while the others had <60% coverage. However, the combined population coverage of these four epitopes was found to be 71.09%. Another uncharacterized protein, Rv3466/MT3572, was predicted to have 4 promiscuous epitopes, of which three were predicted to have >70% population coverage. The combined population coverage of the all four promiscuous epitopes in Rv3466 was 91.68%.

Epitopes from two proteins (Rv3466 and Rv3852) were found to have world-wide population coverage of more than 90%, indicating their potential significance as useful candidates for vaccine design. Deenadayalan et al. (2010) reported population coverage for some well known antigenic proteins of *M. tuberculosis*, viz., ESAT-6, CFP-10, and two other proteins AcpM and PpiA. They estimated a coverage of 79.3% for ESAT-6 and 77.62% for CFP-10. Three proteins identified in the present study were found to have >80% population coverage and thus, Rv3852, Rv2706c and Rv3466 appear to be better in this context than the well known antigenic proteins, ESAT-6 and CFP-10. Most of the shortlisted epitopes were found to be conserved in five of the six virulent MTB strains studied. Thus, these proteins and their promiscuous epitopes are worthy of further investigation for their immunogenic relevance.

5. Conclusion

Though the findings of this report are based on a simple and straight forward assumption, 180 proteins have been identified as unique to MTB and most likely to be responsible for provoking the most robust immunological response in humans. Fifty four of these proteins have already been reported as either antigenic or preferentially expressed during specific conditions such as dormancy by other investigators. We also demonstrated the presence of promiscuous epitopes with wider population coverage and conservation in several virulent strains of MTB for four of the selected antigenic proteins using immunoinformatics tools. We would benefit from experimental studies aimed at investigating the immunological relevance of these proteins and novel epitopes.

Acknowledgment

We acknowledge Indian Council of Medical Research for the facility provided through the "ICMR-Biomedical Informatics Project".

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.meegid.2012.03.023.

References

Altschul, S.F. et al., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402.
Amela, I. et al., 2007. Pathogen proteins eliciting antibodies do not share epitopes with host proteins: a bioinformatics approach. PLoS ONE 2 (6), e512.
Bacon, J. et al., 2004. The influence of reduced oxygen availability on pathogenicity and gene expression in *Mycobacterium tuberculosis*. Tuberculosis 84, 205–217.
Betts, J.C. et al., 2002. Evaluation of a nutrient starvation model of *Mycobacterium tuberculosis* persistence by gene and protein expression profiling. Mol. Microbiol. 43, 717–731.
Bui, H.H. et al., 2006. Predicting population coverage of T-cell epitope-based diagnostics and vaccines. BMC Bioinformatics 7, 153.
Chaitra, M.G. et al., 2007a. Modulation of immune responses in mice to recombinant antigens from PE and PPE families of proteins of *Mycobacterium tuberculosis* by the Ribi adjuvant. Vaccine 25, 7168–7176.
Chaitra, M.G. et al., 2007b. Evaluation of T-cell responses to peptides with MHC class I-binding motifs derived from PE_PGRS 33 protein of *Mycobacterium tuberculosis*. J. Med. Microbiol. 56, 466–474.
Chaitra, M.G. et al., 2008. Characterization of T-cell immunogenicity of two PE/PPE proteins of *Mycobacterium tuberculosis*. J. Med. Microbiol. 57, 1079–1086.
Cole, S.T. et al., 1998. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. Nature 393, 537–544.
De Groot, A.S., 2006. Immunomics: discovering new targets for vaccines and therapeutics. Drug Discov. Today 11, 203–209.
De Groot, A.S. et al., 2005. Developing an epitope-driven tuberculosis (TB) vaccine. Vaccine 23 (17–18), 2121–2131.
Deenadayalan, A. et al., 2010. Immunological and proteomic analysis of preparative isoelectric focusing separated culture filtrate antigens of *Mycobacterium tuberculosis*. Exp. Mol. Pathol. 88, 156–162.
Fisher, M.A. et al., 2002. Microarray analysis of the *Mycobacterium tuberculosis* transcriptional response to the acidic conditions found in phagosomes. J. Bacteriol. 184, 4025–4032.
Flower, D.R., 2003. Towards in silico prediction of immunogenic epitopes. Trends Immunol. 24, 667–674.
Gupta, S.K. et al., 2010. In silico CD4+ T-cell epitope prediction and HLA distribution analysis for the potential proteins of *Neisseria meningitidis* Serogroup B – a clue for vaccine development. Vaccine 28, 7092–7097.
Kanduc, D., 2010. The self/nonself issue: a confrontation between proteomes. Self Nonself. 1 (3), 255–258.
Kanduc, D. et al., 2007. Non-redundant peptidomes from DAPs: towards "The Vaccine". Autoimmun. Rev. 6 (5), 290–294.
Kumar, M. et al., 2010a. Immune response to *Mycobacterium tuberculosis* specific antigen ESAT-6 among south Indians. Tuberculosis 90, 60–69.
Kumar, M. et al., 2010b. Cellular immune response to *Mycobacterium tuberculosis*-specific antigen culture filtrate protein-10 in south India. Med. Microbiol. Immunol. 199, 11–25.
Li, Y. et al., 2010. A proteome-scale identification of novel antigenic proteins in *Mycobacterium tuberculosis* toward diagnostic and vaccine development. J. Proteome Res. 9, 4812–4822.
Lienhardt, C., Zumla, A., 2005. BCG: the story continues. Lancet 366, 1414–1416.
Lin, H.H., et al., 2008. Evaluation of MHC class I peptide binding prediction servers: applications for vaccine research. BMC Immunol. 16, 9:8.
Lucchese, G. et al., 2010. Proposing low-similarity peptide vaccines against *Mycobacterium tuberculosis*. J. Biomed. Biotechnol. Article ID 832341, 8.
Ludin, P. et al., 2011. Genome-wide identification of molecular mimicry candidates in parasites. PLoS ONE 6 (3), e17546.
McMurry, J. et al., 2005. Analyzing *Mycobacterium tuberculosis* proteomes for candidate vaccine epitopes. Tuberculosis 85, 95–105.
Mustafa, A.S., Shaban, F.A., 2006. ProPred analysis and experimental evaluation of promiscuous T-cell epitopes of three major secreted antigens of *Mycobacterium tuberculosis*. Tuberculosis 86, 115–124.
Muttucumaru, D.G. et al., 2004. Gene expression profile of *Mycobacterium tuberculosis* in a non-replicating state. Tuberculosis (Edinb) 84, 239–246.

