INFECTION AND IMMUNITY, Dec. 2007, p. 5798–5805 Vol. 75, No. 12
0019-9567/07/$08.00+0 doi:10.1128/IAI.00335-07

DNA Polymorphisms in the *pepA* and PPE18 Genes among Clinical Strains of *Mycobacterium tuberculosis*: Implications for Vaccine Efficacy▽

Andrea M. Hebert,[1] Sarah Talarico,[1] Dong Yang,[1] Riza Durmaz,[2] Carl F. Marrs,[1] Lixin Zhang,[1] Betsy Foxman,[1] and Zhenhua Yang[1]*

Department of Epidemiology, University of Michigan, Ann Arbor, Michigan 48109,[1] and Medical Microbiology, Medical Faculty, Inonu University, Malatya, Turkey[2]

Received 2 March 2007/Returned for modification 2 May 2007/Accepted 3 August 2007

Tuberculosis continues to be a leading cause of death worldwide. Development of an effective vaccine against *Mycobacterium tuberculosis* is necessary to reduce the global burden of this disease. Mtb72F, consisting of the protein products of the *pepA* and PPE18 genes, is the first subunit tuberculosis vaccine to undergo phase I clinical trials. To obtain insight into the ability of Mtb72F to induce an immune response capable of recognizing different strains of *M. tuberculosis*, we investigated the genomic diversity of the *pepA* and PPE18 genes among 225 clinical strains of *M. tuberculosis* from two different geographical locations, Arkansas and Turkey, representing a broad range of genotypes of *M. tuberculosis*. A combination of single nucleotide polymorphisms (SNPs) and insertion/deletions resulting in amino acid changes in the PPE18 protein occurred in 47 (20.9%) of the 225 study strains, whereas SNPs resulted in amino acid changes in the PepA protein in 14 (6.2%) of the 225 study strains. Of the 122 Arkansas study strains and the 103 Turkey study strains, 32 (26.2%) and 15 (14.6%), respectively, had at least one genetic change leading to an alteration of the amino acid sequence of the PPE18 protein, and many of the changes occurred in regions previously reported to be potential T-cell epitopes. Thus, immunity induced by Mtb72F may not recognize a proportion of *M. tuberculosis* clinical strains.

One-third of the world's population is infected with *Mycobacterium tuberculosis*, the causative agent of tuberculosis (TB) (12). The *Mycobacterium bovis* bacille Calmette-Guérin (BCG) vaccine, the most widely used vaccine in the WHO Expanded Programme for Immunization, greatly reduces the risk that infants will acquire severe forms of TB (4, 9, 16, 24). However, trials to determine the efficacy of the BCG vaccine have produced variable results, especially when workers have examined the efficacy of BCG in adults against pulmonary TB, the most common form of the disease (4). This variability of protection highlights the need for a new vaccine against TB.

In recent years, with advances in our knowledge of genomics and proteomics of *Mycobacterium* species and an increased commitment among major research organizations, vaccine research and development have increased (2, 10, 15). Mtb72F, designed using *M. tuberculosis* strain H37Rv as a template, is the first subunit polyprotein vaccine candidate for protection against TB evaluated in clinical trials (26). This vaccine candidate consists of a fusion of protein products of *pepA* (Rv0125) divided into two segments with the product of the PPE18 (Rv1196) gene engineered in the middle. PepA, a putative serine protease, is believed to be a secreted protein, and PPE18, a member of the PPE protein family with an undetermined function, is thought to be associated with the cell membrane (11, 27). These proteins selectively stimulate peripheral blood mononuclear cells (PBMC) from healthy purified protein derivative (PPD)-positive donors to proliferate and secrete gamma interferon (IFN-γ) (11, 27). Studies with animal models challenged with one of the two *M. tuberculosis* laboratory reference strains, H37Rv and Erdmann, demonstrated that Mtb72F had a protective effect, both as a DNA vaccine and as a fusion protein vaccine (22, 23, 26, 30). In addition, priming and boosting effects were observed when Mtb72F was administered after BCG (5, 22, 23, 30). Phase I clinical trials were completed in the United States and Europe (28).

To ensure a high protective efficacy rate among vaccinated hosts, the components of a subunit vaccine should be conserved among infectious strains of an organism (25). The genetic diversity of specific antigens among different strains of human immunodeficiency virus, for example, has made the development of a vaccine to prevent human immunodeficiency virus infection challenging (20). Previous studies concluded that the *M. tuberculosis* genome is relatively stable (21, 29) and that it has fewer polymorphisms than the genomes of other bacteria (14). However, recent comparative genomics studies of *M. tuberculosis* clinical isolates have revealed genetic variations with biological implications (13). Importantly, integration of comparative genomic and epidemiological analysis data has demonstrated that variations in certain genes are associated with various clinical manifestations (17–19, 33).

The genomic variability of the PPE gene family is well documented, as is its possible role as a major source of antigenic variation (10, 14), yet little information regarding possible genetic variability has been documented for *pepA* and the PPE18 gene specifically. If these two genes are highly variable, the effectiveness of the Mtb72F subunit vaccine would potentially be undermined. To obtain insight into the ability of Mtb72F to induce an immune response capable of recognizing different

* Corresponding author. Mailing address: Epidemiology Department, School of Public Health, University of Michigan, 109 S. Observatory Street, Ann Arbor, MI 48109-2029. Phone: (734) 763-4296. Fax: (734) 764-3192. E-mail: zhenhua@umich.edu.

▽ Published ahead of print on 24 September 2007.

strains of *M. tuberculosis* in a natural population, we investigated the genomic diversity of the genes encoding the vaccine's two components, *pepA* and the PPE18 gene, among 122 clinical strains collected from Arkansas and 103 clinical strains collected from Turkey.

MATERIALS AND METHODS

***M. tuberculosis* isolates.** Initially, a study sample consisting of 122 *M. tuberculosis* strains was selected to represent the broad range of strains present in a population-based sample of 705 isolates collected in Arkansas between 1996 and 2000. We then extended our study to obtain broader insight into the global diversity of *M. tuberculosis* clinical strains by including a sample of 103 clinical strains from a convenience sample consisting of 174 isolates from Turkey collected in 2000 and 2003. In order to assess and compare the variability of *pepA* and the PPE18 gene in strains from two geographic regions, thereby providing an assessment of the potential impact of the genetic variability in the two vaccine targets on future vaccination with Mtb72F in different populations, we conducted analyses of the two study samples separately.

