Isolation and characterization of an insertion element-like repetitive sequence specific for *Mycobacterium tuberculosis* complex

Sujatha Narayanan, R. Sahadevan and P. R. Narayanan

Tuberculosis Research Centre, Mayor V. R. Ramanathan Road, Chetput, Madras 600 031, India

We report the characterization of an insertion-like repetitive sequence containing the clone of *Mycobacterium tuberculosis*. This repetitive sequence contains seven inverted repeats. Restriction fragment length polymorphism studies using this probe have shown that it is not a highly polymorphic probe but rather shows conservative fingerprint pattern. Out of the 150 strains tested, only three showed different fingerprint patterns. It has several direct and inverted repeats. Homology studies of the putative protein coding region show that this repeat element might code for a metalloproteinase of *M. tuberculosis*. Homology studies also implicate this repeat element to be from a very essential region of the *M. tuberculosis* genome participating in recombination. This repeat has been found to be an ideal target for polymerase chain reaction to detect *M. tuberculosis*.

Disease caused by mycobacterial infection is a worldwide problem. Despite their highly pathogenic nature, progress towards an understanding of gene structure, organization and expression in mycobacteria has been slow. Over the last decade, study of the basic biology of the mycobacterial pathogen has benefited greatly from a molecular biological approach. Repeat elements have been identified in various species ranging from prokaryotes to eukaryotes. Analysis of the repetitive elements has led to the identification of putative insertion elements. Insertion elements (IS) are discrete segments of DNA which are able to transpose to numerous sites on bacterial plasmids and chromosomes, usually to give rise to their copies[1]. IS elements can also promote rearrangements of genomes or replicons.

Repeated DNA sequences have been identified in a range of mycobacterial species, including pathogens. There are several reports of repetitive sequences of *Mycobacterium tuberculosis* complex[2,3], and in a number of cases, analysis of these repeats has shown IS elements such as IS6110 or IS986 which were identified in *M. tuberculosis*[4–6]. Pathogenic strains of *M. avium* have multiple copies of an atypical IS element IS900 and IS901[7]. However, mycobacterial repeats without IS-like elements have also been reported like the RLEP elements in *M. leprae*[3,8]. Such highly repeated elements are used as templates in polymerase chain reaction for detection of mycobacteria from clinical specimens and also extensively used in molecular epidemiology.

Here we report the characterization of a novel IS-like repeat element from *M. tuberculosis* which has been found to be specific for *M. tuberculosis* with an implication of an important role in recombination events.

Materials and methods

Bacterial strains

Reference and clinical isolates of *M. tuberculosis* and reference strains of atypical mycobacteria were obtained from the Bacteriology Department of Tuberculosis Research Centre, Madras.

Bacterial growth and chromosomal DNA isolation

Mycobacterial strains were grown in 10 ml Middlebrook's 7H9 medium supplemented with 5% (w/v) albumin–dextrose complex (Difco Lab) at 37°C in a stationary state. The three-week-old culture was heated at 80°C for 20 min to kill the cells. After centrifugation the cell pellet was resuspended in 500 µl TE buffer (10 mM Tris-HCl, 1 mM EDTA pH 8.0). Lysozyme was added to a final concentration of 1 mg ml^{-1}, and the tube was incubated for one hour at 37°C, 70 µl of 10% (w/v) SDS and 6 µl ml^{-1} of a 10 mg ml^{-1} proteinase K (Boehringer Mannheim) were added, and the mixture was incubated for 10 min at 65°C. 100 µl of 5 M sodium chloride and 8 µl of 10% (w/v) *N*-cetyl-*N*,*N*,*N*-trimethyl ammonium bromide in 4.1% (w/v) NaCl solution were added. The tubes were mixed and incubated for 10 min at 65°C, and equal volumes of chloroform/isoamyl alcohol (24 : 1 v/v) was added and mixed. After centrifugation for 5 min, the supernatant was transferred to a fresh tube and 0.6 volume of isopropanol was added. The tubes were kept at –20°C for 30 min to precipitate the DNA. After centrifugation for 15 min the pellet was washed twice with 70% (v/v) alcohol and dissolved in 50 µl of TE buffer.

DNA manipulation

Restriction enzymes and other modifying enzymes were purchased from Boehringer Mannheim and New England Biolabs. All DNA manipulations were performed under standard conditions as described by Maniatis *et al.*[9].

Southern blotting and hybridization

The DNA fragments resolved by gel electrophoresis were transferred on to charged nylon membrane (Dupont, NEN Research Product) by vacuum blotting[10,11] using Trans-vac, TE 80 (Hoeffer Scientific Instruments), depurinated in 0.25 M HCl, and denatured in transfer buffer containing 0.4 M NaOH and 0.6 M NaCl and the membrane was rinsed in 2 × SSC.

The blots were pre-hybridized for 30 min at 65°C and hybridized overnight in the same hybridization buffer containing heat denatured radiolabelled probe DNA. Blots hybridized with radiolabelled probe were washed twice in 2 × SSC-0.5% SDS (w/v) for 30 min, wrapped in a cling film (INTACT, Flexo Film wraps) and exposed to X-ray film (Indu Film, Hindustan Photo Films) at −70°C for varying lengths of time in a cassette containing an intensifying screen. Autoradiograms were developed and fixed by standard procedures. The signals generated were visually analysed.

