The PE-PGRS glycine-rich proteins of *Mycobacterium tuberculosis*: a new family of fibronectin-binding proteins?

Clara Espitia,[1] Juan Pedro Laclette,[1] Mariana Mondragón-Palomino,[1] Angelica Amador,[1] Jaime Campuzano,[1] Anke Martens,[2] Mahavir Singh,[2] Raul Cicero,[3] Ying Zhang[4]† and Carlos Moreno[4]‡

Author for correspondence: Clara Espitia. Tel: +52 5 6223884. Fax: +52 5 6223369.
e-mail: espitia@servidor.unam.mx

[1] Departamento de Inmunología, Instituto de Investigaciones Biomédicas, UNAM, Apartado Postal 70228, 04510 México DF, Mexico

[2] GBF, German National Research Center for Biotechnology, 38124 Braunschweig, Germany

[3] Hospital General de México, SSA, Mexico

[4] MRC Tuberculosis and Related Infections Unit, Clinical Sciences Centre, Hammersmith Hospital, Du Cane Road, London, UK

A clone was isolated by screening of a cosmid library of *Mycobacterium tuberculosis* with an oligonucleotide designed from the N-terminal sequence of a previously reported proline-rich protein. Characterization of the 4481 bp insert showed the presence of polymorphic CG-repetitive sequences (PGRSs) with an ORF of 2·7 kb, encoding a 81·3 kDa protein (PE-PGRS81). Southern blot analysis and BLAST-p searches revealed several homologous sequences in the genome of *M. tuberculosis*. The deduced amino acid sequence was highly similar to a stretch of about 98 residues in the N-terminus present in several members of the PE-PGRS family available in the GenBank database, including 100% identity with the partial amino acid sequence of the potential protein encoded by orf3′ as well as with the Rv0278c sequence. A neighbour-joining analysis of the 99 PE-PGRS sequences available in the database indicated that PE-PGRS81 is included in a group where its closest relatives are the sequences orf3′, Rv0278c, Rv0279c, Rv1759c, Rv3652 and Rv0747. Probing with the complete coding regions of PE-PGRS81 and Rv1759c in Southern blot assays, on samples of genomic DNA from *M. tuberculosis* H37Rv, *Mycobacterium bovis* BCG and *M. tuberculosis* clinical isolates, showed a complex hybridization pattern for all strains. This shows the existence of intrastrain PGRS variability as reported for other PGRS members. In contrast, probing with the short conserved N-terminal region of Rv1759c reduced the hybridization to a single band. This marker allowed identification of *M. tuberculosis* clinical strains that lack Rv1759c. A recombinant C-terminal fragment of Rv1759c showed fibronectin-binding properties and was recognized by sera from patients infected with *M. tuberculosis*, suggesting that at least this member of the PE-PGRS is expressed in tuberculosis infection.

Keywords: *Mycobacterium tuberculosis*, PGRS family genes

INTRODUCTION

The knowledge of the biology of *M. tuberculosis*, the causative agent of human tuberculosis, is growing rapidly and many genes have been isolated and characterized (Andersen & Brennan, 1994). Sequencing of the complete genome of *M. tuberculosis* has been recently accomplished (Cole *et al.*, 1998).

Poulet & Cole (1995) described a polymorphic GC-repetitive sequence (PGRS) in the *M. tuberculosis* genome. This sequence, called orf3′, was cloned and sequenced and showed homology with a group of sequences characterized by the presence of short ORFs. With the *M. tuberculosis* genome sequence now available, two large unrelated families of putative proteins,

†**Present address:** Dept of Molecular Microbiology and Immunology, Johns Hopkins University, School of Hygiene and Public Health, 615 N Wolfe Street, Baltimore, MD 21205, USA.

‡**Present address:** Dept of Immunology, King's College School of Medicine and Dentistry, Bessemer Road, London SE5 9PJ, UK.

Abbreviations: Fn, fibronectin; PGRS, polymorphic GC-repetitive sequence.

The GenBank accession number for the sequence reported in this paper is AF071081.

PE and PPE, have been identified. These putative proteins have conserved N-terminal sequences, with PE and PPE motifs, respectively, present in the majority of them (Cole *et al*., 1998).

PGRSs have been used as a genetic markers of polymorphism for epidemiological studies on *M. tuberculosis* and *M. bovis* (van Soolingen *et al*., 1993; Romano *et al*., 1996; Chaves *et al*., 1996; Torrea *et al*., 1996; Skuce *et al*., 1996; Cousins *et al*., 1998; Strassle *et al*., 1997), the highly repeated DNA element (pTBN12) of *M. tuberculosis* (Ross *et al*., 1992) and the PGRS DNA from *M. bovis* (pMBA2) (Bigi *et al*., 1995) being the most frequently used. The knowledge of the PGRS polymorphism could prove useful in at least three aspects of mycobacterial biology: (1) sources of genetic variation in *M. tuberculosis* and its epidemiological and pathological implications; (2) the biological role that such putative proteins might have if they are really expressed, including their pathological and immunological significance; and (3) the evolution of the genome.

A previous report has suggested that at least one member of the PGRS family is expressed in *M. tuberculosis*: one coding sequence, isolated from an *M. tuberculosis* expression library by immunoscreening with an antibody raised against the *M. bovis* 85 complex, encodes a functional fibronectin (Fn)-binding protein (Abou-Zeid *et al*., 1991). The PGRS member Rv1759c from the genome sequence project has been related to this Fn-binding protein (Cole *et al*., 1998).

During the screening of a cosmid library of *M. tuberculosis* directed to the isolation of the coding sequence of a proline-rich protein (Espitia *et al*., 1995), a clone was isolated containing an insert with an ORF encoding a glycine-rich protein of 81·3 kDa (PE-PGRS81). Here we provide evidence that Rv1759c, a close relative of PE-PGRS81, is a new member of the PE-PGRS glycine-rich protein family, is a functional Fn-binding protein and is expressed in tuberculosis infection. To our knowledge, this is the first direct evidence that a Fn-binding protein member of the PE-PGRS family is expressed in *M. tuberculosis*.

METHODS

Bacteria. *M. tuberculosis* strain H37Rv and *M. bovis* BCG Glaxo were obtained from the mycobacterial culture collection at the Tuberculosis and Related Infections Unit, Medical Research Council, Clinical Sciences Centre (London, UK). The *M. tuberculosis* clinical isolates were obtained from patients with confirmed pulmonary tuberculosis from the Hospital General in Mexico City.

Sera. Sera from patients with pulmonary tuberculosis diagnosed by smear and/or culture of sputum were obtained from the Hospital General in Mexico City. Rabbit antiserum against *M. tuberculosis* H37Rv was obtained as described before (Espitia *et al*., 1991). Human Fn from Boehringer was labelled with biotin (Boehringer) following instructions from the manufacturer.