The 225 study strains belong to all three of the principal genetic groups used to describe the evolutionary lineages of *M. tuberculosis* isolates, based on single nucleotide polymorphisms (SNPs) in the *katG* and *gyrA* genes as described by Sreevatsan and coworkers (29). However, the distributions of the principal genetic groups differed for the two geographic regions. Of the 122 strains collected in Arkansas, 15 (12.3%) belonged to principal genetic group 1, 77 (63.1%) belonged to principal genetic group 2, and 30 (24.6%) belonged to principal genetic group 3. In contrast, of the 103 strains collected in Turkey, 2 (1.9%) belonged to principal genetic group 1, 63 (61.2%) belonged to principal genetic group 2, and 38 (36.9%) belonged to principal genetic group 3.

The sample of 225 strains also represented both clustered and unique isolates, based on IS*6110* fingerprinting using standardized protocols (31). Of the 122 strains collected in Arkansas, 43 were clustered and 79 were unique strains defined by a combination of IS*6110* restriction fragment length polymorphism analysis and pTBN12 secondary fingerprinting (1, 8, 31). Only one isolate from each cluster was included in the study sample. The demographic data collected indicate that 19 strains included in our study sample were collected from foreign nationals, 101 strains were collected from United States citizens, and 2 were collected from people of unknown origin. Of the 103 strains collected in Turkey, 34 were classified as clustered and the remaining 69 were classified as unique, based on IS*6110* restriction fragment length polymorphism analysis (31).

PCR of the *pepA* and PPE18 genes. Both the *pepA* and PPE18 genes were PCR amplified for DNA sequencing using a BD Advantage-GC 2 PCR kit (BD Biosciences Clontech, Palo Alto, CA). The primers used for *pepA* were pepA-F (5′-TGAGCTGGCGATCTGGACTACG-3′) located 103 bp upstream of the *pepA* gene and pepA-R (5′-CACGCGCACGGGAGACGGAACT-3′) located 94 bp downstream from the end of the *pepA* gene. For the PPE18 gene, the primers were PPE18-F (5′-AAGTGGGCGCTGATTGGGAAGA-3′) located 239 bp upstream of the PE13 gene, the gene directly upstream of the PPE18 gene, and PPE18-R (5′-TGTTGAACCTGGACCTAATACCTG-3′) located 120 bp downstream from the end of the PPE18 gene sequence. The inclusion of the regions adjacent to the *pepA* and PPE18 genes allowed further confirmation that the PCR products obtained were specific for their respective genes. For the PCR protocol, *M. tuberculosis* H37Rv was used as a positive control, and PCR-grade water was used as a negative control. Each standard 50-μl reaction mixture consisted of 10 μl of 5× reaction buffer, 5 μl of GC melt, 20 pmol of each primer in 2 μl, 1 μl of a 50× deoxyribonucleoside triphosphate mixture, 1 μl of 50× BD Advantage 2 polymerase mixture, 5 μl of a DNA solution containing 50 ng of DNA template, and 24 μl of PCR-grade water. The thermocycling program used for each gene was one cycle at 94°C for 1 min; 28 cycles of 94°C for 30 s, 62°C for 30 s, and 72°C for 2.5 min; and a final cycle of 72°C for 10 min. PCR products were examined to determine whether they were the appropriate size by 0.8% (wt/vol) agarose gel electrophoresis performed with 1× Tris-borate-EDTA buffer.

Automated DNA sequencing. PCR products were sequenced to identify any insertions/deletions (in/dels) or SNPs in the *pepA* and PPE18 gene sequences in the study sample. The PCR products used for DNA sequencing were purified using a QIAquick PCR purification kit according to the manufacturer's instructions (QIAGEN Inc., Valencia, CA). DNA sequencing was first performed using the pepA-F and pepA-R primers and the PPE18-F and PPE18-R primers that were used for the two PCR protocols. After completion of the first round of sequence analysis, SNPs were confirmed by double-strand sequencing using the following primers: pepA-C1 (5′-TGGTCAGCACGACACCGTTGGG-3′) located 276 bp downstream from the start of *pepA* and pepA-C2 (5′-CAGGCGT CGGATTCGCTGACC-3′) located 541 bp downstream from the start of *pepA*; and PPE18-C1 (5′-TCGTCGGCGGGTCTGATGGTGG-3′) located 995 bp upstream from the end of the PPE18 gene and PPE18-C2 (5′-TGTTGCAGCGC CTGGGGCACAT-3′) located 628 bp downstream of the start of the PPE18 gene. Sequencing was performed with Applied Biosystems DNA sequencers (models 3700 and 3730) at the Sequencing Core of the University of Michigan. Both the *pepA* and PPE18 gene sequences were compared to those of *M. tuberculosis* reference strain H37Rv (GenBank accession number BX842575) using the BLAST and BLAST2 nucleotide sequence alignment program of the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/BLAST) and the Mutation Surveyor DNA variant analysis software (Demo version) of Softgenetics LLC (www.softgenetics.com).

Prediction of the potential immunogenic effect of the observed amino acid change. In order to describe the possible impact of the changes observed in the amino acid sequence of the PPE18 protein on the ability of a host immunized with Mtb72F to recognize such a protein, it is important to know if the amino acid changes are immunologically significant. To predict the immunological significance of the observed amino acid variations in the PPE18 protein, immunological data for the proliferative response of PBMC from PPD^+ donors incubated with 38 overlapping recombinant PPE18 19-residue peptide fragments reported by Dillon and coworkers (11) were used to generate a T-cell epitope map of PPE18. After the information for the stimulatory responses of the four PPD^+ donors reported by Dillon et al. was combined, each of the overlapping peptide segments was given a T-cell epitope likelihood rating between 0 and 100 that reflected the proportion of donors that had a PBMC stimulatory response of more than 5 stimulation index units when preparations were stimulated with the recombinant PPE18 peptide fragments. The amino acid changes that were observed in our study sample were then mapped to the epitope regions.