Subcloning of the TRC4 fragments

The pTRC4 clone was digested with *Eco*RI and *Pst*I enzymes. The resulting fragments, *Eco*RI–*Pst*I (EP4) and *Pst*I–*Pst*I (PP4), 1 kb and 1.1 kb respectively in size, were separated on agarose gel and purified with GENE CLEAN kit (Bio 101 Inc.). The fragments EP4 and PP4 were subcloned independently into the plasmid vector pGEM-4Z which was digested with *Eco*RI–*Pst*I and *Pst*I enzymes respectively, and was dephosphorylated with calf-intestinal alkaline phosphatase. Ligation was carried out at 15°C. The ligated DNA was used for transformation of HB101 competent cells. Ampicillin-resistant colonies on the LB-agar plates were screened by colony hybridization for the presence of DNA sequences which hybridized to ^{32}P labelled pTRC4. The recombinant clones were further confirmed by mini plasmid preparation and restriction digestion with *Eco*RI and *Hin*dIII enzymes and analysis on 0.8% (w/v) agarose gel.

Sequencing strategy

The *M. tuberculosis* fragment in pTRC4 and its subfragments in pEP4 and pPP4 were independently sequenced, directly from each side using primers for SP6 and T7 polymerase promoter sequences present in the vector pGEM4Z, using automatic sequencing (Applied Biosystems).

Nucleotide sequence accession number

The nucleotide sequence of TRC4 has been assigned GenBank Accession No. U84405.

Homology searches

Sequence data was stored, assembled and analysed using various softwares. Homology searches were performed using DNasis, BLAST, T-fasta, Prosite and several other software programs from GenBank and EMBL data bases.

The DNA sequence was analysed with the GCG program provided by the Genetics Computer Group, University of Wisconsin. The e-mail servers of NCBI running the Blast Program[12] and the FASTA servers were also used for sequence comparisons.

Results

A genomic library of *M. tuberculosis* was made in pGEM4Z from which clones were selected on the basis of their strong signals with ^{32}P labelled *M. tuberculosis* DNA. From the 10 clones named pTRC1–pTRC10, pTRC4 was found to be specific for *M. tuberculosis* complex, and did not cross react with any of the nonmycobacterial species and 17 atypical mycobacteria tested. The pTRC4 clone has a 2.1 kb mycobacterial fragment. This clone, besides its specificity for *M. tuberculosis* complex, has been found to be a repetitive element.

pTRC4 in RFLP studies

Experiments were carried out initially to find out whether the cloned fragment in pTRC4 is a repetitive element. *M. tuberculosis* genomic DNA from clinical isolates was restriction digested with combinations of *Pst*I and *Sal*I enzymes and subjected to Southern blot analysis. The nick translated pTRC4 DNA, hybridized with multiple bands in all the clinical isolates. The number and size of the bands were similar in all the strains tested. All the 150 clinical isolates used for Southern hybridization studies hybridized with the radiolabelled pTRC4 clone, and showed similar pattern, except 3 strains indicating that this repeat element is less polymorphic (Figure 1). To find out the actual number of TRC4 copies present in the *M. tuberculosis* genome, an enzyme *Bgl*II which does not have site within the clone was chosen. The radioactive ^{32}P-labelled pTRC4 hybridized with multiple DNA fragments of a *Bgl*II digest of

genomic DNA from standard strains of *M. tuberculosis* H37Rv, *M. tuberculosis* H37Ra, *M. tuberculosis* South Indian low virulent strain (SILV) and *M. bovis* BCG. Southern blot analysis revealed at least four major bands with additional minor bands. All the four standard strains tested showed an identical banding pattern (Figure 2 *a*). Similarly, except for one out of 28 clinical isolates of *M. tuberculosis* from South Indian patients, all other strains showed an identical banding pattern in Southern blot analysis of the *Bgl*II digested genomic DNA. One strain showed a shift in the size of the two lower bands (1.8 and 1.6 kb) but the number of bands were the same. Apart from these four very strong bands, more than three minor bands have also been uniformly found in all the strains tested (Figure 2 *b*).

The two subclones, pEP4 and pPP4 carrying the DNA fragment extending from *Eco*RI to the *Pst*I site and *Pst*I to the *Pst*I site which form the left and right half respectively of the original clone pTRC4 were used as probes to determine the RFLP pattern of 4 standard strains as mentioned above. pTRC4 was subcloned as pPP4 and pEP4 using the *Pst*I site in the middle of the clone. Southern blot of *M. tuberculosis* by all the 3 clones pTRC4, pEP4 and pPP4 revealed that the banding pattern was unique to each and some of the repeats were imperfect (data not shown). Out of the four bands which hybridized with pTRC4, three bands were obtained with pEP4 and two bands hybridized with pPP4. This hybridization pattern can be explained by the presence of imperfect repeats as shown in Figure 3. Only one *Bgl*II fragment, approximately 3.8 kb in size, uniformly lighted up with all the three probes. The two bands, of approximately 3.6 kb and 1.8 kb size, that hybridized with pEP4 but not with pPP4 were relatively weaker in intensity, probably due to its shorter length. One band approximately 1.6 kb in size hybridized only with pPP4 and not with pEP4.

The pattern was compared with the DNA patterns obtained with the insertion element IS6110 and direct repeat probe (DR). For this, the same blots were used for hybridization with either labelled IS6110 or labelled DR probes. The DR and IS6110 banding patterns differed greatly. Interestingly, the TRC4 fragment and DR probe identified strains of *M. tuberculosis* that did not have IS6110 copy (data not shown).