DNA preparations. DNA from *M. tuberculosis* H37Rv and *M. bovis* BCG, and from *M. tuberculosis* clinical isolates, was isolated as described by van Embden *et al*. (1993). In brief, bacteria were killed by heating at 80 °C for 20 min, digested with lysozyme and a mixture of SDS and proteinase K, and then treated with CTAB/NaCl (10% *N*-acetyl-*N*,*N*,*N*-trimethylammonium bromide, Merck, in 7 mM NaCl). DNA was extracted with chloroform/isoamyl alcohol and precipitated with 2-propanol.

Screening of the cosmid library. A 30-mer oligonucleotide probe based on the N-terminal amino acid sequence of *M. tuberculosis* proline-rich protein (Espitia *et al*., 1995) was synthesized: 5′-CGC/G CGC/G CGG/C CAC/G CGG/C CGG/C CGC/G CGG/C CTC CGG/C-3′ The oligonucleotide was radiolabelled at its 5′ end with T4 kinase (Pharmacia) and [^{32}P]ATP (Amersham), and used for screening a cosmid library of *M. tuberculosis* H37Rv in Tropist3 (De Smet *et al*., 1993), by colony plaque blotting on nitrocellulose filters. Filters were prehybridized and probed at 42 °C in 6 × SSC (900 mM NaCl, 90 mM sodium citrate; pH 7·0), 1 mM sodium phosphate, 1 mM EDTA, 0·05% skimmed milk, 0·5% SDS, for 2 and 4 h, respectively. Afterwards, filters were washed twice in 2 × SSC for 15 min each, and once in 2 × SSC, 0·3% SDS for 15 min, and autoradiographed by exposure to an X-ray film (Kodak). Ten positive colonies were picked and grown in overnight cultures in LB-kanamycin (25 μg ml^{-1}). Cosmid DNA was isolated using the Qiagen Plasmid Isolation kit following the manufacturer's instructions.

Southern blot assays. DNA samples from the positive cosmid colonies were initially screened by digestion with *Eco*RI, separated on 0·8% agarose gels and transferred by blotting onto nylon filters (Amersham). The membrane was probed with the labelled oligonucleotide and autoradiographed as described above. From the three cosmid samples that were positive and contained similar inserts, one (randomly chosen) was digested with other restriction enzymes and probed again. A 3·9 kb *Sal*I DNA fragment was purified by excision of the band from the gel using the Stratagene Gene-Clean kit and subcloned for expansion into pUC19; the resulting plasmid was designated pUC19S/3.9. Since the 3·9 kb *Sal*I insert did not contain the complete gene of PE-PGRS81, a 0·5 kb *Kpn*I–*Sal*I DNA fragment from the cosmid was subcloned in pUC18. The 3·9 kb and 0·5 kb inserts were sequenced in both directions using an ABI 373 DNA sequencer and Prism dideoxy cycle sequencing kit (Perkin Elmer). The sequence data were analysed using the GeneWorks program (IntelliGenetics).

RFLP analysis. Southern blot assays were also carried out using restriction-enzyme-digested genomic DNA samples from *M. tuberculosis* H37Rv and *M. bovis* BCG, and also from clinical isolates of *M. tuberculosis*. The DNA probes were labelled with the Amersham direct nucleic acid labelling and detection system. The following probes were used. (1) A 3·9 kb *Sal*I–*Sal*I fragment from pUC19S/3.9, described above, including derived fragments (Fig. 1): a 693 bp *Kpn*I–*Bam*HI fragment, which contains 147 bp of a neighbouring PGRS encoding the C-terminus of Rv0279c, 249 bp of noncoding sequence upstream of the start codon, plus 297 bp of coding sequence for PE-PGRS81; and three consecutive fragments of 339 bp (*Bam*HI–*Bam*HI), 1·2 kb (*Bam*HI–*Kpn*I) and 850 bp (*Kpn*I–*Kpn*I). (2) A 4·0 kb *Sal*I–*Sal*I fragment from pUC19S/4.0 (Fig. 1); this fragment was isolated from an *M. tuberculosis*