RESULTS

Genetic diversity of the *pepA* gene and resulting amino acid sequence variations. Analysis of the sequence variation of *pepA* in the study strains revealed that SNPs were the only cause of genetic variability. Compared to the H37Rv *pepA* sequence, eight SNPs were observed in 40 (17.8%) of the 225 strains in the combined study sample. Five of the eight SNPs were observed exclusively in six (4.9%) of the Arkansas strains, two SNPs were observed exclusively in 26 (25.2%) of the Turkish strains, and one SNP was observed in seven (5.7%) of the Arkansas strains and one (1.0%) of the Turkish strains. Of the six SNPs observed in the Arkansas study sample, one was a synonymous SNP (sSNP) observed in one (0.8%) of the 122 Arkansas strains and five were nonsynonymous SNPs (nsSNPs) causing amino acid changes in 12 (9.8%) of the 122 Arkansas strains. One of the two SNPs found exclusively in the Turkish study sample was an sSNP that was observed in 26 (25.2%) of the 103 Turkish strains. Two (1.9%) of the Turkish strains each had one nsSNP. The one nsSNP that was observed in study strains from both Arkansas and Turkey occurred in the region coding for a putative signal peptide reported to be composed of the first 32 amino acids of the amino acid sequence encoded by *pepA* (27). The remaining nsSNPs were found throughout the gene. With the exception of one Turkish strain that had an sSNP and an nsSNP, none of the strains in the two study samples had more than one nucleotide variation.

Of the 12 Arkansas strains that had genetic variation resulting in amino acid changes, 2 (16.7%) belonged to principal genetic group 2 and 10 (83.3%) belonged to principal genetic group 3. Of the two Turkish strains with genetic variations resulting in amino acid changes, one belonged to principal genetic group 2 and the other belonged to principal genetic group 3.

TABLE 1. Amino acid variation of PPE18 among 122 study strains collected in Arkansas from 1996 to 2000[a]

Amino acid sequence variant	No. of nsSNPs	No. of other amino acid changes (no. of amino acids, location)[b]			No. of strains (%)	Genetic group of affected strains
		Insertion	Deletion	Frameshift[c]		
H37Rv type	0				90 (73.8)	
1	1				5 (4.1)	1
2	1				3 (2.5)	2
3	4				4 (3.3)	1
4	3				2 (1.6)	1
5	36	1 (4, 294)	1 (1, 274)	1 (3, 221–223)	1 (0.8)	2
6	7				1 (0.8)	2
7	3				1 (0.8)	2
8	1	1 (4, 294)			1 (0.8)	3
9	2				1 (0.8)	2
10	31	1 (4, 294)	1 (1, 274)	1 (3, 221–223)	1 (0.8)	3
11			1 (18, 17–34)		2 (1.6)	2
12	11				1 (0.8)	2
13	6				1 (0.8)	3
14	31	1 (4, 294)			1 (0.8)	3
15		1 (2, 162)			1 (0.8)	2
16	35		1 (1, 274)	1 (3, 221–223)	1 (0.8)	1
17	9				1 (0.8)	1
18			1 (2, 161–162)		1 (0.8)	2
19[d]	13				1 (0.8)	2
20	4				1 (0.8)	2
21	3				1 (0.8)	1

[a] The amino acid variants are described based on the number of nsSNPs, the in/dels, and the principal genetic group.
[b] in/dels are described based on the number of amino acids affected and the location in the amino acid sequence, which in H37Rv is 391 amino acids long.
[c] Two consecutive frameshifts were found in the same three strains. An insertion of a guanine at bp 659 within the gene, occurring with deletion of a guanine at bp 669, caused the frameshift to be reverted and ultimately affected three amino acids.
[d] A codon change at position 155 from UAC to UAA resulted in an amino acid change from tyrosine to termination. The remaining 12 nsSNPs shown occur downstream.

Genetic diversity of the PPE18 gene and resulting amino acid sequence variations. Genetic variability of the PPE18 gene, including SNPs and in/dels, was observed in 83 (36.9%) of the 225 study strains. The proportions of strains with PPE18 DNA sequence variation were 40.2 and 33% for the Arkansas sample and the Turkish sample, respectively.

When sequences were compared to the H37Rv PPE18 gene sequence, 105 SNPs were observed in the combined study sample. Of the 105 SNPs, 79 were observed exclusively in the Arkansas strains, 11 were observed exclusively in the Turkish strains, and 15 were observed in both the Arkansas strains and the Turkish strains. Eighty (35.6%) of the 225 strains had at least one SNP (46 [37.7%] strains in the Arkansas study sample and 34 [33%] in the Turkey study sample). Seventy-one (67.6%) of the 105 SNPs were nonsynonymous and were observed in 43 (19.1%) of the 225 strains. Of these 71 nsSNPs, 55 were observed exclusively in the Arkansas strains, 9 were observed exclusively in the Turkish strains, and 7 were observed in both the Arkansas and Turkish strains. in/dels occurred in 11 (4.9%) of the 225 strains (9 [7.4%] of the Arkansas strains and 2 [1.9%] of the Turkey strains). A 12-bp insertion was observed in two (1.9%) of the Turkey strains and four (3.3%) of the Arkansas strains.

When nsSNPs were combined with in/dels, 47 (20.9%) of the combined study strains (32 [26.2%] of the Arkansas study strains and 15 [14.6%] of the Turkey study strains) had at least one genetic change altering the amino acid sequence of the PPE18 protein. When alleles were grouped based on identical amino acid sequences, 21 distinct amino acid sequences were observed for the PPE18 protein in the Arkansas study sample (Table 1) and eight distinct amino acid sequences were observed for the PPE18 protein in the Turkey study sample (Table 2).

The most common amino acid sequence variation in the Arkansas study sample, present in 16 (13.1%) of the 122 strains investigated and in 8 of the 21 amino acid variant groups, was a change from arginine to glutamine. Insertion of a guanine at position 659 in the gene occurred with deletion of a guanine at position 669, causing a frameshift for three amino acids that then reverted to the original sequence. This change occurred in three (2.5%) of the 122 strains, each representing a different amino acid sequence variant due to dissimilarities in the re-

TABLE 2. Amino acid variation of PPE18 among 103 study strains collected in Turkey in 2000 and 2003[a]

Amino acid sequence variant	No. of nsSNPs	No. of insertions (no. of amino acids, location)[b]	No. of strains (%)	Genetic group(s) of affected strains
H37Rv type	0		88 (85.4)	
1	2		5 (4.9)	2 (4 strains), 3
2	3		1 (1.0)	1
3	4		2 (1.9)	2
4	1		2 (1.9)	2, 3
5	1		2 (1.9)	3
6	2	1 (4, 294)	1 (1.0)	3
7	3		1 (1.0)	3
8	8	1 (4, 294)	1 (1.0)	3

[a] The amino acid variants are described based on the number of nsSNPs, the in/dels, and the principal genetic group.
[b] Insertions are described based on the number of amino acids affected and the location in the amino acid sequence, which in H37Rv is 391 amino acids long.

maining mutations. One strain had an nsSNP at nucleotide 465, resulting in a protein product that was 236 amino acids shorter than the PPE18 protein of H37Rv. Eleven of the 21 amino acid sequence variants, representing 12 strains and 9.8% of the study sample, had genetic variations causing five or more amino acid changes in the PPE18 protein sequence.