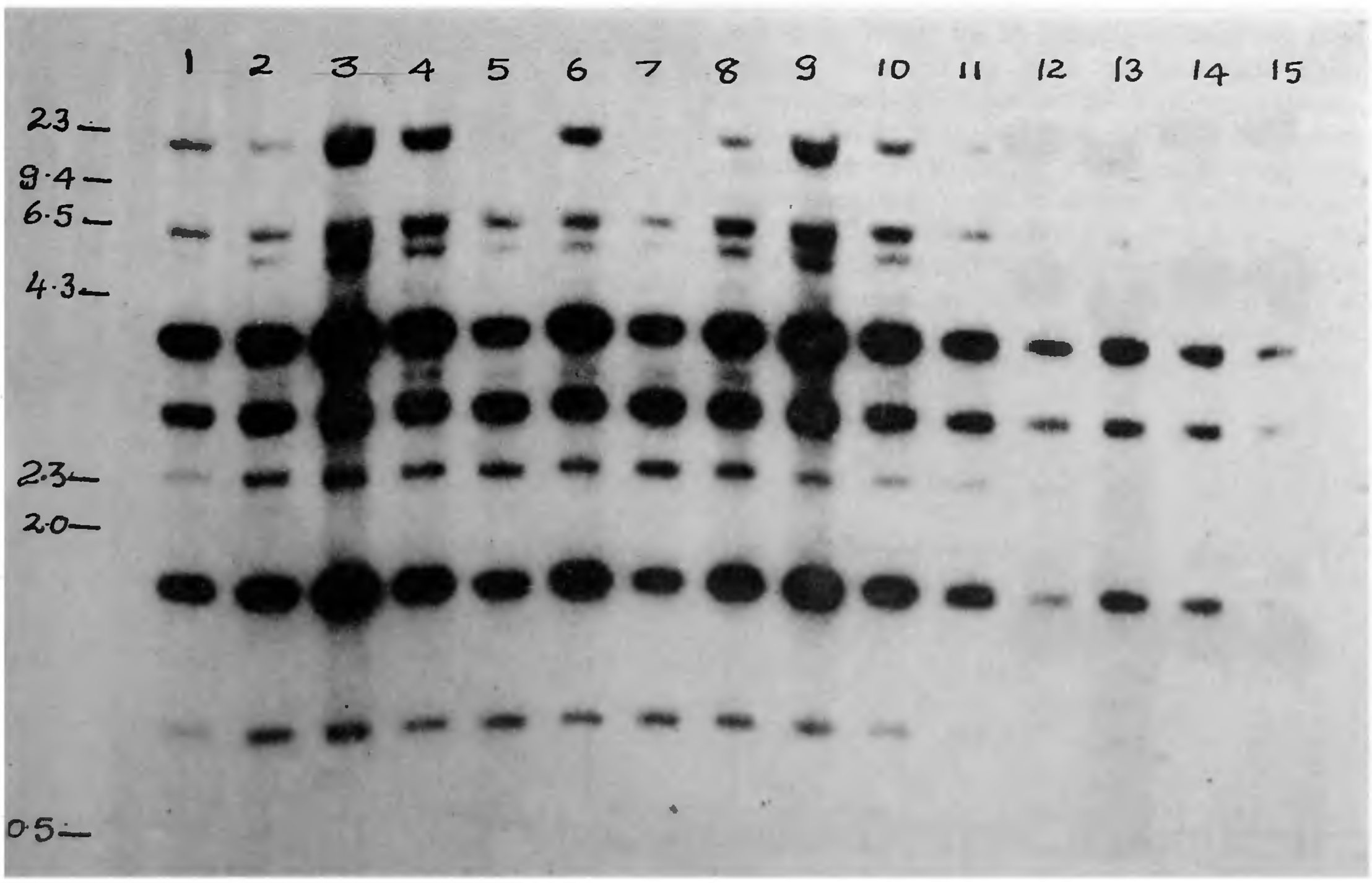

Figure 1. Southern hybridization of ^{32}P-labelled pTRC4 with *Pst*I–*Sal*I digested DNA from clinical isolates of *M. tuberculosis*. Lanes. 1–15, *M. tuberculosis* clinical isolates.

Nucleotide sequencing of TRC4

The nucleotide sequence of this 2.1 kb fragment has been deduced and the sequencing strategy has been described in the previous section. The entire nucleotide sequence of the cloned *M. tuberculosis* fragment TRC4 is shown in Figure 4. The G+C content of TRC4 is 63% which approximates that of the global G+C ratio already determined for *M. tuberculosis* genome[13].

Homology studies

A small portion of the TRC4 sequence ranging from 48 to 76 base pairs showed 68–80% homology to other sequences. The 'T-fasta' revealed that TRC4 has a borderline significant homology with the region in *M. leprae* gene cluster coding for several ribosomal proteins and subunits of RNA polymerase. The homologous region in *M. leprae* genome is about 2 kb and this region lies between an open reading frame (ORF) 220 and the rpsL gene.

The sequence was simply translated into protein sequences in all six possible frames, and homology searches were performed using 'T-fasta' in the GenBank and EMBL data bases. There were no significant homologies to any known IS or transposon sequences. The 'T-fasta'