Table 1. Features of the PE-PGRS family

PE-PGRS	Mol. mass (kDa)	I* (%)	Putative RBS	PE-PGRS	Mol. mass (kDa)	I* (%)	Putative RBS
PE-PGRS81[1]	81.3	-	CGTATTGGGAGGTGTCAGATG	Rv1430	54.5	58	GTTGGTCGGAAGGTCGGTATG
orf 3'[2]	9.4	100	CGTATTGGGAGGTGTCAGATG	Rv3511	59.9	58	ACGGTTTGGAGCTGGTCCGTG
Rv0278c	83.8	100	CGTATTGGGAGGTGTCAGATG	Rv1068c	39.2	57	AATCAGGGAGAGGAAACCGTG
Rv0279c	68.9	94	CGTATTGGGAGGTGTCAGATG	Rv2591	46.2	57	GGGTTGGGAGGGTGGCTGATG
Rv1759c	74.2	78	GTCGTGGGAGGTTTTCAGATG	Rv1803c[5]	54.8	57	ACGACAACGGCCCAGGAGGTG
Rv3652	4.9	76	GGTGGTCGGAGGTGTCCGATG	Rv1450c	107.9	56	CGTGGCGGGAGGTCTGTGATG
Rv0747	65.3	76	GGGGTTGGGAGGCATGCGATG	Rv2490c	132.9	56	CTATTGGGTAGGTGCGAGATG
Rv0742	15.5	76	CCTCTTGGGAGGTGTCACATA	Rv2340c	51.2	56	ACGGGCAGGTGCTCGCTTATG
Rv1091	66.8	72	ATTGATCGGAGGTAAACGATG	Rv1441c	40.7	56	AGGTTTTAGACTGCAGCGATG
Rv0124	41	71	GGGTTGGGGAGGTGTGTGATG	Rv2853	51.2	56	GGGAGGTAACGATGTTGTATG
Rv2741	44.1	70	CGTGTTTGGAGGTCTCAGATG	Rv1172c	30.9	56	ACCAGGAGAAGGTACGAGATG
Rv1087	62.4	69	ACTGATCGGAGTAAGGCGATG	Rv0578c	105.8	55	GGTATCGGGAGGTGCGAGGTG
Rv3595c	37.1	69	GCTGTTCGGAGGTGGCCAATG	Rv2615c	39.2	55	GCGGTCAGGAGGATTTCGATG
Rv1325c	49.5	68	AGCTGATCGGAGGAAAGCATG	Rv3097c	44.8	55	GAGAAATAGGGACACGTAATG
Rv3345c	91.9	67	GCCATTGGTGTTGGAGGAATG	Rv3650	9.5	55	CCGGGCGGGAGGTGACACATG
Rv3388[3]	60.3	67	ATTGAAGGGAGAACAGCCATG	Rv0978c	30.9	55	GCGGTGAGGAGGATTTCGATG
Rv2396	30.6	67	CGTGGCGGGTGGACTGTGATG	Rv1452c	61	54	CGTGGCGGGTGGACTGTGATG
Rv2162c	44.4	66	GCTGGTCGGAGGTGCGGGATG	Rv0980c	42.2	54	GCGGTCAGGAGGATTTCGATG
Rv1768	51.9	66	TGAGGCGGGAGGTCCCTGATG	Rv2371	5.9	51	AGCGTTCAGGAGGTCTCGATG
Rv1468c	32.1	66	TCATTGCGGAGGTGCCAGATG	Rv0754	56.8	51	AACACCCGGGGGTAAGCGATG
Rv1840c	43.8	66	TCCCGTGAGGAGCGCGTCATG	Rv1651c	88.3	50	ACCACATGTCCTTTGTGAATG
Rv0297[4]	49	65	AGTCGACGGGAGGTTCCCATG	Rv0977	81.5	50	CCCCAGGAGGTCAGCGCCATG
Rv1396c	47.8	64	GGATTCAGGGGGACGGTCATG	Rv1788	9.5	50	AACTAAGCAGGAGATCGCATG
Rv0746	65.3	64	TCGGGTGGGAGGTGGCACATG	Rv1040c	26.1	45	ACACACAAGGAGCAATGGATG
Rv0834c	71.6	64	CCAGCTGGGAGGTGTCGGATG	RV3812	51.7	49	CCAGTCGAGAGGAACAACGTG
Rv1818c	40.7	63	GTGTTGACGAGGTGCCAGATG	Rv0160c	52.9	49	GTTGGGAAGGACGGGGAGATG
Rv2487c	57.1	62	GGTATCGGGAGGTGCGAGGTG	Rv1791	9.5	49	TTGGGAGAGGAAGACAGCATG
Rv0109	12.2	62	AAACGGCAGAGGTCCCTGATG	Rv1195	9.6	48	GTTTGAGGAGGAGTGCACATG
Rv3590c	34.4	61	CGCGGAAAGGGGTGTCAGATG	Rv0159c	46.8	48	CGTCGAACAGGAGACCAGATG
Rv3508	147.4	61	CCGGTTGGGAGTTCGCCCATG	Rv2519	48.7	48	GCGGGGAAGAGGTTAGTTGTG
Rv3514	115.6	61	GGTTGCACCGGCGCAAAGATG	Rv2328	36.6	47	GAGTCGAGAGGACCGAGTATG
Rv1243c	47.2	60	CCGCGAGAGAGGCCCCGCGAT	Rv1646	30.1	46	TGTACAAGGAGTCGGGCTATG
Rv0532	50.7	60	GTTGCGGGAAGGTGTGCGATG	Rv1088	15.3	45	GTGGGTAAAGGATTGCGGATG
Rv3367	49.6	60	GCCTGGGGGTGGGTTGTGATG	Rv3477	9.7	43	AAATGAGGAGGAGCACGCATG
Rv2634c	63	59	CGATGATCCGCGGCGACCATG	Rv3622c	9.5	43	TAGAGGGAGGAAAATACCATG
Rv1214c	9.9	59	GCAGCCCCAACCACCGCGATG	Rv2107	10.3	38	GGAGCGGCGTGGTGTCTTATG
Rv1983	53.6	59	TCGAGGCGGAGGGGCGGCATG	Rv0916c	9.8	37	CGCGTCAGGTGGTATCCGATG
Rv0832	13	59	TGTGGGAGGAGATGTCGCGTG	Rv2099c	5.5	37	GTTGGTGGGAGGTGCGCAATG
Rv3507	110.5	59	GCTGGTGCGAGGTACCGGATG	Rv2431c	10.6	34	TCGCGGAGTGGAGTCGGCATG
Rv0872c	50.2	59	CCCTGCGGGAGGACAGCGATG	Rv1806	9.8	31	TCTGTCAGGAGAAGACCCATG
Rv1067c	54.6	58	AATCAGGGAGAGGAAACCGTG	Rv2769c	26.3	30	ACCAGGAGAAGGTACGAGATG

*Identity calculated on the 98 amino acids at the N-terminus of PE-PGRS.

[1] This study (AF071081).

[2] Accession: S76843.

[3] Homologous to pTBN12 (accession: M95490).

[4] Homologous to pMBA2 (accession: Z34263).

[5] Homologous to MBHRD (accession: x70687).

λEMBL3 library using the coding sequence for a Fn-binding protein (TB1) as probe (Abou-Zeid *et al.*, 1991). A derived fragment of this was also used as a probe, the 647 bp *Pvu*II–*Bam*HI fragment containing 219 bp of noncoding region upstream of the start codon plus 428 bp starting at the putative N-terminus of Rv1759c (see below). (3) A 1·27 kb *Sph*I fragment containing the *pstS-1* gene cloned from an *M. tuberculosis* pYUB328 cosmid library was used as a control.

Recombinant expression and purification of a peptide corresponding to the C-terminus of Rv1759c. Restriction analysis of the 4·0 kb DNA fragment described above indicated that it contained the putative Rv1759c protein (not shown). In contrast, restriction analysis of the fragment containing the partial coding sequence for the TB1 Fn-binding protein (Abou-Zeid *et al.*, 1991) shows several differences from Rv1759c. Therefore, we used a 1·5 kb *Sma*I–*Sal*I fragment from pUC19S/4.0 to be fused downstream of the PQE32 plasmid vector sequence (Qiagen) encoding 481 amino acids, including hexahistidine. After ligation and transformation of XL-1 Blue cells by electroporation, recombinant expression was induced as recommended by the plasmid manufacturer. The recombinant peptide (Rv1759c-C) was purified by metal affinity chromatography using the Qiagen express system (QIAexpress) following the instructions of the manufacturer.

SDS-PAGE and immunoblotting. Rv1759c-C recombinant protein (10 μg) was run in preparative 12 % SDS-PAGE and transferred to Immobilon-P membranes (Millipore). Each membrane was cut into strips and incubated for 1 h with sera from tuberculosis patients diluted 1/50 or with biotin-labelled human Fn diluted 1/50 in PBS containing 3 % BSA and 0·3 % Tween 20. After washing, strips were incubated for 30 min at room temperature with protein A/peroxidase diluted 1/2000 (Sigma) and with streptavidin/peroxidase (Zymed) diluted 1/2000 for detection of biotin-labelled Fn. Peroxidase activity was revealed with 3,3′-diaminobenzidine and hydrogen peroxide in PBS.

Phylogenetic analysis. Alignment of all distinct PE-PGRS family sequences available in the GenBank database was initially performed with the DNAMAN program (version 2.6, Lynnon BioSoft 1994–97, Montreal, Canada), and then visually adjusted. The alignment is available from the corresponding author. Phylogenetic analysis of the 82 sequences (see Table 1), was performed using the N-terminal 98 residues that could be reliably aligned. Distances were evaluated using the method of Poisson correction, where gaps and missing information data sites were removed only in pairwise comparisons. The tree was generated by neighbour-joining using MEGA (Kumar *et al.*, 1993). All *M. tuberculosis* sequence data used in this work were downloaded from the NCBI database (http://www3.ncbi.nlm.nih.gov/Entrez/Genome/).