The most common amino acid change found in the Turkish sample occurred in seven (6.8%) strains and was due to two nsSNPs that occurred in the same codon (codon 322) and were not observed in the Arkansas study sample.

The variant strains in the Arkansas study sample represented all three principal genetic groups. In total, 14 (43.8%) of the 32 variant strains were in genetic group 1, 14 (43.8%) were in genetic group 2, and 4 (12.5%) were in genetic group 3. When the results were stratified by genetic group, variations in the PPE18 gene occurred in most (14/15 [93%]) of the principal genetic group 1 strains. The 14 genetic group 1 strains with amino acid variations included strains with six amino acid variants, amino acid variants 1, 3, 4, 16, 17, and 21 (Table 1). Of the 19 clinical strains included in the study sample from foreign nationals, 10 (52.6%) had amino acid variations.

Of the 15 Turkish strains having PPE18 DNA sequence variations that resulted in amino acid changes, 1 (6.7%), 7 (46.7%), and 7 (46.7%) belonged to principal genetic groups 1, 2, and 3, respectively. One (50%) of the two principal genetic group 1 strains, 7 (11.1%) of the 63 principal genetic group 2 strains, and 7 (18.4%) of the 38 principal genetic group 3 strains from Turkey had amino acid changes.

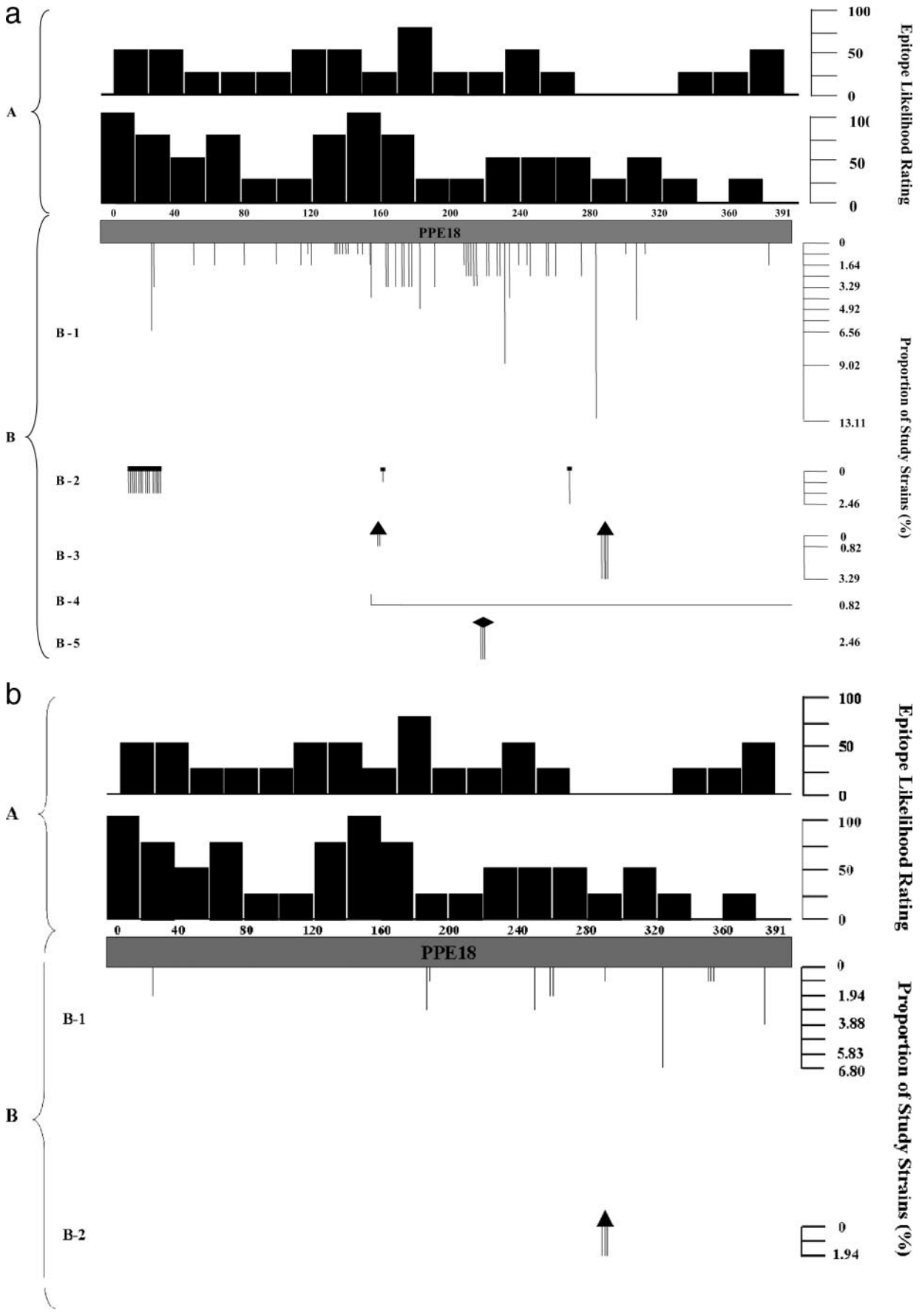
Mapping of amino acid variations to hypothesized T-cell epitopes in the PPE18 amino acid sequence. In order to describe the ability of an immune system primed with Mtb72F to recognize the variations in the amino acid sequence of the PPE18 protein observed in the study sample, an understanding of the potential impact of the amino acid changes on the immunological response induced by vaccination is important. Based on the information published by Dillon and coworkers (11) regarding the immunogenic responses of four PPD^+ donors' PBMC to peptide fragments of PPE18, the amino acid changes observed in the study sample were determined to occur in regions considered to be potential T-cell epitopes. Using the epitope likelihood scale described above, 18 of the 38 overlapping peptide fragments used in the study of Dillon et al. had an epitope likelihood rating of 50 or higher, occurring most frequently in the first 27 fragments or 280 amino acids (Fig. 1a and 1b). In this region, a total of 73 amino acid changes were observed, 48 of which were due to nsSNPs and 26 of which were caused by in/dels. Of the 47 strains with amino acid variations observed in the Arkansas and Turkey study samples, 33 (70.2%) had at least one amino acid change occurring in this region. Two peptide fragments, one containing residues 1 to 21 and one containing residues 141 to 160, resulted in a stimulatory response of >5 for all four PBMC donors and were therefore given an epitope likelihood rating of 100. An 18-amino-acid deletion, occurring from residue 17 to residue 34 and thus including five residues of the first epitope fragment, was found to occur in two strains (1.6%) in the Arkansas study sample. Seven amino acid changes were observed in the fragment from residue 141 to residue 160, occurring in six (4.9%) strains of the Arkansas study sample. One of the amino acid changes observed created a termination codon at residue 155 which occurred in one (0.8%) of the Arkansas study strains.