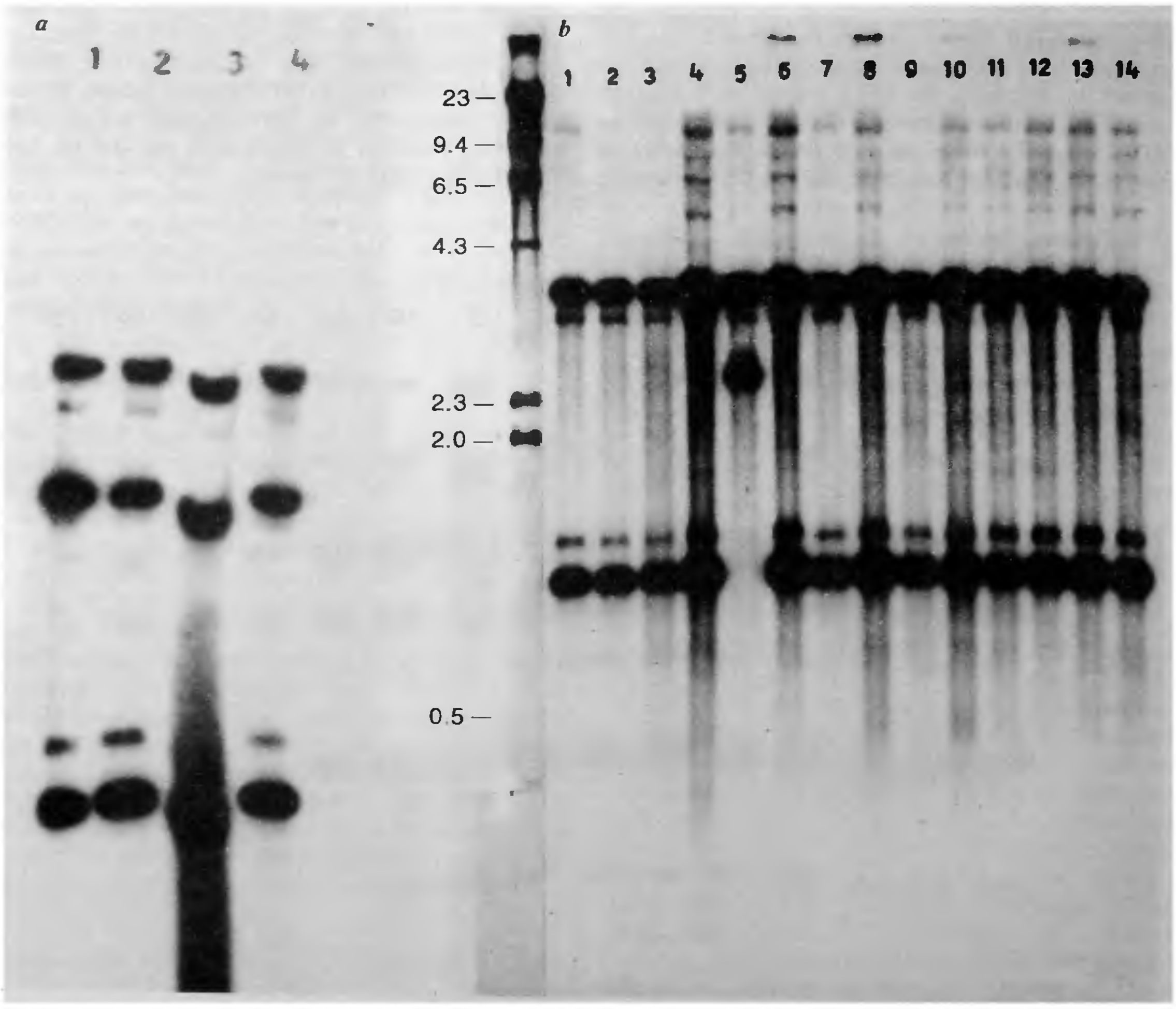

Figure 2. *a*, Southern hybridization of ^{32}P-labelled pTRC4 with *Bgl*II, digested DNA from standard strains. Lanes: 1, *M. tuberculosis* H37Rv; 2, *M. tuberculosis* SILV; 3, *M. tuberculosis* H37Ra; 4, *M. bovis* BCG. *b*, RFLP pattern of *Bgl*II digested DNA from clinical isolates of *M. tuberculosis* using pTRC4 probe. Lanes 1 to 14, *M. tuberculosis* clinical isolates. Numbers at the left indicate sizes of the standard DNA fragments in kilobase pairs.

revealed a borderline significant homology with a region of *M. leprae* gene cluster coding for several ribosomal proteins and subunits of RNA polymerase, which has the accession No. Z14314 (ref. 14). The homology lies in an approximately 2000 bp 'gap' between an ORF called ORF220 and rpsL gene. The authors did not assign an ORF or any other relevant feature to this sequence. But TRC4 has a borderline significant similarity to this region of the *M. leprae* chromosome. It is possible that related but not identical elements may be present in *M. leprae*. Nevertheless the features of the sequences show it to be a transposable element. Further work should confirm this.

TRC4 sequence position 1211 to 1251 showed less significant homology with a locus in an insertion element 3411 from *E. coli* (GenBank, No. Tn3411, 1992), which flanks the citrate utilization determinant of transposon Tn3411, and *Shigella sonnei* insertion sequence IS629. 'BLASTN' revealed few nearly-identical sequences showing homology to *M. tuberculosis* cosmid Y339 and Y210, TBC2, *Streptomyces coelicolor* gene for metalloproteinase and *lysR* type transcriptional activator, *Saccharopolispora erythraea* ferredoxin (*fdxA*) gene, 5′ end and mouse mRNA for P-cadherin. Recent searches in the GenBank revealed that TRC4 sequence has 97.5% identity in a 650 bp overlap to *M. tuberculosis* cosmid clone Y339. The region homologous might code for aminotransferase, diacyl pyrophosphatase or cytochrome P450.

The sequence of TRC4 has three major putative ORFs (Figure 4). The long ORF1 starts with an ATG codon at position 427, and ends with a translational stop codon TGA at position 799 on the normal strand and might encode a protein consisting of 124 amino acids with a molecular mass of 13.5 kDa. Upstream to this ATG codon, a probable ribosome-binding site, TGGGG extends from 413 to 417. The sequence of the putative translational product did not have homology to any other protein in the SWISS PROT Data Bank.

ORF2 extends from a GTG start codon at position 481 to TGA stop codon at position 799 (Figure 4) and its product would be a protein 106 amino acids long, with a molecular mass of 11.6 kDa. Analysis of the complementary strand revealed one long ORF3, encoding a protein of 103 amino acids from nucleotide 574C (GTG) to nucleotide 188C (TGA) with an expected molecular mass of 11 kDa. A possible purine-rich Shine-Dalgarno sequence GAAGGG from position 598 to 590 is present upstream to the GTG initiation codon.

