JOURNAL OF CLINICAL MICROBIOLOGY, Oct. 2005, p. 4954–4960 Vol. 43, No. 10
0095-1137/05/$08.00+0 doi:10.1128/JCM.43.10.4954–4960.2005

Variation of the *Mycobacterium tuberculosis* PE_PGRS33 Gene among Clinical Isolates

Sarah Talarico,[1] M. Donald Cave,[2,3] Carl F. Marrs,[1] Betsy Foxman,[1] Lixin Zhang,[1] and Zhenhua Yang[1]*

Department of Epidemiology, School of Public Health, University of Michigan, Ann Arbor, Michigan,[1] and Department of Neurobiology and Developmental Sciences, College of Medicine, University of Arkansas for Medical Sciences,[2] and Central Arkansas Veterans Healthcare Center,[3] Little Rock, Arkansas

Received 11 May 2005/Returned for modification 16 June 2005/Accepted 5 July 2005

PE_PGRS33, one of about 60 PE_PGRS genes in the *Mycobacterium tuberculosis* genome, encodes a surface-expressed protein that may be involved in the antigenic variation of *M. tuberculosis* strains and evasion of the host immune system. While genetic differences between the PE_PGRS33 genes of H37Rv and CDC1551 have been noted, genetic variation in this gene among clinical isolates has not been evaluated. In order to gain a better understanding of the genetic basis for the role of PE_PGRS in antigenic variation and evasion of the host immune system, we investigated the genetic diversity of the PE_PGRS33 gene among 123 clinical *M. tuberculosis* isolates from a population-based study, using PCR and DNA sequencing. The 123 isolates belonged to principal genetic groups 1, 2, and 3 and had IS*6110* copy numbers ranging from 1 to 22. Eighty-four (68.3%) of the 123 isolates were found to have at least one sequence variation in the PE_PGRS33 gene, relative to that of H37Rv. Twenty-five different sequence variations were observed and included three insertions (ranging from 9 to 87 bp), nine deletions (ranging from 1 to 273 bp), one insertion-and-deletion event, and 12 single-nucleotide polymorphisms (six synonymous and six nonsynonymous). Analysis of the relationships among the different PE_PGRS33 gene sequence variations suggests that polymorphisms in the gene are shifting along evolutionary lineages. The observed genetic diversity of the PE_PGRS33 gene supports its role in antigenic variation and can serve as a basis for future investigations of the function of the PE_PGRS33 gene among clinical isolates.

Sequencing of the *Mycobacterium tuberculosis* genome revealed a unique family of genes, designated PE/PE_PGRS, thought to be involved in antigenic variation of *M. tuberculosis* strains and evasion of the host immune system (6, 13). About 100 PE genes appear in the sequenced *M. tuberculosis* laboratory strain H37Rv and the sequenced *M. tuberculosis* clinical isolate CDC1551 (6, 9). Of these, 37 encode PE alone, while the remainder are associated with a PGRS domain. The highly homologous PE domain is approximately 110 amino acids in length, with a unique proline-glutamic acid (PE) sequence near the amino terminus (3). The PGRS domain varies in size and contains many repeats of alanine and glycine in the form Gly-Gly-Ala-Gly-Gly (3).

Differences among *M. tuberculosis* clinical isolates in restriction fragment length polymorphism patterns determined using PGRS probes and Western analysis patterns determined using cross-reactive antibodies demonstrate variation in size of both the PE_PGRS genes and the expressed PE_PGRS proteins (1, 14, 17), suggesting a possible role in antigenic variation. Comparison of the PE and PE_PGRS gene sequences between H37Rv and CDC1551 showed that all 37 of the PE genes were the same in both strains. However, 39 of the 62 common PE_PGRS genes had differences that would result in the absence of the protein due to frameshift mutations or a difference in size due to insertion-and-deletion events (1). Genetic variation of the PE_PGRS genes is thought to be mediated by deletions or insertions in the glycine-alanine repeats (3).

To date, most of the research on the PE_PGRS of *M. tuberculosis* has been done on PE_PGRS33 (Rv1818c). PE_PGRS33, which has a transmembrane domain, is associated with the cell wall and expressed on the cell surface during infection (7, 8). Transposon mutagenesis of the BCG homologue of the *M. tuberculosis* Rv1818c gene resulted in dispersed growth in liquid media and decreased ability to enter into or survive within macrophages. Complementation of the mutant with the wild-type Rv1818c gene restored the wild-type phenotype, suggesting that PE_PGRS33 may play an important role in the interactions with other mycobacterial cells as well as with macrophages (4).