DISCUSSION

TB remains a major cause of morbidity and mortality worldwide, and with the documented variability of BCG protection against *M. tuberculosis* infection, a new vaccination strategy is needed to prevent continued transmission. To obtain insight into the ability of Mtb72F to induce an immune response capable of recognizing different clinical strains of *M. tuberculosis*, we investigated the genomic diversity for this vaccine's two components, the products of the *pepA* and PPE18 genes, among 225 clinical strains from two different geographic locations. Examining the genetic diversity of clinical isolates can provide insight into an ideal subunit vaccine candidate that provokes an immune response across many different strains. In an investigation to describe the extent of variation in the Mtb72F vaccine target genes among clinical study samples from two different geographical locations, a high frequency of variation was observed in the PPE18 gene of clinical strains that caused amino acid changes in the PPE18 protein product. These changes occurred throughout the protein product, including areas previously proposed to be T-cell epitope regions (11).

As seen in other pathogens, the diversity of specific antigens among strains can hinder the development of efficacious vaccines (20). While only one strain had a genetic mutation that resulted in premature termination of the amino acid sequence, the presence of amino acid changes in the PPE18 protein product in 26.2% of the clinical strains in the Arkansas study sample and in 14.6% of the clinical strains in the Turkey study sample suggests that antigenic diversity of vaccine targets in *M. tuberculosis* could potentially be an issue. Two differences were observed when the genetic similarity between H37Rv and clinical strains from Arkansas and Turkey was examined. First, there was a higher proportion of strains with amino acid changes in the Arkansas study sample than in the Turkey study sample (26.2 and 14.6%, respectively). Second, in the strains with amino acid changes, change was more frequent in the Arkansas strains than in the Turkey strains. For example, the strain with the most frequent amino acid changes in the Arkansas data had 43 amino acids changes, whereas the Turkey strain with the most amino acid changes had 12 changes.

The lower percentage of Turkish strains with amino acid changes may have been due to a number of factors, including the fact that strains belonging to principal genetic group 1 were underrepresented in the Turkish collection and therefore made up only 1.9% of the study sample, compared to the 12.3% of the Arkansas sample strains that were defined as principal genetic group 1 strains, which is representative of the proportion of isolates that are group 1 strains in the population-based sample from Arkansas. If the comparison of the levels of diversity was restricted to only principal genetic groups 2 and 3, the levels appeared to be similar for the two geographic study samples (18.2 and 11.1% for principal genetic group 2 and 13.3 and 18.4% for principal genetic group 3). Furthermore, because the Arkansas study strains were selected from a population-based sample and the Turkey strains were selected from a convenience sample, this may have implica-


a
A
B
B-1
B-2
B-3
B-4
B-5
PPE18
0
40
80
120
160
200
240
280
320
360
391
100
50
0
Epitope Likelihood Rating
0
1.64
3.29
4.92
6.56
9.02
13.11
2.46
0.82
3.29
0.82
2.46
Proportion of Study Strains (%)
b
A
B
B-1
B-2
PPE18
0
40
80
120
160
200
240
280
320
360
391
100
50
0
Epitope Likelihood Rating
0
1.94
3.88
5.83
6.80
1.94
Proportion of Study Strains (%)

tions for the conclusions that can be drawn from examining the Turkey strains alone. Future studies with a population-based sample would allow a more complete inference to be made.

Furthermore, the fact that we observed identical nsSNPs and insertions in both the Arkansas and Turkey strains (seven nsSNPs and one 4-amino-acid insertion affecting 25 of 225 strains [11.1%]) indicates that some of these genetic changes are likely to be found in a large proportion of *M. tuberculosis* clinical strains globally. In addition, the similar distributions of amino acid changes in the last one-third of the PPE18 protein sequence in both the Arkansas and Turkey study samples suggest that this region is affected by amino acid changes consistently.

In vitro studies have indicated that the Mtb72F protein subunits cause variable immune responses when they are used to stimulate the proliferation of PBMC from populations infected with clinical strains of *M. tuberculosis* (11, 27, 32). The original study identifying the PPE18 protein as a potential vaccine candidate demonstrated that there were strong T-cell proliferative and IFN-γ responses in 75% of the 12 PPD^+ donors included in the study (11). The original study identifying the PepA protein as a possible vaccine candidate showed that there was a proliferative PBMC response in up to 50% of the 14 PPD^+ donors included in the study (27). A study conducted in The Gambia showed that there was a variable degree of IFN-γ production by the PBMC of 33 TB patients when they were stimulated with both recombinant PPE18 and PepA proteins, with values ranging from 0 to 5,264 pg/ml (10th to 90th percentiles) (32). Variation in the in vitro proliferation of PPD^+ donor PBMC when the donors were exposed to the PPE18 and *pepA* proteins could have been due to the genetic diversity among clinical isolates of *M. tuberculosis*; however, the strains infecting the PPD^+ donors were not characterized in these previous studies.

The findings regarding the high degree of diversity of the PPE18 gene reflect previous conclusions concerning the PPE gene family as a whole (10, 14), as well as specific PPE genes (7). However, to our knowledge, this is the first time that the diversity of the PPE18 gene among clinical isolates has been described.