The predicted iso-electric point of the proteins are 12.15, 12.51 and 10.96. The significance of these three ORFs remains to be elucidated. The other observed features of the TRC4 sequence are eight direct repeat sequences of 9–12 base pair long and seven palindromic sequences (Figure 4). Among the various palindromic sequences, the GC-rich palindrome which is 14 base pair long at position 936 seems to function as a rho independent terminator. The special feature of this sequence is that there are seven inverted repeats of 14 base pair long with 2 mismatches (Table 1). The shortest inverted repeat has 4 base pair target site duplication (Figure 4). Homology studies implicate a probable important role for these inverted repeats as discussed below.

Unique feature of TRC4 sequence

Subsequent sequence analysis of the cloned fragment revealed that among the 35 homologous sequences from GenBank, the trend has been that the initial region encompassing 1–500 bases and the distal region encompassing 1500–2126 bases of the TRC4 sequence show homology to several proteins mainly metalloproteinases and enzymes using metals as cofactor. The central part of the sequence does not show homology to any of the known protein except an activator/sigma factor. One striking feature is that most of the inverted repeats of TRC4 are located between 400 bp and 1500 bp. These inverted repeats seem to bracket putative promoter-like sequences.

TRC4 in diagnosis

From the deduced sequence of TRC4 which is 2.126 kb several primers were designed and optimized for

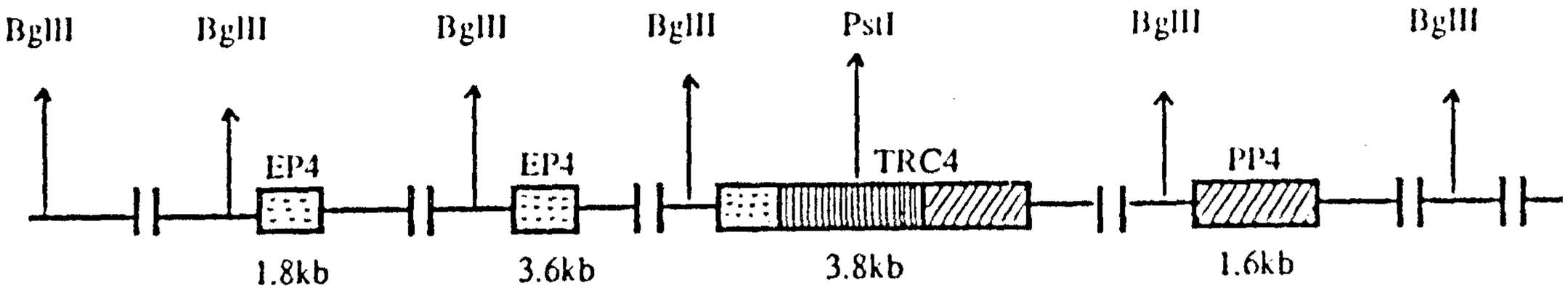

Figure 3. Hypothetical representation of TRC4 in *M. tuberculosis* genome. The four different *Bgl*II fragments, sized 3.8 kb, 3.6 kb, 1.8 kb and 1.6 kb, hybridizing with TRC4 are shown. ▥ Sequence hybridizing with 3.8 kb fragment; ▩ Sequence hybridizing with 3.6 and 1.8 kb fragments. ▨ Sequence hybridizing with 1.6 kb fragment.