The Epstein-Barr virus nuclear antigen 1, which has homology to the PGRS domain, interferes with antigen processing and presentation through the major histocompatibility complex class I pathway via a small glycine- and alanine-rich peptide (6, 11). DNA vaccine studies suggest that the PGRS domain of PE_PGRS33 may also be involved in the prevention of antigen processing and presentation of the PE domain in *M. tuberculosis* (7).

In comparison to the PE_PGRS33 of H37Rv, which contains 32 Gly-Gly-Ala-Gly-Gly repeats, the PE_PGRS33 of CDC1551 would show a loss of 30 amino acids and four glycine-alanine repeats at one position and the gain of three amino acids and one glycine-alanine repeat at another (3). However, genetic variation in the PE_PGRS33 gene and any other *M. tuberculosis* PE_PGRS gene among clinical isolates has not been characterized to date.

* 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.

Because of the potential role of PE_PGRS in antigenic variation, the genetic diversity of the PE_PGRS genes among clinical isolates is of interest. Also, because PE_PGRS33 may be involved in prevention of antigen processing (3), genetic variation of the PE_PGRS33 gene among clinical isolates could potentially account for some of the differences in their ability to evade the host immune system. In order to gain a better understanding of the genetic basis of the interactions between *M. tuberculosis* and the host and the role of PE_PGRS in antigenic variation and evasion of the host immune system, we investigated the genetic diversity of the PE_PGRS33 gene among 123 clinical *M. tuberculosis* isolates.

MATERIALS AND METHODS

***M. tuberculosis* isolates.** A study sample of 123 *M. tuberculosis* isolates was selected from 705 isolates collected in Arkansas between 1996 and 2000 to represent a broad range of strains found in the population-based collection of isolates from the Arkansas Department of Health molecular epidemiology database. The 123 isolates contained 43 isolates representing 43 clusters and 80 unique isolates defined by a combination of IS*6110* restriction fragment length polymorphism analysis and pTBN12 secondary fingerprinting (2, 5, 16).

Assignment of strains to three principal genetic groups. Genotypic grouping based on single-nucleotide polymorphisms (SNPs) found in the *katG* and *gyrA* genes, as described by Sreevatsan et al. (15), was also used to assess the genetic relatedness of the isolates. Codon 463 of the *katG* gene and codon 95 of the *gyrA* gene were PCR amplified using the BD Advantage 2 PCR kit (BD Biosciences Clontech, Palo Alto, CA). The primers used to amplify the portion of the *katG* gene were kat*G*F (5′-AGC CGC CTT TGC TGC TTT CTC TA-3′) and kat*G*R (5′-TGC TGG CCA CTG ACC TCT CGC T-3′), located 547 bp and 1,094 bp upstream of the end of the *katG* gene sequence, respectively. The primers used to amplify the portion of the *gyrA* gene were gyr*A*F (5′-AAC CGG TTG ACA TCG AGC AGG AGA-3′) and gyr*A*R (5′-ATT TCC CTC AGC ATC TCC ATC-3′), located 47 bp and 434 bp downstream of the beginning of the *gyrA* gene sequence, respectively. Each standard 50-μl reaction mixture consisted of 5 μl of 10× reaction buffer, 20 pmol of each primer in 2 μl, 1 μl of a 50× deoxyribonucleoside triphosphate mix, 1 μl of 50× BD Advantage 2 polymerase mix, 4 μl DNA solution containing 40 ng DNA template, and 35 μl PCR-grade water. The thermocycling program used was 1 cycle at 94°C for 1 min; 26 cycles of 94°C for 30 seconds, 68°C for 30 seconds, and 72°C for 1.5 min; and a final cycle at 72°C for 10 min. The PCR products were purified using a QIAquick PCR purification kit according to the manufacturer's instructions (QIAGEN Inc., Valencia, CA) and sequenced using the same primers used for PCR amplification.