The PPE18 gene is a member of a multigene family that includes approximately 70 genes sharing high sequence similarity (6). For example, the PPE19 and PPE60 genes were found to have high levels of nucleotide similarity to the PPE18 gene (92 and 90%, respectively). We ensured the specificity of the PCR product by including the flanking regions of the PPE18 gene. Because the flanking regions of the PPE18 gene are not similar to the flanking regions of the PPE19 and PPE60 genes, we determined that the sequenced PCR products did not include the PPE19 or PPE60 gene. In addition, the reverse primer sequence for the PPE18 gene PCR and the sequence immediately upstream were found to occur twice in the H37Rv genome. However, the sequence similar to the reverse primer region is not immediately adjacent to any other PPE gene, ensuring the accuracy of our PCR amplification.

There was little variation in the *pepA* gene in both study samples that would create an amino acid change. The most common amino acid change, observed in seven (5.7%) of the Arkansas study strains and 0.9% of the Turkey study sample strains, occurred in what has previously been reported to be the signal sequence of PepA (27). The possible implications of this amino acid change for the level of PepA expression remain to be determined.

Future studies to identify the structure, localization, expression, and biological function of the PPE18 protein should allow better assessment of the effect of the observed amino acid diversity among clinical strains on Mtb72F's potential to protect humans from infection by *M. tuberculosis* strains. Specifically, this information should enhance our ability to determine whether the amino acid changes in the PPE18 protein of the study strains change the ability of a primed immune response to recognize the immunogenic portions of the protein. If this occurs, classification of key amino acid changes in PPE18 that lead to host immune system evasion could occur. In addition, immunological studies to identify specific PPE18 peptide fragments as T-cell epitopes in a more numerous PPD^+ donor population could further specify epitope regions where amino acid changes identified could impact the protective immunity induced by Mtb72F vaccination, leading to more specific evaluation of this vaccine candidate.

A large proportion (43.8%) of the *M. tuberculosis* strains from the Arkansas study sample with variations in PPE18 belonged to principal genetic group 1. Although the sample size of the genetic group 1 strains in this study was small, 93.3% (14/15) of the strains classified as members of genetic group 1 had genetic variation causing amino acid differences compared to the amino acid sequence of H37Rv. We decided to examine additional genetic group 1 isolates from our Arkansas collection to see whether we continued to observe such a high frequency of isolates with amino acid changes (data not shown). When comparing 23 additional principal genetic group 1 isolates to H37Rv, we observed amino acid changes in the PPE18

FIG. 1. Mapping of the amino acid changes in the PPE18 protein that were found among the 122 study strains of *M. tuberculosis* from Arkansas (a) and the 103 study strains of *M. tuberculosis* from Turkey (b) in relation to the PPE18 protein sequence of reference strain H37Rv and the T-cell epitope likelihood scale of PPE18 generated using the data reported by Dillon et al. (11). (Panels A) T-cell epitope likelihood score distributions for 38 overlapping peptides of PPE18 described by Dillon et al. (11). The top graph shows the T-cell epitope likelihood score distribution for the overlapping peptide fragments having odd numbers of amino acids (amino acids 11 to 30, 31 to 50, etc.), and the bottom graph shows the T-cell epitope likelihood score distribution for the overlapping peptide fragments having even numbers of amino acids (amino acids 1 to 21, 21 to 40, etc.). The scale on the right ranges from 0 to 100; 0 is equivalent to a stimulatory index of ≤5 for all four donors, and 100 is equivalent to a stimulatory index of >5 for all four donors. (Panels B) Maps of locations of the different amino acid variations among the study strains in relation to the PPE18 protein of H37Rv. Vertical lines extending from the PPE18 amino acid map represent amino acid changes resulting from nsSNPs (B-1). Squares represent amino acid changes caused by nucleotide deletions (panel a, B-2), triangles represent amino acid changes resulting from nucleotide insertions (panel a, B-3; panel b, B-2), the horizontal line represents early termination of transcription (B-4), and the diamond represents two consecutive frameshifts, resulting in three amino acid changes (B-5). Each vertical line represents one amino acid change. The length of a vertical line indicates the proportion of the study strains having the specific change. Percentage scales are shown on the right.

amino acid sequence of 22 (95.7%) of the isolates. In total, 36 of the 38 (94.7%) genetic group 1 isolates from the Arkansas clinical isolate collection examined had at least one amino acid change. The most common amino acid variant, which was also seen in the original Arkansas data set, was a single amino acid change that occurred in the last one-third of the PPE18 amino acid sequence (codon 287) and was an arginine-to-glutamic acid change. While no new SNPs were observed, three new amino acid sequence variants with combinations of previously observed SNPs were found.

While these results are not surprising if one considers that H37Rv is classified as a member of genetic group 3 (29), the diversity among the principal genetic groups with regard to PPE18 could potentially prevent a host's immune system that is primed with Mtb72F from recognizing *M. tuberculosis* strains characterized as members of genetic group 1. This is a major concern due to the fact that genetic group 1 is the predominate genetic group in geographical regions with high TB endemicity (3). Thus, further studies of the diversity of the PPE18 gene using a larger sample of strains from regions where TB is endemic, especially the principal genetic group 1 strains, are in order.

Because our study samples were selected to represent a diverse panel of clinical strains in our collection, our findings may not correspond to the frequency of change of the PPE18 gene and *pepA* in the natural *M. tuberculosis* population. However, Arkansas strains with amino acid changes in PPE18 were responsible for a wide range of cluster sizes, as well as unique strains. Twelve of the 43 (27.9%) clustered strains in the Arkansas sample had amino acid variation. When the PPE18 gene sequences of a few selected isolates from the same cluster were compared, the amino acid changes observed were identical (data not shown). Therefore, it can be assumed that isolates belonging to the same cluster in the Arkansas population-based sample collection are likely to have identical PPE18 gene sequences. Based on this assumption, the estimated frequency of amino acid changes in the clustered strains in the population-based sample is 34.8% (70 of 201 strains), which is similar to the distribution observed in the 43 clustered strains examined (27.9%). Because the Turkish isolate collection was not population based, we were not able to make the same assumption.

The findings of this study suggest that *M. tuberculosis* has genomic regions in which there is high variability, and such genetic variability must be considered when potential vaccine candidates are selected and evaluated. Evaluation of the genetic diversity of clinical strains, like the evaluation done here, could provide important information regarding the ability of the immune response induced by a vaccine candidate to recognize different strains of *M. tuberculosis*.