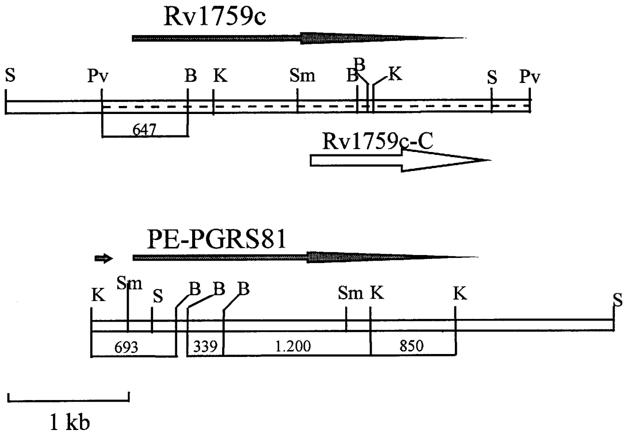

Fig. 1. Restriction maps of Rv1759c and PE-PGRS81. The arrows above the maps indicate the coding regions. Numbers in boxes show the length of the restriction fragments used as probes. The broken line indicates the 3·3 *Pvu*II fragment where the 647 bp *Pvu*II–*Bam*HI probe of Rv1759c anneals. The open arrow indicates the DNA fragment that was expressed and tested with the patient sera. Abbreviations: S, *Sal*I, Pv, *Pvu*II, B, *Bam*HI, Sm, *Sma*I, K, *Kpn*I.

RESULTS

DNA sequence analysis

Two ORFs were found in the 4·4 kb insert from the cosmid clone that bound to the oligonucleotide probe based on the N-terminal amino acid sequence of this proline-rich protein (Espitia *et al.*, 1995). The first ORF encodes a putative proline-alanine-rich protein of 724 amino acids with a calculated molecular mass of 70 kDa. The second ORF, of 2718 kb, was found in the antisense direction and encodes a putative 81·3 kDa glycine-rich protein (PE-PGRS81) starting at its 5′ end. A clone was isolated containing an insert with the ORF encoding PE-PGRS81.

BLAST-p searches of the predicted amino acid sequence of PE-PGRS81 showed significant identity to the N-terminus of members of the PE protein family. Moreover, the predicted amino acid sequence showed 100 % identity to the partial protein sequence encoded by the previously reported orf3′ (Poulet & Cole, 1995). At the DNA level, there were only two differences between orf3′ and PE-PGRS81 in 647 bp overlap. The amino acid sequence of PE-PGRS81 also showed 96 % identity with the 83·8 kDa amino acid sequence deduced from the ORF in Rv0278c. A total of 57 differences were found between the coding sequences of PE-PGRS81 and Rv0278c, including the insertion of 12 amino acids (NGGAGGNGGAGG) in position 464 of Rv0278c, encoded by 36 bp in two identical repeats of 18 bp (5′-CAACGGCGGCGCCGGCGG-3′). This insertion contains the triplet consensus sequences present in PGRSs described elsewhere, CGGCGGCAA (Ross *et al.*, 1992; Doran *et al.*, 1993; Poulet & Cole, 1995).

The putative site for ribosome binding, GGAGG, described by Poulet & Cole (1995), is present in 38 of the 82 PE glycine-rich protein sequences; 40 show similar sequences and 4 do not have the consensus sequence (Table 1), raising the question whether these proteins are expressed in mycobacteria. The same consensus sequence has been reported for protein genes of *M. tuberculosis*, *M. leprae* and *M. paratuberculosis* (Dale & Patki, 1990; Bannantine *et al.*, 1997), as well as in some *Streptomyces* protein-coding genes (Strohl *et al.*, 1992).

RFLP analysis

RFLP assays were done on membranes blotted with *Pvu*II-restricted genomic DNA of *M. tuberculosis*, *M. bovis* BCG and *M. tuberculosis* clinical isolates, using several PGRS restriction fragments as a probes (Fig. 1). A polymorphic pattern of about 20 positive bands was observed in all cases when with full-length PE-PGRS81 was used as probe (Fig. 2a). An identical pattern was also observed when blots were probed with a variety of fragments from PE-PGRS81 or with the full-length Rv1759c (not shown). In contrast, only one hybridization band of about 3·3 kb was detected in five out of seven clinical isolates with the 647 bp *Pvu*II–*Bam*HI probe, corresponding to the conserved N-terminal region of Rv1759c (Fig. 2b). Since Rv1759c contains a fragment of the same size flanked by *Pvu*II sites (Fig. 1), the 647 bp probe is specifically annealing to the Rv1759c gene. These results indicate that the observed polymorphism is due to the GC-rich region that contains the consensus repeats, and suggest that the same basic sequence is responsible for the polymorphism in all PE-PGRSs. Hybridization of the specific *M. tuberculosis* complex probe with the clinical isolates confirmed that all of them were *M. tuberculosis* strains (Fig. 2c).

Although the 693 bp *Kpn*I–*Bam*HI probe was also derived from the N-terminus of PE-PGRS81 (Fig. 1), it