PCR of the PE_PGRS33 gene. The PE_PGRS33 gene was PCR amplified for DNA sequencing using the BD Advantage-GC 2 PCR kit (BD Biosciences Clontech, Palo Alto, CA). We selected primers specific for amplification of the PE_PGRS33 gene using the BLAST program of the National Center for Biotechnology Information (www.ncbi.nlm.nih.gov/BLAST). The primers were PE_PGRS33F (5′-CTA CGG TAA CCC GTT CAT CCC-3′), located at the end of the PE_PGRS33 gene sequence, and PE_PGRS33R (5′-GCG CCC GCC GAA GTG TAA G-3′), located 152 bp upstream of the beginning of the PE_PGRS33 gene sequence. The inclusion of the 152-bp flanking region, positions 2062675 through 2062826 of the H37Rv complete genome sequence (NC_000962), allowed for further confirmation during sequence analysis that the sequences obtained were specific for the PE_PGRS33 gene. *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 mix, 1 μl of 50× BD Advantage 2 polymerase mix, 6 μl DNA solution containing 60 ng DNA template, and 23 μl PCR-grade water. The thermocycling program used was 1 cycle at 94°C for 1 min; 30 cycles of 94°C for 30 seconds, 64°C for 30 seconds, and 72°C for 2.5 min; and a final cycle at 72°C for 10 min. All PCR amplification was performed using a 96-well Perkin-Elmer thermocycler (P-E 960; Applied Biosystems, Foster City, CA). PCR products were examined 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, or SNPs in the PE_PGRS33 gene sequence. 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 PE_PGRS33F primer and the PE_PGRS33R primer that were used for the PCR. After the completion of the first round of sequence analysis, a second round of sequencing was performed using the PGRS0660R primer (5′-CGG CGG AGA CGG CGG GTT GTT-3′), located 739 bp upstream of the end of the PE_PGRS33 gene sequence, to sequence the end of the PE_PGRS33 gene and also to confirm the SNPs found during the first round of sequencing. The primers PGRS0778F (5′-CAC CAA TAC CGC CCA CCC CAC CAC-3′) and PGRS0778R (5′-GTG GTG GGG TGG GCG GTA TTG GTG-3′), located 857 bp upstream of the end of the PE_PGRS33 gene, were also used to confirm SNPs. All SNPs were confirmed by double-strand sequencing. Sequencing was performed in Applied Biosystems DNA sequencers (models 3700 and 3730 sequencers) at the Sequencing Core of the University of Michigan. The PE_PGRS33 gene sequences were compared to that of the *M. tuberculosis* reference strain H37Rv using the BLAST program of the National Center for Biotechnology Information (www.ncbi.nlm.nih.gov/BLAST).

Data analysis. A dendrogram illustrating the relationships between the 23 different PE_PGRS33 alleles found among the 123 isolates was constructed using MEGA version 3.0 (10). The dendrogram was generated using the neighbor-joining method, and the distance was calculated using the number of sequence variations found in each of the 23 PE_PGRS33 alleles. Deletions, insertions, and SNPs were included in the analysis, and each was scored as one change in sequence.

RESULTS

Genetic relatedness of study isolates. The IS*6110* copy numbers of the 123 isolates ranged from 1 to 22, representing 111 different IS*6110* banding patterns. Isolates having identical IS*6110* banding patterns with fewer than six hybridizing bands were differentiated by pTBN12 typing. Of the 123 isolates, 38 (30.9%) were low-copy-number strains, having five or fewer copies of IS*6110*, and 85 (69.1%) were high-copy-number strains.

Genotypic grouping based on the SNPs found in the *katG* and *gyrA* genes, as described by Sreevatsan et al. (15), placed 16 (13.0%), 79 (64.2%), and 28 (22.8%) of the 123 isolates into principal genetic groups 1, 2, and 3, respectively. Groups 1 and 2 contained both low- and high-copy-number isolates; in contrast, the group 3 isolates were all high-copy-number isolates.

Genetic diversity of the PE_PGRS33 gene. Relative to the sequence of H37Rv, 84 (68.3%) of the 123 isolates had at least one sequence variation in the PE_PGRS33 gene, and these 84 isolates included both high-copy-number and low-copy-number strains. The 39 (31.7%) isolates that did not have any sequence variations were all high-copy-number strains, harboring more than five copies of IS*6110*. A total of 25 different sequence variations, designated A through Y, were observed. The observed sequence variations included three insertions, nine deletions, one insertion-and-deletion event, and 12 SNPs (Table 1).

Insertions and deletions in the PE_PGRS33 gene. The insertions and deletions all occurred within the PGRS domain, and all had an effect on one or more of the glycine-alanine repeats (Fig. 1). The insertions were repeats of 9, 18, and 87 bp of the region adjacent to the insertion site. The deletions ranged in size from 1 to 273 bp (Table 1). One of the isolates had a deletion event and an insertion event at the same position (sequence variation M). In contrast to the other insertions, this insertion was a repeat of a region of the PE_PGRS33 gene (positions 611 through 642 of the H37Rv PE_PGRS33 gene sequence) that was not adjacent to the position of the insertion.