From the point of view of vaccine candidate evaluation, comparative genomics of clinical strains provides an additional tool for preclinical evaluation of new DNA and subunit vaccines, which is complementary to the traditional preclinical animal and immunological studies of new vaccine candidates. Thus, investigation of the genetic diversity of new vaccine candidate genes in the natural population of *M. tuberculosis* should be considered a further screening method for vaccine candidates proposed for clinical trials in order to avoid conducting clinical trials for a vaccine candidate that may not be effective in some populations due to the genetic diversity of the vaccine target. Given the considerable resources necessary for development and conduct of clinical trials of a new vaccine, the use of comparative genomics as an additional screening tool during preclinical development could potentially allow more efficient use of resources in vaccine development.

To summarize, a significant proportion of both study samples, especially the strains characterized as genetic group 1 strains, had at least one genetic change leading to alteration of the amino acid sequence of the PPE18 protein, and many of the changes occurred in regions previously reported to be potential T-cell epitopes. This suggests that immunity induced by Mtb72F may not recognize a proportion of *M. tuberculosis* strains present in the natural population, especially strains belonging to genetic group 1, which predominate in regions where TB is endemic.

ACKNOWLEDGMENTS

This study was supported by grant NIH-R01-AI151975 from the National Institutes of Health.

We thank Donald Cave of the University of Arkansas for Medical Sciences for helpful discussions during preparation of the manuscript.

REFERENCES

1. **Barnes, P. F., Z. Yang, S. Preston-Martin, J. M. Pogoda, B. E. Jones, M. Otaya, K. D. Eisenach, L. Knowles, S. Harvey, and M. D. Cave.** 1997. Patterns of tuberculosis transmission in central Los Angeles. JAMA **278:** 1159–1163.
2. **Behr, M. A., M. A. Wilson, W. P. Gill, H. Salamon, G. K. Schoolnik, S. Rane, and P. M. Small.** 1999. Comparative genomics of BCG vaccines by whole-genome DNA microarray. Science **284:**1520–1523.
3. **Bifani, P. J., B. Mathema, N. E. Kurepina, and B. N. Kreiswirth.** 2002. Global dissemination of the *Mycobacterium tuberculosis* W-Beijing family strains. Trends Microbiol. **10:**45–52.
4. **Bloom, B., and P. E. M. Fine.** 1994. The BCG experience: implications for future vaccines against tuberculosis, p. 531–557. *In* B. Bloom (ed.), Tuberculosis: protection, pathogenesis, and control. ASM Press, Washington, DC.
5. **Brandt, L., Y. A. Skeiky, M. R. Alderson, Y. Lobet, W. Dalemans, O. C. Turner, R. J. Basaraba, A. A. Izzo, T. M. Lasco, P. L. Chapman, S. G. Reed, and I. M. Orme.** 2004. The protective effect of the *Mycobacterium bovis* BCG vaccine is increased by coadministration with the *Mycobacterium tuberculosis* 72-kilodalton fusion polyprotein Mtb72F in *M. tuberculosis*-infected guinea pigs. Infect. Immun. **72:**6622–6632.
6. **Brennan, M., N. C. Gey van Pittius, and C. Espitia.** 2005. The PE and PPE multigene families of mycobacteria, p. 513–525. *In* S. Cole (ed.), Tuberculosis and the tubercle bacillus. ASM Press, Washington, DC.
7. **Chakhaiyar, P., Y. Nagalakshmi, B. Aruna, K. J. Murthy, V. M. Katoch, and S. E. Hasnain.** 2004. Regions of high antigenicity within the hypothetical PPE major polymorphic tandem repeat open-reading frame, Rv2608, show a differential humoral response and a low T cell response in various categories of patients with tuberculosis. J. Infect. Dis. **190:**1237–1244.
8. **Chaves, F., Z. Yang, H. el Hajj, M. Alonso, W. J. Burman, K. D. Eisenach, F. Dronda, J. H. Bates, and M. D. Cave.** 1996. Usefulness of the secondary probe pTBN12 in DNA fingerprinting of *Mycobacterium tuberculosis*. J. Clin. Microbiol. **34:**1118–1123.
9. **Colditz, G. A., C. S. Berkey, F. Mosteller, T. F. Brewer, M. E. Wilson, E. Burdick, and H. V. Fineberg.** 1995. The efficacy of bacillus Calmette-Guerin vaccination of newborns and infants in the prevention of tuberculosis: meta-analyses of the published literature. Pediatrics **96:**29–35.
10. **Cole, S., R. Brosch, J. Parkhill, T. Garnier, C. Churcher, D. Harris, S. Gordon, K. Eiglmeier, S. Gas, C. E. Barry III, F. Tekaia, K. Badcock, D. Basham, D. Brown, T. Chillingworth, R. Connor, R. Davies, K. Devlin, T. Feltwell, S. Gentles, N. Hamlin, S. Holroyd, T. Hornby, K. Jagels, A. Krogh, J. McLean, S. Moule, L. Murphy, K. Oliver, J. Osborne, M. Quail, M. Rajandream, J. Rogers, S. Rutter, K. Seeger, J. Skelton, R. Squares, S. Squares, J. Sulston, K. Taylor, S. Whitehead, and B. G. Barrell.** 1998. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. Nature **393:**537–544.
11. **Dillon, D. C., M. R. Alderson, C. H. Day, D. M. Lewinsohn, R. Coler, T. Bement, A. Campos-Neto, Y. A. Skeiky, I. M. Orme, A. Roberts, S. Steen, W. Dalemans, R. Badaro, and S. G. Reed.** 1999. Molecular characterization and human T-cell responses to a member of a novel *Mycobacterium tuberculosis* mtb39 gene family. Infect. Immun. **67:**2941–2950.
12. **Dye, C., S. Scheele, P. Dolin, V. Pathania, and M. C. Raviglione.** 1999.

Consensus statement. Global burden of tuberculosis: estimated incidence, prevalence, and mortality by country. WHO Global Surveillance and Monitoring Project. JAMA **282:**677–686.