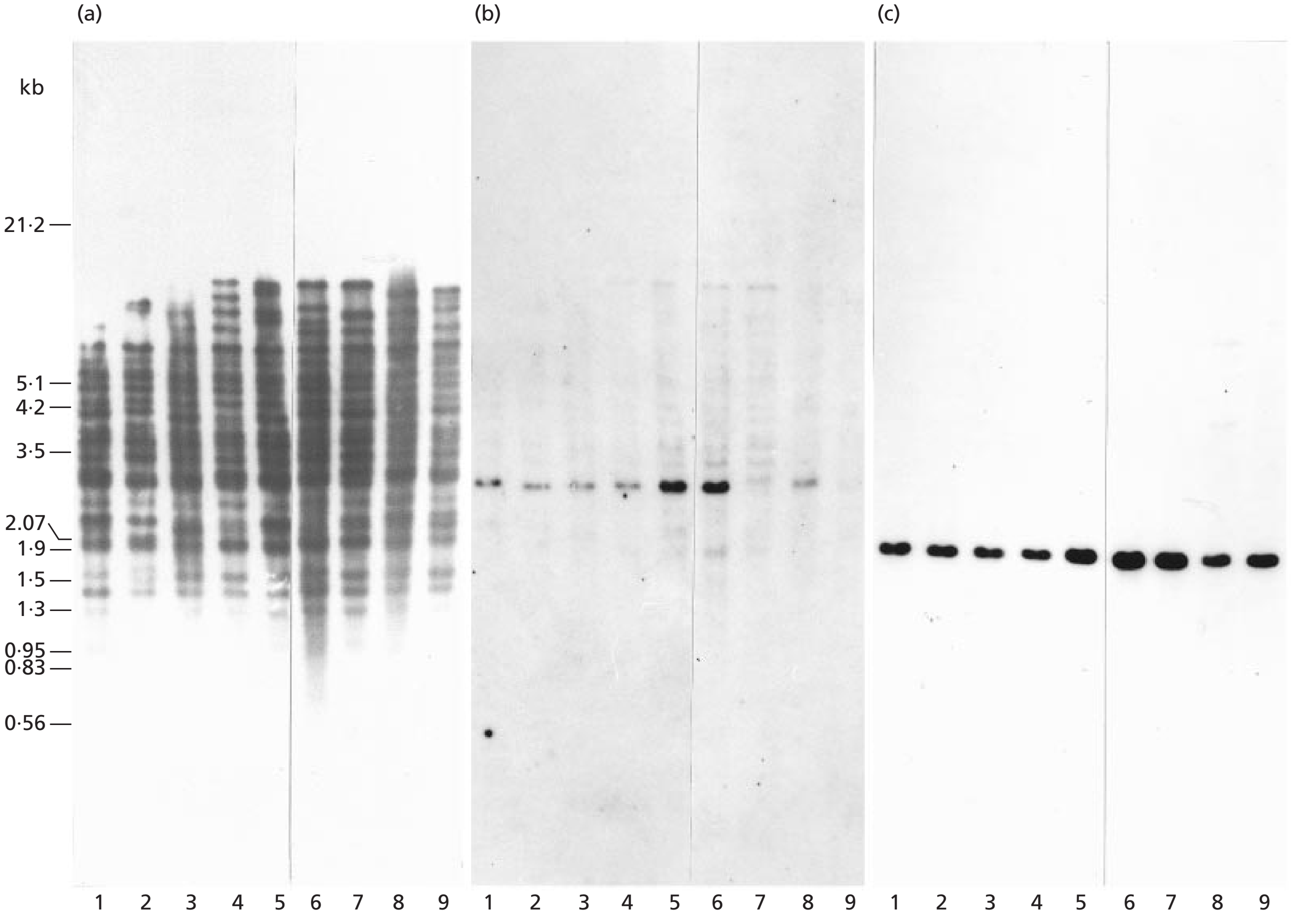

Fig. 2. RFLP analysis of genomic mycobacterial DNA: *M. tuberculosis* H37Rv (lane 1), *M. bovis* BCG (lane 2) and seven *M. tuberculosis* clinical isolates (lanes 3–9). *Pvu*II-digested DNA from each strain was hybridized with: (a) the insert of pUC19S/3.9 (which includes the coding sequence for PGRS81); (b) the 647 bp *Pvu*II–*Bam*HI fragment in Rv1759; and (c) the 1272 bp *Sph*I fragment containing the *pstS-1* gene. Fragment sizes were determined using DNA markers.

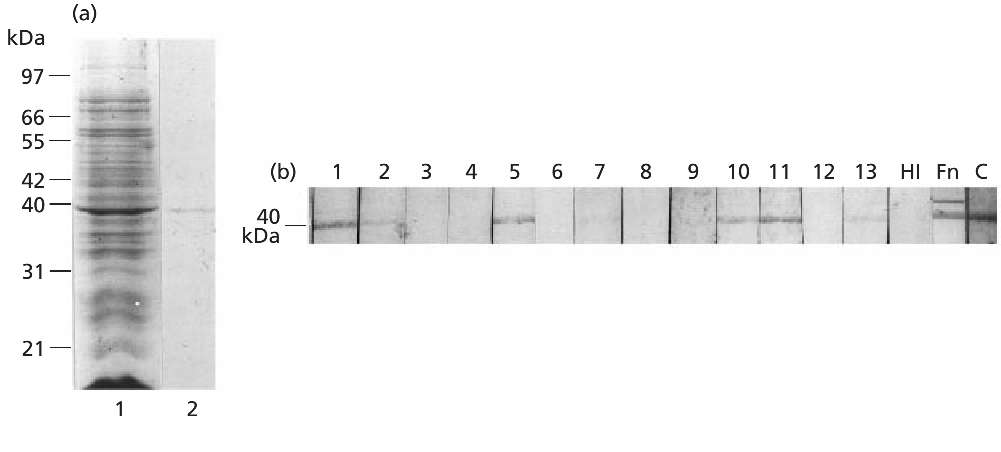

Fig. 3. Recombinant expression and purification of Rv1759c-C. (a) Coomassie-blue-stained 12 % SDS-PAGE gel with the transformed bacterial lysate (lane 1) and affinity-purified Rv1759c-C (lane 2). (b) Western blot antibody recognition of the *M. tuberculosis* recombinant Rv1759c-C by sera from individuals with pulmonary tuberculosis (lanes 1–13). Lanes HI and Fn were incubated with rabbit hyperimmune anti-*M. tuberculosis* serum (HI) or with biotin-labelled Fn, respectively. Lane C shows the Coomassie blue staining of the purified Rv1759c-C transferred to the nylon membrane.