While most of the insertions and deletions were in frame,

TABLE 1. Characterization of sequence variations within the PE_PGRS33 gene among 123 clinical *M. tuberculosis* isolates

Category	Sequence variation	Genomic change	Position(s)[a]	Effect on amino acid sequence	Loss or gain of glycine-alanine repeat unit[b]	No. (%) of strains (n = 123)
Insertion	A	9-bp insertion	1240	Repeat of Gly, Ala, Gly	+1	65 (52.8)
	B	18-bp insertion	597	Repeat of Gly, Ala, Gly, Gly, Ala, Gly	+2	1 (0.8)
	C	87-bp insertion	1039	Repeat of 29 amino acids	+5	2 (1.6)
Deletion	D	1-bp deletion	1014	Frameshift resulting in different amino acid sequence and premature stop	−12	16 (13.0)
	E	9-bp deletion	651–659	Deleted Gly, Ala, Gly	−1	7 (5.7)
	F	9-bp deletion	681–689	Deleted Gly, Ala, Gly	−1	2 (1.6)
	G	9-bp deletion	556–564	Deleted Gly, Gly, Ala	−1	1 (0.8)
	H	18-bp deletion	591–608	Deletion of Ala, Gly, Gly, Ala, Gly, Gly	−2	1 (0.8)
	I	90-bp deletion	550–639	Deletion of 30 amino acids	−4	4 (3.3)
	J	96-bp deletion	1114–1209	Deletion of 32 amino acids	−2	1 (0.8)
	K	144-bp deletion	585–728	Deletion of 48 amino acids	−5	1 (0.8)
	L	273-bp deletion	709–981	Deletion of 91 amino acids	−9	1 (0.8)
Insertion and deletion	M	107-bp deletion with 32-bp insertion	710–816 deleted; insertion at 817	Deletion of 36 amino acids and addition of 11 amino acids	−3	1 (0.8)
SNP	N	T → C	717	Ala → Ala	—[c]	65 (52.8)
	O	T → C	837	Asn → Asn	—	1 (0.8)
	P	T → C	1224	Gly → Gly	—	1 (0.8)
	Q	C → T	81	Thr → Thr	—	1 (0.8)
	R	C → T	48	Thr → Thr	—	1 (0.8)
	S	C → T	47	Thr → Ile	—	1 (0.8)
	T	G → A	1166	Gly → Asp	—	3 (2.4)
	U	T → G	529	Ser → Ala	—	3 (2.4)
	V	C → T	1172	Thr → Ile	—	1 (0.8)
	W	T → C	701	Val → Ala	+1	2 (1.6)
	X	G → C	959	Gly → Ala	−1	1 (0.8)
	Y	G → C	1068	Gly → Gly	Neutral	1 (0.8)

[a] Positions based on the *M. tuberculosis* H37Rv PE_PGRS33 gene sequence (Rv1818c).
[b] The glycine-alanine repeat unit was defined as Gly-Gly-Ala-Gly-Gly.
[c] —, not found within a glycine-alanine repeat.

the deletion of 1 bp at position 1014 (sequence variation D) resulted in a frameshift. The change in frame would result in a change in sequence for 36 amino acids and then a premature stop codon with the loss of 124 amino acids (Table 1). This was also the only sequence variation found that would result in a change to the 258 bp that encode the carboxy-terminal end of the PGRS domain.

SNPs in the PE_PGRS33 gene. Of the 12 SNPs observed, 9 occurred in the PGRS domain and 3 in the PE domain (Fig. 1). Of the nine SNPs found in the PGRS domain, four were synonymous and five were nonsynonymous. Three of the nine SNPs in the PGRS domain were located within a glycine-alanine repeat unit, defined as Gly-Gly-Ala-Gly-Gly (3). Of these three, one was synonymous, and the other two would result in a loss of a repeat and the gain of a repeat unit, respectively. Of the three SNPs that were found in the PE domain (sequence variations Q, R, and S), two were synonymous and one was nonsynonymous. Two of the SNPs found in the PE domain (sequence variations R and S) were in the same isolate in consecutive positions (Table 1).

Combinations of sequence variations among the 123 isolates. Thirty-nine (31.7%) of the 123 isolates had PE_PGRS33 gene sequences identical to that of H37Rv. Among the remaining 84 (68.3%) isolates, 22 different combinations of sequence variations were observed. Sixty-eight (81.0%) of the 84 isolates had more than one sequence variation (Fig. 2 and Table 2).