13. **Filliol, I., A. S. Motiwala, M. Cavatore, W. Qi, M. H. Hazbon, M. Bobadilla del Valle, J. Fyfe, L. Garcia-Garcia, N. Rastogi, C. Sola, T. Zozio, M. I. Guerrero, C. I. Leon, J. Crabtree, S. Angiuoli, K. D. Eisenach, R. Durmaz, M. L. Joloba, A. Rendon, J. Sifuentes-Osornio, A. Ponce de Leon, M. D. Cave, R. Fleischmann, T. S. Whittam, and D. Alland.** 2006. Global phylogeny of *Mycobacterium tuberculosis* based on single nucleotide polymorphism (SNP) analysis: insights into tuberculosis evolution, phylogenetic accuracy of other DNA fingerprinting systems, and recommendations for a minimal standard SNP set. J. Bacteriol. **188:**759–772.
14. **Fleischmann, R. D., D. Alland, J. A. Eisen, L. Carpenter, O. White, J. Peterson, R. DeBoy, R. Dodson, M. Gwinn, D. Haft, E. Hickey, J. F. Kolonay, W. C. Nelson, L. A. Umayam, M. Ermolaeva, S. L. Salzberg, A. Delcher, T. Utterback, J. Weidman, H. Khouri, J. Gill, A. Mikula, W. Bishai, W. R. Jacobs, Jr., J. C. Venter, and C. M. Fraser.** 2002. Whole-genome comparison of *Mycobacterium tuberculosis* clinical and laboratory strains. J. Bacteriol. **184:**5479–5490.
15. **Ginsberg, A.** 2002. What's new in tuberculosis vaccines? Bull. W. H. O. **80:**483–488.
16. **Ginsberg, A. M.** 2000. A proposed national strategy for tuberculosis vaccine development. Clin. Infect. Dis. **30**(Suppl. 3)**:**S233–S242.
17. **Kato-Maeda, M., P. J. Bifani, B. N. Kreiswirth, and P. M. Small.** 2001. The nature and consequence of genetic variability within *Mycobacterium tuberculosis*. J. Clin. Investig. **107:**533–537.
18. **Kong, Y., M. D. Cave, L. Zhang, B. Foxman, C. F. Marrs, J. H. Bates, and Z. H. Yang.** 2007. Association between *Mycobacterium tuberculosis* Beijing/W lineage strain infection and extrathoracic tuberculosis: insights from epidemiologic and clinical characterization of the three principal genetic groups of *M. tuberculosis* clinical isolates. J. Clin. Microbiol. **45:**409–414.
19. **Kong, Y., M. D. Cave, L. Zhang, B. Foxman, C. F. Marrs, J. H. Bates, and Z. H. Yang.** 2006. Population-based study of deletions in five different genomic regions of *Mycobacterium tuberculosis* and possible clinical relevance of the deletions. J. Clin. Microbiol. **44:**3940–3946.
20. **Korber, B., B. Gaschen, K. Yusim, R. Thakallapally, C. Kesmir, and V. Detours.** 2001. Evolutionary and immunological implications of contemporary HIV-1 variation. Br. Med. Bull. **58:**19–42.
21. **Musser, J. M., A. Amin, and S. Ramaswamy.** 2000. Negligible genetic diversity of *Mycobacterium tuberculosis* host immune system protein targets: evidence of limited selective pressure. Genetics **155:**7–16.
22. **Reed, S., and Y. Lobet.** 2005. Tuberculosis vaccine development; from mouse to man. Microbes Infect. **7:**922–931.
23. **Reed, S. G., M. R. Alderson, W. Dalemans, Y. Lobet, and Y. A. Skeiky.** 2003. Prospects for a better vaccine against tuberculosis. Tuberculosis **83:**213–219.
24. **Rodrigues, L. C., V. K. Diwan, and J. G. Wheeler.** 1993. Protective effect of BCG against tuberculous meningitis and miliary tuberculosis: a meta-analysis. Int. J. Epidemiol. **22:**1154–1158.
25. **Scarselli, M., M. M. Giuliani, J. Adu-Bobie, M. Pizza, and R. Rappuoli.** 2005. The impact of genomics on vaccine design. Trends Biotechnol. **23:** 84–91.
26. **Skeiky, Y. A., M. R. Alderson, P. J. Ovendale, J. A. Guderian, L. Brandt, D. C. Dillon, A. Campos-Neto, Y. Lobet, W. Dalemans, I. M. Orme, and S. G. Reed.** 2004. Differential immune responses and protective efficacy induced by components of a tuberculosis polyprotein vaccine, Mtb72F, delivered as naked DNA or recombinant protein. J. Immunol. **172:**7618–7628.
27. **Skeiky, Y. A., M. J. Lodes, J. A. Guderian, R. Mohamath, T. Bement, M. R. Alderson, and S. G. Reed.** 1999. Cloning, expression, and immunological evaluation of two putative secreted serine protease antigens of *Mycobacterium tuberculosis*. Infect. Immun. **67:**3998–4007.
28. **Skeiky, Y. A., and J. C. Sadoff.** 2006. Advances in tuberculosis vaccine strategies. Nat. Rev. Microbiol. **4:**469–476.
29. **Sreevatsan, S., X. Pan, K. E. Stockbauer, N. D. Connell, B. N. Kreiswirth, T. S. Whittam, and J. M. Musser.** 1997. Restricted structural gene polymorphism in the *Mycobacterium tuberculosis* complex indicates evolutionarily recent global dissemination. Proc. Natl. Acad. Sci. USA **94:**9869–9874.
30. **Tsenova, L., R. Harbacheuski, A. L. Moreira, E. Ellison, W. Dalemans, M. R. Alderson, B. Mathema, S. G. Reed, Y. A. Skeiky, and G. Kaplan.** 2006. Evaluation of the Mtb72F polyprotein vaccine in a rabbit model of tuberculous meningitis. Infect. Immun. **74:**2392–2401.
31. **van Embden, J. D., M. D. Cave, J. T. Crawford, J. W. Dale, K. D. Eisenach, B. Gicquel, P. Hermans, C. Martin, R. McAdam, T. M. Shinnick, et al.** 1993. Strain identification of *Mycobacterium tuberculosis* by DNA fingerprinting: recommendations for a standardized methodology. J. Clin. Microbiol. **31:** 406–409.
32. **Vekemans, J., M. O. Ota, J. Sillah, K. Fielding, M. R. Alderson, Y. A. Skeiky, W. Dalemans, K. P. McAdam, C. Lienhardt, and A. Marchant.** 2004. Immune responses to mycobacterial antigens in the Gambian population: implications for vaccines and immunodiagnostic test design. Infect. Immun. **72:**381–388.
33. **Yang, Z., D. Yang, Y. Kong, L. Zhang, C. F. Marrs, B. Foxman, J. H. Bates, F. Wilson, and M. D. Cave.** 2005. Clinical relevance of *Mycobacterium tuberculosis plcD* gene mutations. Am. J. Respir. Crit. Care Med. **171:**1436–1442.

Editor: J. L. Flynn