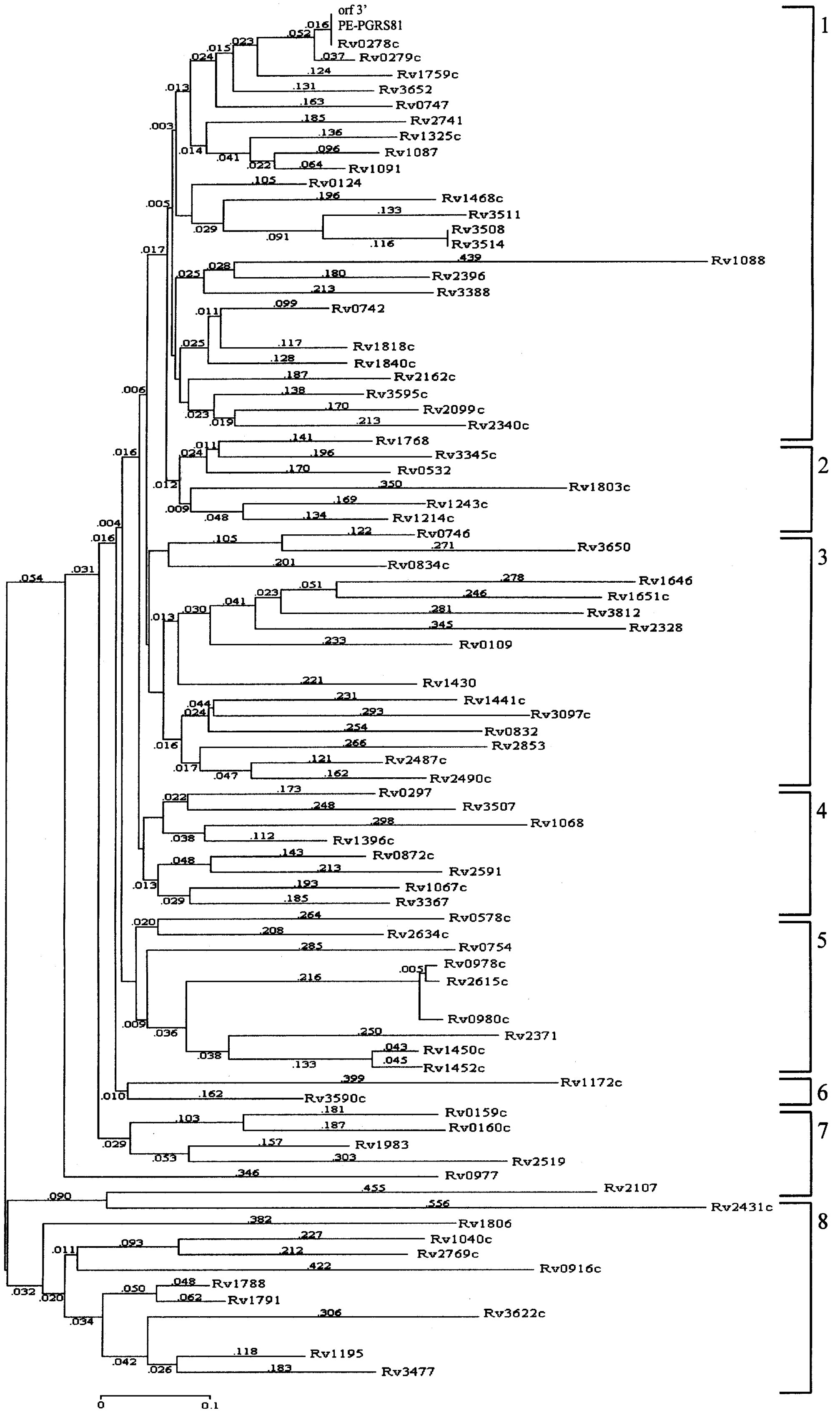

Fig. 4. For legend see facing page.

showed the same polymorphic pattern as found with probes outside the conserved region. DNA comparison of the 693 bp *Kpn*I–*Bam*HI from PE-PGRS81 and the 647 bp *Pvu*II–*Bam*HI from Rv1759c showed important differences in the number of CGGCGGCAA repeats: abundant in PE-PGRS81 and scarce in the coding region of Rv1759c (CGGCGGCGG, CGGCGGCAA 2×, CGGCGG and CGGCAACGG). This could explain the moderate background observed with the 647 bp probe (Fig. 2b).

Expression of the C-terminus of Rv1759c

The 1·5 kb (*Sma*I–*Sal*I) fragment from pUC19S/4.0 used for expression encodes the putative C-terminus of a Fn-binding protein. The recombinant product was expressed as a fusion protein with polyhistidine purification tags. After purification, the recombinant fragment of about 38·5 kDa (Fig. 3a) was recognized in Western blotting assays by antibodies present in 6 out of 12 sera from tuberculosis patients (Fig. 3b). In contrast, it was not recognized by a rabbit hyperimmune serum raised against a crude extract of *M. tuberculosis* H37Rv. The recombinant fusion product on the membrane also bound biotin-labelled Fn (Fig. 3b).

Analysis of PGRS sequences

All 99 sequences referred to as PE in the *M. tuberculosis* chromosomal linear map (Cole *et al.*, 1998) were individually downloaded from the GenBank database. From these, 19 sequences were excluded because initial alignments of the deduced amino acid sequences showed that 12 sequences could not be reliably aligned (Rv0151c, Rv0152c, Rv0285, Rv0335, Rv1386, Rv2408, Rv2769, Rv3020c, Rv3539, Rv3746, Rv3872 and Rv3893c) and the other 7 did not have the conserved N-terminus. (Rv0833, Rv1089, Rv2098c, Rv2126c, Rv3344, Rv3512 and Rv3652). Orf3′ and our PE-PGRS81 sequence were also included in the phylogenetic analysis, making a total of 82 sequences. Two distinct regions were identified in all 82 putative true PE-PGRS family members: the highly conserved N-terminal region of about 100 residues and a high (20–36%) content of alanine, and the more variable contiguous glycine-rich region characterized by pentapeptide GGXGG motifs and glycine content of 35–43%.

Phylogenetic analysis of the 82 sequences, performed on a stretch of 98 amino acids at the N-terminus, that could be reliably aligned, resulted in a tree that suggests a complex evolutionary history for the PGRS family (Fig. 4). Eight groups were arbitrarily defined on the basis of sister branching in the tree. The potential protein PGRS81 fell within group 1 in a small subgroup including orf3′, Rv0278c, Rv0279c, Rv1759c, Rv3652 and Rv0747.

DISCUSSION

The rapid progress in the characterization of the *M. tuberculosis* genome is resulting in the identification of related sequences that can be organized in gene and protein families. Here, we have described a new potential coding sequence (PE-PGRS81) related to the PE-PGRS glycine-rich proteins of *M. tuberculosis*. BLAST-p searches of the amino acid sequence of PE-PGRS81 as well as phylogenetic analysis of the PE family indicate that its closest relatives are a group of at least six sequences, including orf3′, Rv0278c, Rv0279c, Rv1759c, Rv3652 and Rv0747.

PGRSs have been widely used as markers of polymorphism in epidemiological studies of the *M. tuberculosis* complex. Comparison of previously described PGRSs with the sequences in the database showed that they share a very high identity with members of the PE-PGRS family. They only differ from those in the database by deletions/insertions and single nucleotide changes. For example, pTBN12, with an insert of 1111 bp, isolated from genomic DNA of *M. tuberculosis* H37Rv, showed 93% identity (14 single changes and one insertion of 64 bp in Rv3388) with the C-terminus of Rv3388; pMBA2, a 746 bp fragment, cloned from an Argentinian isolate of *M. bovis*, has 94% identity (8 single nucleotide changes and a deletion of about 37 bp in Rv0297) with the C-terminus of Rv0297; and MBHR, a 2365 bp fragment, cloned from a genomic library of *M. bovis* AN5, has 94% identity (14 single nucleotide changes and an insertion of 164 bp in Rv1803c) with Rv1803c. Finally, orf3′, a 1435 bp fragment isolated from a cosmid library of *M. tuberculosis* H37Rv DNA, shows 92% identity with Rv0278c and Rv0279c (which are contiguous sequences). The major changes in these sequences are 15 single changes and a 49 bp insertion following by a 45 bp deletion near to the 3′ end of Rv0279c.

From these observations, it is clear that PE-PGRS81, orf3′ and Rv0278c on the one hand, and pTBN12 and Rv3388 on the other, all originating from *M. tuberculosis* strain H37Rv, are highly homologous sequences. The changes (mutations, deletions/insertions) in these sequences could be the result of intergenic or intragenic recombinational events between the repeat regions of these PGRS members (Cole *et al.*, 1998). A similar mechanism has been described for protein M in group A of *Streptococcus*. This protein is a virulence factor and a major surface protein which exhibits size variation in strains of the same serotype. Analysis of variants shows that insertions/deletion mutations arise in a single strain by homologous recombination events. These events can lead to the generation of antigenic variation (Hollingshead *et al.*, 1987).