The two most frequent sequence variations (sequence variations A and N) were found together in 64 (52.0%) of the 123 isolates. Although these two sequence variations were distributed throughout the high-IS*6110*-copy-number isolates, all 38 of the low-copy-number isolates had either sequence variation A or N. The sequence variations A, N, and I are also found in the *M. tuberculosis* isolate CDC1551. Four (3.3%) of the 123 isolates shared the same PE_PGRS33 gene sequence variation pattern with CDC1551. Sequence variation D, the deletion of 1 bp resulting in a premature stop codon, always occurred with either sequence variation A or N (Fig. 2 and Table 2) and was found exclusively in 16 (42.1%) of 38 low-copy-number isolates.

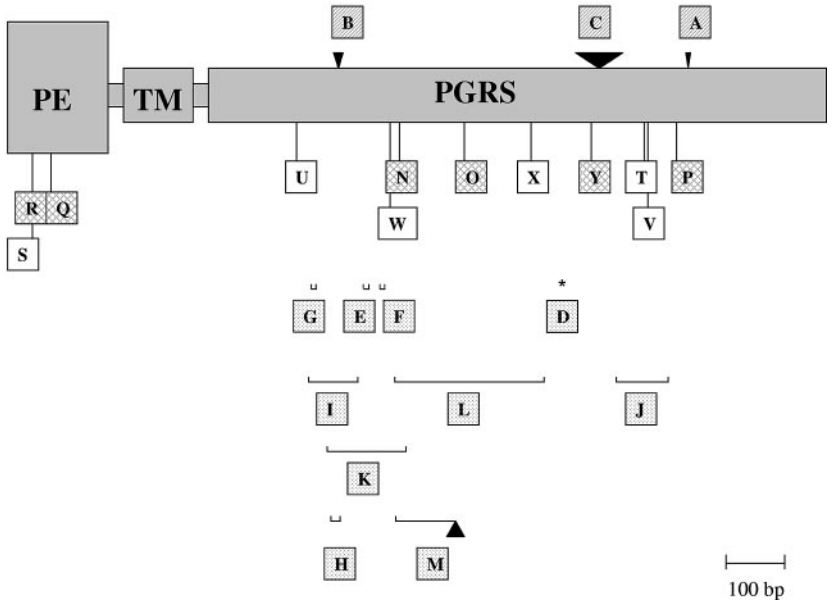

FIG. 1. Map of the positions of different sequence variations found in the *M. tuberculosis* PE_PGRS33 gene. Triangles represent insertions, brackets represent deletions, and lines extending from the PE_PGRS map represent SNPs. The asterisk represents a deletion of 1 bp. The sequence variations were designated A through Y and are described in Table 1.

Relationships among the PE_PGRS33 sequence variations. Based on the analysis of SNPs in the *katG* and *gyrA* genes, *M. tuberculosis* strains can be placed in three principal genetic groups. Principal genetic group 1 is ancestral to groups 2 and 3, and principal genetic group 2 is ancestral to group 3 (15). A dendrogram illustrating the relationships among the 23 different alleles of the PE_PGRS33 gene shows a clear separation of principal genetic group 1 from group 3 but no separation between groups 1 and 2 or between groups 2 and 3 (Fig. 2).

Based on two SNPs, one in the *katG* gene and one in the *gyrA* gene, the *M. tuberculosis* 210 strain (a widespread member of the Beijing family), CDC1551, and H37Rv belong to principal genetic groups 1, 2, and 3, respectively. The 16 group 1 isolates all had either sequence variation A or N, and 14

TABLE 2. Frequency of different PE_PGRS33 alleles found among 123 clinical *M. tuberculosis* isolates and the effects of the combinations of sequence variations on the PE_PGRS33 gene product

Sequence variation(s)	No. (%) of isolates (n = 123)	Net change in no. of amino acids	Net change in glycine-alanine repeat units	Change in amino acid
None (identical to Rv1818c)[a]	39 (31.7)	Reference	Reference	Reference
A and N	40 (32.5)	+3	+1	
A, N, and D	10 (8.1)	−124	−12	
A, N, D, and G	1 (0.8)	−127	−13	
A, N, D, and T	3 (2.4)	−124	−12	
A, N, and I	4 (3.3)	−27	−3	
A, N, and B	1 (0.8)	+9	+3	
A, N, and V	1 (0.8)	+3	+1	Thr → Ile
A, N, and O	1 (0.8)	+3	+1	
A, N, and Q	1 (0.8)	+3	+1	
A, N, and H	1 (0.8)	−3	−1	
A, N, R, and S	1 (0.8)	+3	+1	Thr → Ile
A, D, W, and L	1 (0.8)	−215	−20	Val → Ala
U	2 (1.6)	0	0	Ser → Ala
U and J	1 (0.8)	−32	−2	Ser → Ala
M and W	1 (0.8)	−25	−2	Val → Ala
N, D, and P	1 (0.8)	−124	−12	
E	7 (5.7)	−3	−1	
F	2 (1.6)	−3	−1	
C	2 (1.6)	+29	+5	
X	1 (0.8)	0	−1	Gly → Ala
Y	1 (0.8)	0	0	
K	1 (0.8)	−48	−5	

[a] The Rv1818c sequence is the reference.