NUCLEOTIDE SEQUENCE OF TRC4

```
5'-
1    GAATTCCCGA CCTCAGCGCT GATCGCCTCG CGCGGCGCCG CCTACCCCGC AACCCAACCG
     CTTAAGGGCT GGAGTCGCGA CTAGCGGAGC GCGCCGCGGC GGATGGGGCG TTGGGTTGGC

61   AACGGGCCAT CCATCCGCAC CATATTGCGG GCGTTGGCAA TTCACCGCCG CCCGTCGGCC
     TTGCCCGGTA GGTAGGCGTG GTATAACGCC CGCAACCGTT AAGTGGCGGC GGGCAGCCGG

121  AACAGTAAGC TCGGCGCGGT AACCACCGAG TTCGACACGT CCGCGCCGCG GATCCTGGAC
     TTGTCATTCG AGCCGCGCCA TTGGTGGCTC AAGCTGTGCA GGCGCGGCGC CTAGGACCTG
             DR1                          DR1
181  CAGCTGCACG CATCGGACGT CATCCGCCTG GACGCATCGG CGCGATGTAT CCTTTCTCCG
     GTCGACGTGC GTAGCCTGCA GTAGGCGGAC CTGCGTAGCC GCGCTACATA GGAAAGAGGC
                              #
241  CAGCACCGAA TCCGCATCGT ATTCAGATCG CCTATGGTGC CCAGGTTTTC GGGCTAATCG
     GTCGTGGCTT AGGCGTAGCA TAAGTCTAGC GGATACCACG GGTCCAAAAG CCCGATTAGC
                                                        DR 4
301  CAGTTTGGGT GTAGGGGGGT CGGGCGTGTT CGTTGTTGCC GGGGTCAAGG TTTTCGATGA
     GTCAAACCCA CATCCCCCCA GCCCGCACAA GCAACAACGG CCCCAGTTCC AAAAGCTACT

361  TGAAGGTTGG TTGGAACAAT CCAGCGGTGA CGCCCTTGGT GGCCGGCACT GGGTTTGGGG
     ACTTCCAACC AACCTTGTTA GGTCGCCACT GCGGGAACCA CCGGCCGTGA CCCAAACCCC
             *ORF 1->
421  TCCACCGCGA TGGGTGAGTA TGGGAAGTGT GGCACGTGTG AGCCGTCGTG GTGCACACGG
     AGGTGGCGCT ACCCACTCAT ACCCTTCACA CCGTGCACAC TCGGCAGCAC CACGTGTGCC
        *ORF 2->
481  CCAGTGGCAG CCCGTTGGCG CCGTCGCCCA ACGTGTTCTG CGGGCGGAAA ATCGGGGGGC
     GGTCACCGTC GGGCAACCGC GGCAGCGGGT TGCACAAGAC GCCCGCCTTT TAGCCCCCCG

541  TCTGATTCTC CGCCCTCAGT TCACGCTCGG TGCCGGTTAG CCTCACGCGG TCAACGTGCA
     AGACTAAGAG GCGGGAGTCA AGTGCGAGCC ACGGCCAATC GGAGTGCGCC AGTTGCACGT
                                        DR8 <-ORF 3*    DR3
601  CCCTTCGGTT ACCCCTGCGA CCCAATGACT CGGGGTCCGG GCGGGCGCTC GCGTGTCGTT
     GGGAAGCCAA TGGGGACGCT GGGTTACTGA GCCCCAGGCC CGCCCGCGAG CGCACAGCAA
                    -35
661  TCAGGACCGG TGCCGATCAG ACTCTCGAGT CTCGGTTGGT CCAGGTCACC ACGTGGAAAG
     AGTCCTGGCC ACGGCTAGTC TGAGAGCTCA GAGCCAACCA GGTCCAGTGG TGCACCTTTC
                       DR7                                  DR8
721  GATCGACGCG GCGGACCGCG TTGGGGCACC GCTGGGCCAC GACGTCGGCA ATCCGGGCGG
     CTAGCTGCGC CGCCTGGCGC AACCCCGTGG CGACCCGGTG CTGCAGCCGT TAGGCCCGCC
                              #                   -35
781  CAGCATCAGC TAAGACTACA GTGATCTGCG CTGGTCCTGT CCTTTGACAA GCCAGAAACC
     GTCGTAGTCG ATTCTGATGT CACTAGACGC GACCAGGACA GGAAACTGTT CGGTCTTTGG
                                                   -35
841  CTAAGCGACA ACGACGTGCG CCTACTCAAA CCAGAAGTCC ACCCACGGAA GTGTCAGAAG
     GATTCGCTGT TGCTGCACGC GGATGAGTTT GGTCTTCAGG TGGGTGCCTT CACAGTCTTC

901  AGCCACTTCT TCGCAGGCCG CCAAACCGCG AAGGCTGGGC CGCGGCCCAC TCCGATGTCA
     TCGGTGAAGA AGCGTCCGGC GGTTTGGCGC TTCCGACCCG GCGCCGGGTG AGGCTACAGT
                DR6
961  GCGCCATCAT CGTCGACCTG ACGACTGCAG TCCGATTGCG GAGCTACGCT AATTCGGTCC
     CGCGGTAGTA GCAGCTGGAC TGCTGACGTC AGGCTAACGC CTCGATGCGA TTAAGCCAGG
          -35                                    -10