The possibility that members of the PE-PGRS family are a source of antigenic variation is relevant in the context

Fig. 4. Fifty percent majority tree from a neighbour-joining analysis with a bootstrap of 100 replicates using the 98 amino acid sequence in the N-terminal end of PE glycine-rich proteins. Numbers on the right show the most obvious groupings of the sequences. The scale distance from 0 to 0·1 units at the bottom was calculated by Poisson correction.

of the immune response against mycobacteria. However, no direct evidence on the *in vivo* expression of PGRS proteins has been obtained. Abou-Zeid *et al.* (1991) cloned a partial coding sequence from an *M. tuberculosis* *λ*gt11 DNA expression library, with an antiserum raised against the antigen 85 complex of *M. bovis* BCG. Proteins from the 85 complex are mycolyltransferases with Fn-binding activity present in all mycobacteria species tested so far (Content *et al.*, 1991; Soini & Viljanen, 1997; Belisle *et al.*, 1997). Recombinant TB1 also binds Fn and is recognized by sera from tuberculosis patients. *TB1* appears to be related to Rv1759c, although our restriction enzyme mapping showed that they are distinct sequences (see Methods); the precise identification of *TB1* as a PGRS requires further characterization.

It is worthy of note that rabbit polyclonal serum raised against *M. tuberculosis* H37Rv did not react with the recombinant Rv1759c-C as the patient sera did, suggesting that the protein is expressed *in vivo* during infection, but not during the *in vitro* growth of the bacteria against which the rabbit serum was raised.

The relationship between the PGRS and the Fn-binding proteins is also manifested by the annealing of the oligonucleotide designed from a proline-rich protein with the DNA encoding PE-PGRS81. These proline-rich proteins present in *M. tuberculosis* and other mycobacteria bind Fn (Schorey *et al.*, 1996). However, no significant similarity in the amino acid sequence can be found between PGRS and the Fn-binding proteins, although they share short CG-rich stretches in their coding sequence.

There is no homology between the two families of mycobacterial Fn-binding proteins described, including different Fn-binding motifs, FEWYYQ for the 85 complex proteins (Naito *et al.*, 1998) and RWFV for the members of the highly homologous Fn attachment protein family (Zhao *et al.*, 1999). Our finding of Fn-binding activity in Rv1759c suggests that PGRS could constitute a third group of Fn-binding proteins with a distinct Fn-binding motif.

The *M. tuberculosis* Fn-binding proteins are immunodominant antigens (Huygen *et al.*, 1988; Content *et al.*, 1991; Espitia *et al.*, 1992, 1995) with the capacity to mediate attachment of whole bacteria to Fn-coated surfaces (Ratliff *et al.*, 1988). Interestingly, antigenic variation in the Fn adhesin has been described in *Streptococcus pyogenes* (Talay *et al.*, 1994). Taken together, these observations indicate that putative Fn-binding PGRS proteins could have an important role in host–bacteria interaction.

RFLP analysis using PE-PGRS81 and Rv1759c as probes showed identical polymorphic patterns (number, position and intensity of hybridization bands) for some *M. tuberculosis* clinical isolates. In contrast, the conserved N-terminus probe derived from Rv1759c identified only one band, indicating that the region of the PGRS which contains the consensus triplet repeats is responsible for the polymorphism, in agreement with previous reports (Poulet & Cole, 1995; Cole *et al.*, 1998). Tandem repeats of CGGCGG, CGGCAA or combinations of both arrangements were found along the coding sequences of PE-PGRS81 and its closest relatives. The polymorphic pattern observed within the 693 bp *Kpn*I–*Bam*HI probe coincided with the stretch of 164 bp which is 100% identical to the C-terminus of Rv0279c. Interestingly the deleted region of PE-PGRS81 contains two identical 18 bp repeats, with these repeats indicating that the variation of PGRS could be given by different mechanisms, one involving the presence or not of specific sequences and the other involving changes due to insertions/deletions in the repetitive regions. In terms of the evolutionary relationships among PGRSs it would be interesting to analyse the rate of changes in clinical isolates of *M. tuberculosis* and its relevance to pathogenicity.

ACKNOWLEDGEMENTS

The Tropist 3 cosmid library of *M. tuberculosis* was kindly provided by K. De Smet (Tuberculosis and Related Infections Unit, Hammersmith Hospital, London, UK). The pYUB328 *M. tuberculosis* cosmid library was kindly provided by W. Jacobs (Albert Einstein College). The plasmid pUC19S/4.0 was kindly provided by C. Abou-Zeid (Imperial College School of Medicine at St Mary's, London, UK). We thank F. Salinas for *M. tuberculosis* clinical isolates and R. Cervantes for technical assistance This research was partially supported by grant no. IN207195 DGAPA-UNAM (C.E. and J.P.L.).

REFERENCES

Abou-Zeid, C., Garbe, T., Latigra, R., Wiker, H. G., Harboe, M., Rook, G. A. W. & Young, D. (1991). Genetic and immunological analysis of *Mycobacterium tuberculosis* fibronectin-binding proteins. *Infect Immun* **59**, 2712–2718.

Andersen, A. B. & Brennan, P. (1994). Protein and antigens of *Mycobacterium tuberculosis*. IV. Physiology of *Mycobacterium tuberculosis*. In *Tuberculosis*: *Pathogenesis*, *Protection and Control*, pp. 307–332. Edited by B. Bloom. Washington, DC: American Society for Microbiology.

Bannantine, J. P., Barletta, R., Thooen, C. & Andrews, J. (1997). Identification of *Mycobacterium paratuberculosis* gene expression signals. *Microbiology* **143**, 921–928.

Belisle, J. T., Vissa, V. D., Sievert, T., Takayama, K., Brennan, P. J. & Besra, G. S. (1997). Role of the major antigen of *Mycobacterium tuberculosis* in cell wall biogenesis. *Science* **276**, 1420–1422.

Bigi, F., Romano, M. I., Alito, A. & Cataldi, A. (1995). Cloning of a novel polymorphic GC-rich repetitive DNA from *Mycobacterium bovis*. *Res Microbiol* **146**, 341–348.

Chaves, F., Yang, Z., el Hajj, H., Alonso, M., Burman, W. J., Eisenach, K. D., Dronda, F., Bates, J. H. & Cave, M. D. (1996). Usefulness of the secondary probe pTBN12 in DNA fingerprinting of *Mycobacterium tuberculosis*. *J Clin Microbiol* **34**, 1118–1123.

Cole, S. T., Brosch, R., Parkhill, J. & 22 other authors (1998). Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* **393**, 537–544.

Content, J., De la Cuvellerie, A., De Wit, L., Vincent-Levy-Frébault, J., Ooms, J. & De Bruyn, J. (1991). The genes coding for the antigen 85 complexes of *Mycobacterium tuberculosis* and *Mycobacterium bovis BCG* are members of the gene family: cloning, sequence determination and genomic organization of the gene coding for the antigen 85-C of *Mycobacterium tuberculosis*. *Infect Immun* **59**, 3205–3212.