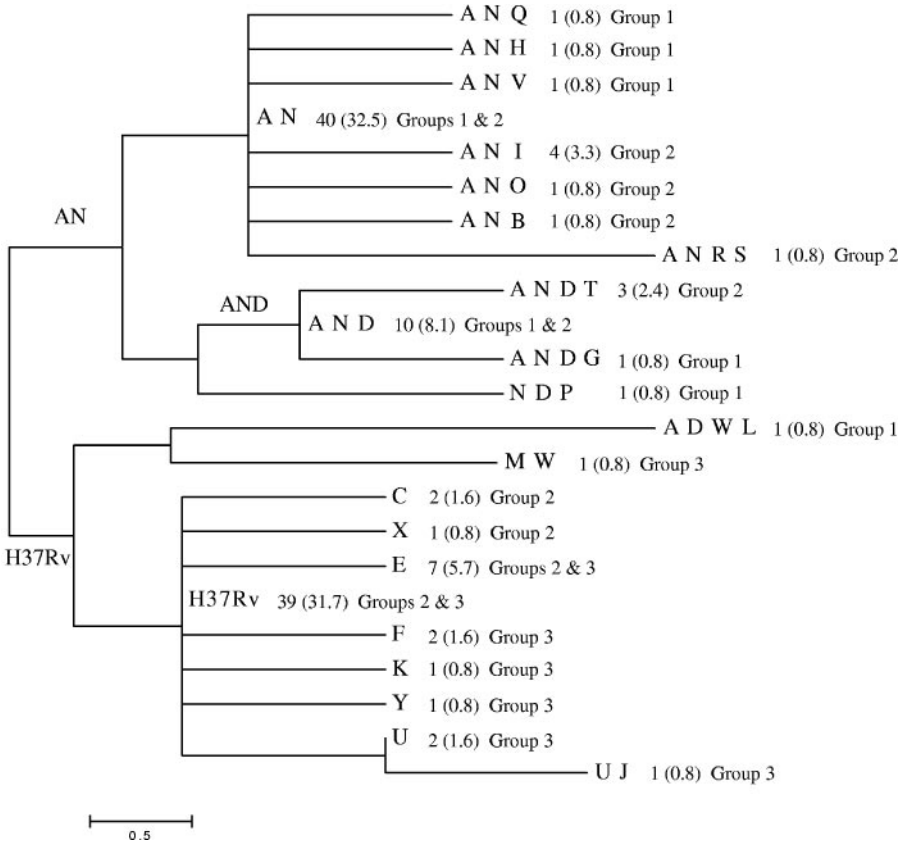

FIG. 2. Dendrogram showing the genetic relationships among the 23 different PE_PGRS33 alleles based on the analysis of deletions, insertions, and SNPs within the PE_PGRS33 gene that were found among the 123 clinical *M. tuberculosis* isolates. The sequence variations found among the 123 isolates were designated A through Y and are described in Table 1. For each of the 23 PE_PGRS33 alleles, the number and percentage of the 123 clinical *M. tuberculosis* isolates having the allele and the principal genetic group based on SNPs in the *katG* and *gyrA* genes (15) to which the isolates belong are indicated.

(87.5%) of these 16 isolates had both sequence variations A and N, which are found in both the 210 strain and CDC1551. Among the 79 group 2 isolates, 50 (63.3%) shared the A and N sequence variations with the 210 strain and CDC1551, and 20 (25.3%) had PE_PGRS33 gene sequences identical to that of H37Rv. In addition, four (5.1%) of the group 2 isolates had PE_PGRS33 gene sequences identical to that of CDC1551. Of the 28 group 3 isolates, 19 (67.9%) had PE_PGRS33 gene sequences identical to that of H37Rv, and none had either the A or N sequence variations found in the 210 strain and CDC1551 (Fig. 2).