1021 CCCTGCTCAA TGCCTGACCG AAATGCACAT TCGGCAGCGA CCGCTGTGAT CGCCCCACTC
     GGGACGAGTT ACGGACTGGC TTTACGTGTA AGCCGTCGCT GGCGACACTA GCGGGGTGAG

1081 GACGAAAGTC ACTGAACAAA TTCGCCCGCG CGGTGGCCCA GCATTACGAT CGTTGCGTGT
     CTGCTTTCAG TGACTTGTTT AAGCGGGCGC GCCACCGGGT CGTAATGCTA GCAACGCACA
                                                        DR 4
1141 GGACCGCGAC TGGTGATCGA CGGAAGGACA GATCCGTCTA TCACCCATAG GTTTTCGATG
     CCTGGCGCTG ACCACTAGCT GCCTTCCTGT CTAGGCAGAT AGTGGGTATC CAAAAGCTAC
                                                DR6                DR7
1201 CCGCGGACCC GACACCTCGG GTCGACGACG GCTCGTGGGT CATCGTCGGT GCCCATTGGG
     GGCGCCTGGG CTGTGGAGCC CAGCTGCTGC CGAGCACCCA GTAGCAGCCA CGGGTAACCC

1261 GCACTACCAC ACAGATGCTG CGATGTCGCC CATACGGCTG GACCGATGCG CGTTGCCGCA
     CGTGATGGTG TGTCTACGAC GCTACAGCGG GTATGCCGAC CTGGCTACGC GCAACGGCGT
                                                        -35
1321 CCGCATAATT CGTGGGCCAA TGCGCTACCC TGGCGCAGGC CCCCGACATC GACAGGTTCA
     GGCGTATTAA GCACCCGGTT ACGCGATGGG ACCGCGTCCG GGGGCTGTAG CTGTCCAAGT
                      DR5                -10
1381 CTGTCGTATC GGTGCTCGAT GCGGACTGGT ATCTGGGGAT CACTCGAGAC CAACGTGATG
     GACAGCATAG CCACGAGCTA CGCCTGACCA TAGACCCCTA GTGAGCTCTG GTTGCACTAC

1441 CGTCCGCGTG CCCGCGGCTG CATGAGCGCC ACCCCGATAT CCGGCCAATC ACGATACCCG
     GCAGGCGCAC GGGCGCCGAC GTACTCGCGG TGGGGCTATA GGCCGGTTAG TGCTATGGGC
                DR2                                        DR5
1501 CCTGTACGAA CGGTGGCCAG CGCGTGATGT GGACCCTAAG GGCCCGCCCG CGTGCTCGAT
     GGACATGCTT GCCACCGGTC GCGCACTACA CCTGGGATTC CCGGGCGGGC GCACGAGCTA

1561 CGGCGGCTTG CGAATTGTGC TCGCCACCCT CCGGTTGGTG GCGGTCGCGG GGTCGTTGTC
     GCCGCCGAAC GCTTAACACG AGCGGTGGGA GGCCAACCAC CGCCAGCGCC CCAGCAACAG
                                                DR3
1621 GGTGATGCCG ATTTTCGAGC CACCGCAGTC GACGCTCGCG TGGGGGCATT TTTCGAGCAT
     CCACTACGGC TAAAAGCTCG GTGGCGTCAG CTGCGAGCGC ACCCCCGTAA AAAGCTCGTA

1681 GGGCCGTCGT CACTAGGGGT CTGAATGTTG TTGGGAACCA ACCATTGGTT GTTGATTTGA
     CCCGGCAGCA GTGATCCCCA GACTTACAAC AACCCTTGGT TGGTAACCAA CAACTAAACT

1741 TCAAGGCATC CCGCTCAGCT CTCGTATGCC CGGACTCACC TGCTTCGAAA TCTTTCTGCC
     AGTTCCGTAG GGCGAGTCGA GAGCATACGG GCCTGAGTGG ACGAAGCTTT AGAAAGACGG

1801 CATCGCTGAG GCCGGCAGTC TTGGCGGCCC CGCACGCGAA CTCGGGTTGA CTCAACAAGC
     GTAGCGACTC CGGCCGTCAG AACCGCCGGG GCGTGCGCTT GAGCCCAACT GAGTTGTTCG

1861 TGTGTCAAGG CGGCTCGCAT CGATGGAGGG CCCAGATCGG GGTGCGATTG GCCATCCGGA
     ACACAGTTCC GCCGAGCGTA GCTACCTCCC GGGTCTAGCC CCACGCTAAC CGGTAGGCCT

1921 CGACACGTGG CTCCCAACTC TCTCCTACCG GCATCGTCGT CGCCGAATGG GCGGCCCGCT
     GCTGTGCACC GAGGGTTGAG AGAGGATGGC CGTAGCAGCA GCGGCTTACC CGCCGGGCGA

1981 TGCTCGAAGT CGCCGACGAG ATCGATGCCG GCCTCGGCTC GCTGCGCACC GAAGTCCGCC
     ACGAGCTTCA GCGGCTGCTC TAGCTACGGC CGGAGCCGAG CGACGCGTGG CTTCAGGCGG
                                DR2
2041 AGCGCATCAG AGTGGTGGCC AGCCAGACGA TAGCCGAACA GCTGATGCCG CATTGGATGC
     TCGCGTAGTC TCACCACCGG TCGGTCTGCT ATCGGCTTGT CGACTACGGC GTAACCTACG

2101 TGTCCTTGCG GGCCGCCGAC ATGCGC-3'
     ACAGGAACGC CCGGCGGCTG TACGCG
```