Ohno, H. et al., 2003. The effects of reactive nitrogen intermediates on gene expression in *Mycobacterium tuberculosis*. Cell. Microbiol. 5, 637–648.
Ramakrishnan, K., Flower, D.R., 2010. Discriminating antigen and non-antigen using proteome dissimilarity: bacterial antigens. Bioinformation 4 (10), 445–447.
Raman, K. et al., 2008. TargetTB: a target identification pipeline for *Mycobacterium tuberculosis* through an interactome, reactome and genome-scale structural analysis. BMC Syst. Biol. 2, 109.
Raviglione, M.C., Smith, I.M., 2007. XDR tuberculosis—implications for global public health. J. Med. 356, 656–659.
Ristori, G. et al., 2000. Compositional bias and mimicry toward the nonself proteome in immunodominant T cell epitopes of self and nonself antigens. FASEB J. 14, 431–438.
Rolland, M. et al., 2007. Recognition of HIV-1 peptides by host CTL Is related to HIV-1 similarity to human proteins. PLoS ONE 9, 823.
Salimi, N. et al., 2010. Design and utilization of epitope-based databases and predictive tools. Immunogenetics 62, 185–196.
Sampson, S.L., 2010. Mycobacterial PE/PPE proteins at the host-pathogen interface. Clin. Dev. Immunol. 10, 1155.
Schnappinger, D. et al., 2003. Transcriptional Adaptation of *Mycobacterium tuberculosis* within macrophages: insights into the phagosomal environment. J. Exp. Med. 198, 693–704.
Shah, N.S. et al., 2007. Worldwide emergence of extensively drug-resistant tuberculosis. Emerg. Infect. Dis. 13, 380–387.
Singh, H., Raghava, G.P., 2003. ProPred1: prediction of promiscuous MHC class-I binding sites. Bioinformatics 19, 1009–1014.
Singh, S.P., Mishra, B.N., 2009. Identification and characterization of merozoite surface protein 1 epitope. Bioinformation 4, 1–5.
Stenger, S., Modlin, R.L., 1999. T cell mediated immunity to *Mycobacterium tuberculosis*. Curr. Opin. Microbiol. 2, 89–93.
Stewart, G.R. et al., 2002. Dissection of the heat-shock response in *Mycobacterium tuberculosis* using mutants and microarrays. Microbiology 148, 3129–3138.
Tang, S.T. et al., 2011. Genome-based in silico identification of new *Mycobacterium tuberculosis* antigens activating polyfunctional CD8+ T cells in human tuberculosis. J. Immunol. 186, 1068–1080.
Vani, J. et al., 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, 738–746.
Voskuil, M.I. et al., 2004. *Mycobacterium tuberculosis* gene expression during adaptation to stationary phase and low-oxygen dormancy. Tuberculosis 84, 218–227.
Wang, J. et al., 2010. Analysis of predicted CD8+ T cell epitopes from proteins encoded by the specific RD regions of *Mycobacterium tuberculosis* for vaccine development and specific diagnosis. Mol. Biol. Rep. 37 (4), 1793–1799.
WHO, 2010. <http://whqlibdoc.who.int/publications/2010/9789241500340_eng.pdf> (accessed 18.01.12).
Zvi, A. et al., 2008. Whole genome identification of *Mycobacterium tuberculosis* vaccine candidates by comprehensive data mining and bioinformatic analyses. BMC Med. Genomics. 1, 18.