Effects of sequence variations on the resulting PE_PGRS33 amino acid sequence. The combination of sequence variations in each isolate was analyzed to examine the overall effect on the PE_PGRS33 gene product (Table 2). Some of the sequence variations occurring with sequence variation D (sequence variations T, P, and A) would not have any effect on the resulting amino acid sequence in these isolates because these sequence variations are downstream of sequence variation D, which results in a premature stop codon. Sequence variations T and P were found exclusively in isolates that also had sequence variation D, and, therefore, these sequence variations would not have any effect on the PE_PGRS33 amino acid sequence in any of the 123 isolates. Eighty-three (98.8%) of the 84 isolates having a sequence variation in the PE_PGRS33 gene would have a resulting PE_PGRS33 amino acid sequence different from that of H37Rv. The isolate containing sequence variation Y was the only isolate that had a sequence variation in the PE_PGRS33 gene but would have a resulting PE_PGRS33 amino acid sequence identical to that of H37Rv.

DISCUSSION

The variation in the PE_PGRS33 gene is extensive compared to the limited genetic variability in the *M. tuberculosis* genome as a whole. Eighty-four (68.3%) of 123 isolates had at least one sequence variation in the PE_PGRS33 gene in relation to the PE_PGRS33 gene of H37Rv. Twenty-five different sequence variations were found and included three insertions, nine deletions, one insertion-and-deletion event, six synonymous SNPs, and six nonsynonymous SNPs. These findings are in contrast to a previous study that found limited genetic diversity among genes that code for antigenic proteins, including some members of the PE and PPE gene families, suggesting that the immune system exerts a limited selective pressure on genes that code for targets of the host immune system (12). The extensive genetic variability found in the PE_PGRS33

gene supports the role of the PE_PGRS family in antigenic variation.

Based on the analysis of SNPs in the *katG* and *gyrA* genes, *M. tuberculosis* strains can be placed in three principal genetic groups. It is proposed that principal genetic group 1, containing the *M. tuberculosis* 210 strain, is ancestral to groups 2 and 3 and that principal genetic group 2, containing *M. tuberculosis* CDC1551, is ancestral to group 3, which contains *M. tuberculosis* H37Rv (15). The dendrogram in Fig. 2 identifies two major branches of the PE_PGRS33 alleles. Genetic group 1 and group 3 isolates are exclusively (except for one genetic group 1 isolate) in one of these two allele branches, while group 2 isolates fall into both of the two branches. The emerging genetic group 3 isolates are associated with new PE_PGRS33 alleles that are distant from alleles found in genetic group 1 isolates but still close to alleles found in genetic group 2 isolates. This suggests that genetic group 1 isolates and genetic group 3 isolates are evolutionarily linked through genetic group 2 isolates. The analysis of the genetic relationships among the different PE_PGRS33 alleles lends further support to the three principal genetic groups proposed by Sreevatsan et al. (15), because the analysis is based on all the sequence variations present within the PE_PGRS33 gene among clinical *M. tuberculosis* isolates, as opposed to previously determined markers based on comparison of the two sequenced genomes.

The positions of the sequence variations found within the PE_PGRS33 gene among the 123 clinical isolates could be informative of the importance of certain regions of the protein. The carboxy-terminal end of the PGRS domain and the transmembrane domain were highly conserved, suggesting that these regions may have an important function. However, although there were no sequence variations found within the 258 bp that encode the last 86 amino acids of the PGRS domain, sequence variation D would result in the loss of the last 124 amino acids resulting from the frameshift-mediated premature stop codon. Sequence variation D was observed in 16 (13.0%) of the 123 isolates; however, the frequency of this sequence variation in the population of *M. tuberculosis* has not been determined. Future studies that investigate the pathogenicity of isolates with this truncated PGRS domain in comparison to that of isolates without this sequence variation may be informative of the role of the different regions of PE_PGRS33 in pathogen-host interactions.

Both slipped-strand mispairing during replication and homologous recombination of repetitive sequences could potentially account for the high frequency of insertions and deletions found in the PGRS domain of the PE_PGRS33 gene (1, 13). With the exception of the frameshift mutation, the insertions and deletions found within the PGRS domain in this study did not change the reading frame. This suggests that PE_PGRS33 may be important to the survival of *M. tuberculosis* and that there is a selective pressure to maintain the reading frame of this gene. It is possible that genetic variation of the PE_PGRS33 gene is advantageous to the tubercle bacilli because it aids in escape from the host immune system, but changes that alter the reading frame are selected against because PE_PGRS33 has other essential functions, such as facilitating interactions with and surviving within macrophages.