Cousins, D., Williams, S., Liebana, E., Aranaz, A., Bunschoten, A., Van Embden, J. & Ellis, T. (1998). Evaluation of four DNA typing techniques in epidemiological investigations of bovine tuberculosis. *J Clin Microbiol* **36**, 168–178.

Dale, W. J. & Patki, A. (1990). Mycobacterial gene expression and regulation. In *Molecular Biology of Mycobacteria*, pp. 173–198. Edited by J. McFadden. Guildford: Surrey University Press.

De Smet, K. A., Jamil, S. & Stoker, N. G. (1993). Tropist3: a cosmid vector for simplified mapping of both G+C-rich and A+T-rich genomic DNA. *Gene* **22**, 215–219.

Doran, T. J., Hodgson, A. L. M., Davies, J. K. & Radford, A. J. (1993). Characterization of a highly repeated DNA sequence of *Mycobacterium bovis*. *FEMS Microbiol Lett* **111**, 147–152.

van Embden, J. D. A., Cave, M. D., Crawford, J. T. & 8 other authors (1993). Strain identification of *Mycobacterium tuberculosis* by DNA fingerprinting: recommendations for a standarized methodology. *J Clin Microbiol* **31**, 406–409.

Espitia, C., Cervera, I. & Mancilla, R. (1991). The antigenic structure of *Mycobacterium tuberculosis* examined by immunoblot and ELISA. Influence of the age of culture and of the obtaining method on the composition of the antigenic extracts. *Arch Invest Med* **22**, 101–107.

Espitia, C., Sciutto, E., Bottasso, O., González-Amaro, R., Hernández-Pando, R. & Mancilla, R. (1992). High antibody levels to the mycobacterial fibronectin-binding antigen of 30–31 kD in tuberculosis and lepromatous leprosy. *Clin Exp Immunol* **87**, 362–367.

Espitia, C., Espinosa, R., Saavedra, R., Mancilla, R., Romain, F., Laqueyrerie, A. & Moreno, C. (1995). Antigenic and structural similarities between *Mycobacterium tuberculosis* 50/55 kDa and *Mycobacterium bovis* BCG 45/47 kDa antigens. *Infect Immun* **63**, 580–584.

Hollingshead, S. K., Fischetti, V. A & Scott, J. R. (1987). Size variation in group A streptococcal M protein is generated by homologous recombination between intragenic repeats. *Mol Gen Genet* **207**, 196–203.

Huygen, K., Van Vooren, J. P., Turneer, M., Bosmans, R., Dierckx, P. & De Bruyn, J. (1988). Specific lymphoproliferation, gamma interferon production, and serum immunoglobulin G directed against a purified 32 kDa mycobacterial protein antigen (P32) in patients with active tuberculosis. *Scand J Immunol* **27**, 187–194.

Kumar, S., Tamura, K. & Nei, M. (1993). MEGA: Molecular Evolutionary Genetics Analysis, version 1.0. University Park, PA: Pennsylvania State University.

Naito, M., Ohara, N., Matsumoto, S. & Yamada, T. (1998). The novel fibronectin-binding motif and key residues of mycobacteria. *J Biol Chem* **273**, 2905–2909.

Poulet, S. & Cole, S. T. (1995). Characterization of the highly abundant polymorphic GC-rich-repetitive sequence (PGRS) present in *Mycobacterium tuberculosis*. *Arch Microbiol* **163**, 87–95.

Ratliff, T. L., McGarr, J. A., Abou-Zeid, C., Rook, G. A., Stanford, J. L., Aslanzadeh, J. & Brown, E. J. (1988). Attachment of mycobacteria to fibronectin-coated surfaces. *J Gen Microbiol* **134**, 1307–1313.

Romano, M. I., Alito, A., Fisanotti, J. C., Bigi, F., Kantor, I., Cicuta, M. E. & Cataldi, A. (1996). Comparison of different genetic markers for molecular epidemiology of bovine tuberculosis. *Vet Microbiol* **50**, 59–71.

Ross, B. C., Raios, K., Jackson, K. & Dwyer, B. (1992). Molecular cloning of highly repeated DNA element from *Mycobacterium tuberculosis* and its use as an epidemiological tool. *J Clin Microbiol* **30**, 942–946.

Schorey, J. S., Holsti, M. A., Ratliff, T. M., Allen, P. M. & Brown, E. J. (1996). Characterization of the fibronectin-attachment protein of *Mycobacterium avium* reveals a fibronectin-binding motif conserved among mycobacteria. *Mol Microbiol* **21**, 321–329.

Skuce, R. A., Brittain, D., Hughes, M. S. & Neill, S. D. (1996). Differentiation of *Mycobacterium bovis* isolates from animals by DNA typing. *J Clin Microbiol* **34**, 2469–2474.

Soini, H. & Viljanen, M. K. (1997). Diversity of 32-kilodalton protein gene may form a basis for species determination of potentially pathogenic mycobacterial species. *J Clin Microbiol* **35**, 769–773.

van Soolingen, D., De Hass, P. E. W., Hermmans, P. W. M., Groenen, P. M. A. & van Embden, J. D. A. (1993). Comparison of various repetitive elements as genetic markers for strain differentiation and epidemiology of *Mycobacterium tuberculosis*. *J Clin Microbiol* **31**, 1987–1995.

Strassle, A., Putnik, J., Weber, R., Fehr-Merhof, A., Wust, J. & Pfyffer, G. E. (1997). Molecular epidemiology of *Mycobacterium tuberculosis* strains isolated from patients in a human immunodeficiency virus cohort in Switzerland. *J Clin Microbiol* **35**, 374–378.

Strohl, W. R. (1992). Compilation and analysis of DNA sequences associated with apparent streptomycete promoters. *Nucleic Acids Res* **20**, 961–974.

Talay, S. R., Valentin-Weigand, P., Timmis, K. N. & Chhatwal, G. S. (1994). Domain structure and conserved epitopes of Sfb protein, the fibronectin-binding adhesin of *Streptococcus pyogenes*. *Mol Microbiol* **13**, 531–539.

Torrea, G., Offredo, C., Simonet, M., Gicquel, B., Berche, P. & Pierre-Audigier, C. (1996). Evaluation of tuberculosis transmission in a community by 1 year of systematic typing of *Mycobacterium tuberculosis* clinical isolates. *J Clin Microbiol* **34**, 1043–1049.

Zhao, W., Schorey, J. S., Groger, R., Allen, P. M., Brown, E. J. & Ratliff, T. L. (1999). Characterization of the fibronectin binding motif for a unique mycobacterial fibronectin attachment protein, FAP. *J Biol Chem* **274**, 4521–4526.

Received 21 June 1999; revised 14 September 1999; accepted 16 September 1999.