Future studies are needed to investigate the effect of these PE_PGRS33 alleles on the interaction of the tubercle bacillus with macrophages and evasion of the host immune system. While transposon mutagenesis of the PE_PGRS33 gene results in a decreased ability of *M. bovis* BCG to enter into or survive within macrophages (4), it is not clear what effect sequence variations in the PE_PGRS33 gene found among *M. tuberculosis* clinical isolates have on interactions with and survival within the host. Some of the PE_PGRS33 gene sequence variations found among clinical isolates could potentially account for some of the differences among strains in their ability to evade the host immune system and can serve as a basis for future investigations of the differences in function of *M. tuberculosis* PE_PGRS33 among clinical isolates. Studies of the associations between the genetic polymorphisms of the PE_PGRS33 gene and the clinical phenotypes of the isolates will also generate information useful for the development of new vaccines and diagnostic and therapeutic agents.

ACKNOWLEDGMENTS

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

We thank Dong Yang for her assistance in DNA preparation and *M. tuberculosis* culturing.

REFERENCES

1. **Banu, S., N. Honore, B. Saint-Joanis, D. Philpott, M. C. Prevost, and S. T. Cole.** 2002. Are the PE-PGRS proteins of *Mycobacterium tuberculosis* variable surface antigens? Mol. Microbiol. **44:**9–19.
2. **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.
3. **Brennan, M. J., and G. Delogu.** 2002. The PE multigene family: a 'molecular mantra' for mycobacteria. Trends Microbiol. **10:**246–249.
4. **Brennan, M. J., G. Delogu, Y. Chen, S. Bardarov, J. Kriakov, M. Alavi, and W. R. Jacobs, Jr.** 2001. Evidence that mycobacterial PE_PGRS proteins are cell surface constituents that influence interactions with other cells. Infect. Immun. **69:**7326–7333.
5. **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.
6. **Cole, S. T., R. Brosch, J. Parkhill, T. Garnier, C. Churcher, D. Harris, S. V. 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. Hornsby, K. Jagels, A. Krogh, J. McLean, S. Moule, L. Murphy, K. Oliver, J. Osborne, M. A. Quail, M.-A. Rajandream, J. Rogers, S. Rutter, K. Seeger, J. Skelton, R. Squares, S. Squares, J. E. 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.
7. **Delogu, G., and M. J. Brennan.** 2001. Comparative immune response to PE and PE_PGRS antigens of *Mycobacterium tuberculosis*. Infect. Immun. **69:** 5606–5611.
8. **Delogu, G., C. Pusceddu, A. Bua, G. Fadda, M. J. Brennan, and S. Zanetti.** 2004. Rv1818c-encoded PE_PGRS protein of *Mycobacterium tuberculosis* is surface exposed and influences bacterial cell structure. Mol. Microbiol. **52:** 725–733.
9. **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.
10. **Kumar, S., K. Tamura, and M. Nei.** 2004. MEGA3: Integrated software for molecular evolutionary genetics analysis and sequence alignment. Brief. Bioinform. **5:**150–163.
11. **Levitskaya, J., A. Sharipo, A. Leonchiks, A. Ciechanover, and M. G. Masucci.** 1997. Inhibition of ubiquitin/proteasome-dependent protein degradation by the Gly-Ala repeat domain of the Epstein-Barr virus nuclear antigen 1. Proc. Natl. Acad. Sci. USA **94:**12616–12621.
12. **Musser, J. M., A. Amin, and S. Ramaswamy.** 2000. Negligible genetic diver-

sity of *Mycobacterium tuberculosis* host immune system protein targets: evidence of limited selective pressure. Genetics **155:**7–16.

13. **Poulet, S., and S. T. Cole.** 1995. Characterization of the highly abundant polymorphic GC-rich-repetitive sequence (PGRS) present in *Mycobacterium tuberculosis*. Arch. Microbiol. **163:**87–95.
14. **Ross, B. C., K. Raios, K. Jackson, and B. Dwyer.** 1992. Molecular cloning of a highly repeated DNA element from *Mycobacterium tuberculosis* and its use as an epidemiological tool. J. Clin. Microbiol. **30:**942–946.
15. **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.
16. **van Embden, J. D. A., M. D. Cave, J. T. Crawford, J. W. Dale, K. D. Eisenach, B. Gicquel, P. Hermans, C. Martin, R. McAdam, T. M. Shinnick, and P. M. Small.** 1993. Strain identification of *Mycobacterium tuberculosis* by DNA fingerprinting: recommendations for a standardized methodology. J. Clin. Microbiol. **31:**406–409.
17. **van Soolingen, D., P. E. W. de Haas, P. W. M. Hermans, P. M. A. Groenen, and J. D. A. van Embden.** 1993. Comparison of various repetitive DNA elements as genetic markers for strain differentiation and epidemiology of *Mycobacterium tuberculosis*. J. Clin. Microbiol. **31:**1987–1995.