Figure 4. Nucleotide sequence of the 2.126 kb TRC4 DNA. The longest open reading frames (ORFs) are marked as ORF 1, ORF 2 and ORF 3. *Start codon (ATG & GTG), #stop codon (TGA). The potential RBS are in bold and boxed. The eight direct repeats (DR1 to DR8) are underlined. Potential promoter hexamers (– 35 and – 10 sequences) are in bold italics. Palindromic sequences are doubly underlined.

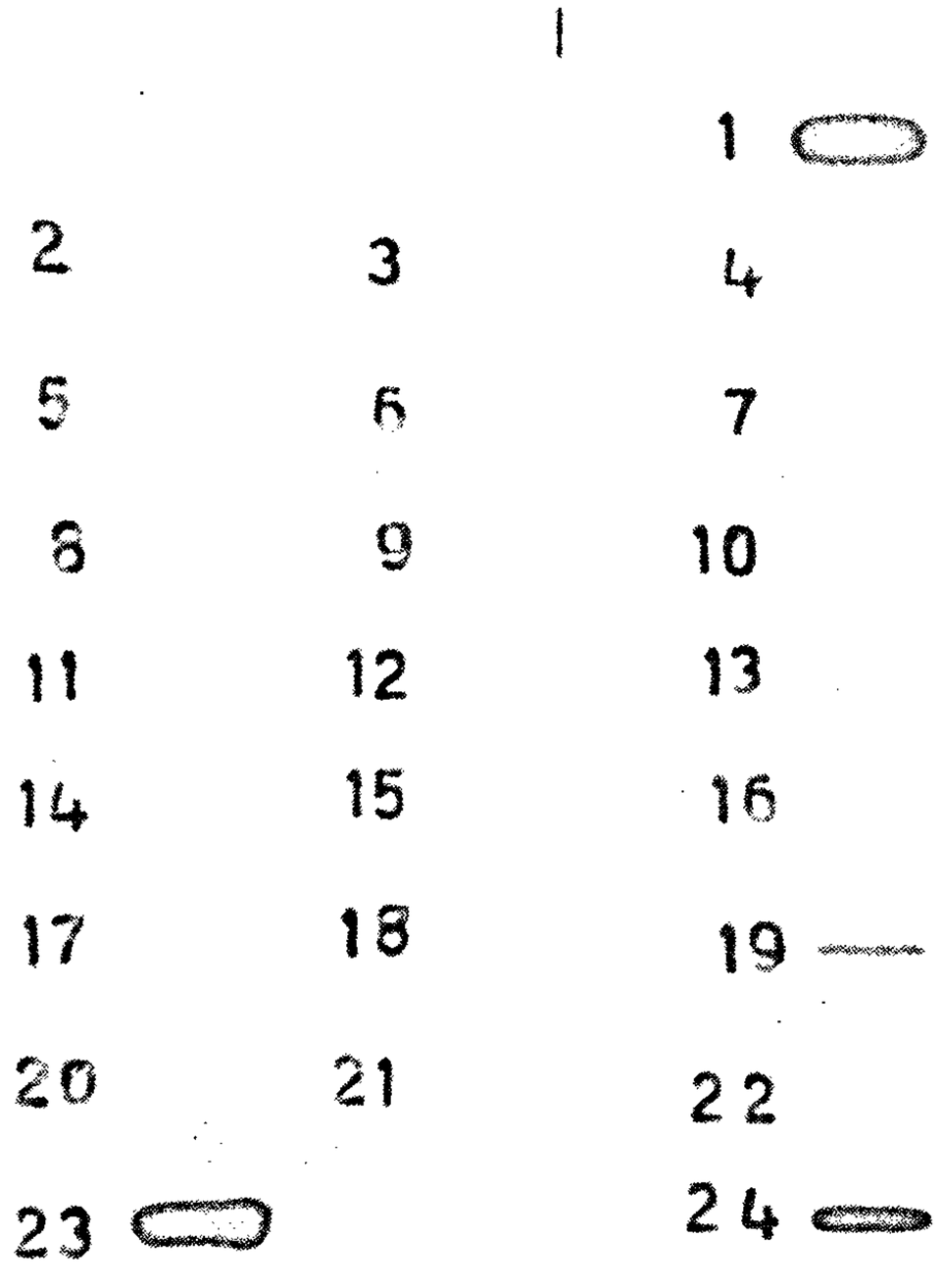

Figure 5. Slot blot hybridization of 658 bp PCR product with total DNA from various organisms. Slots: 1, *M. tuberculosis* (1 µg); 2, *M. thermoresistible*; 3, *M. gastri*; 4, *M. dieonhoferri*; 5, *M. gordonae*; 6, *M. flavescens*; 7, *M. scrofulaceum*; 8, *M. terrae*; 9, *M. fortuitum*; 10, *M. smegmatis*; 11, *M. vaccae*; 12, *M. avium-intracellulare*; 13, *M. chelonae abscesses*; 14, *M. chelonae chelone*; 15, *M. phlei*; 16, *M. simiae*; 17, *M. chitae*; 18, *M. microti*; 19, *M. bovis*; 20, *M. aurum*; 21, *E. coli*; 22, Human; 23, *M. tuberculosis* (1 µg); 24, *M. tuberculosis* (200 ng). The probe DNA was labelled with fluorescein-HRP conjugate using ECL-random-prime-labelling and detection system (Amersham International, UK).

Table 1. Inverted repeats of TRC4

No.	Sequence	Position
1	5′-TCGTGGTGCACACG-3′ 3′-TCGCGGTCCACACG-5′	466, 1136c
2	5′-GCGGGCGGAAAATC-3′ 3′-AGGGGCGGAGAATC-5′	520, 544c
3	5′-TCGCAGGCCGCCAA-3′ 3′-TCGCAAGCCGCCGA-5′	911, 1560c
4	5′-CATCGTCGACCTGA-3′ 3′-CGTCGTCGAGGCGA-5′	968, 1217c
5	5′-GGTGATCGACGGAA-3′ 3′-GGTGATAGACGGAT-5′	1149, 1182c
6	5′-CGCTGAGGCCGGCA-3′ 3′-AGCCGAGGCCGGCA-5′	1799, 2005c
7	5′-CGCGCGGCGCCGCC-3′ 3′-CGTGCGGGCCCGCC-5′	29, 1823c

to detect *M. tuberculosis* from clinical specimens. ıation is being carried out and the results will be ted shortly. One set of primers yielding a PCR product of 658 bp was radiolabelled. This was used for hybridization with the DNA from various atypical mycobacteria and *M. tuberculosis* to confirm the specificity. Figure 5 shows that this product is absolutely specific for *M. tuberculosis* complex.

Discussion

An insertion-like element TRC4 was found in *M. tuberculosis*. It is 2126 bp long with three putative ORFs, several direct repeats and seven inverted repeats of 14 bp long with two mismatches. Southern blot analysis showed that it is a repeat element with a low degree of polymorphism and is conserved among the various clinical isolates of *M. tuberculosis* and specific for *M. tuberculosis* complex. Among the four bands lighting up in the Southern blots (using *Bgl*II enzyme) two hybridization bands with weaker intensity could be due to the presence of partial copies of TRC4 sequence (Figure 3).

This repeat element is an ideal target for polymerase chain reaction to identify *M. tuberculosis* from clinical specimens including extrapulmonary tuberculosis, especially to detect strains carrying no copy of IS6110 (ref. 15). Several primer pairs have been designed and are being evaluated on a large scale with clinical specimens. Homology studies of DNA and protein sequences indicate that TRC4 is not related to any known IS elements except the insertion sequences of *E. coli* and *Shigella sonneii* with which it shows partial homology. The absence of any large nucleotide sequence identities between TRC4 and other characterized prokaryotic IS element classifies TRC4 as a new genetic element. The structural features qualify TRC4 as an insertion element which has seven inverted repeats, one of them generating a 4 bp target site duplication on integration.

Repetitive elements can act as agents of chromosomal rearrangement. As regions of portable homology, insertion elements can also be substrates for host recombinative pathways, giving rise to large scale deletions, duplications and inversions[1]. Thus, IS elements can promote rearrangements of genomes or replicons. Such a recombination event could be either recA and recBC dependent or independent[1].

'T-fasta' search revealed that among various other minor homologies, there is a borderline significant homology with RNA polymerase (rboB) gene of *M. leprae* at the indicated location. It will be interesting to explore whether this region of *M. leprae* also has related insertion element-like features[13].

The trend of the homology to known proteins is restricted to the initial and distal ends of the fragment ranging from 1 to 500 bp and 1200 to 2126 bp. The region between 600 and 1200 does not show much homology to known proteins except an activator and sigma factor. This simulates certain sequence elements of lambda phage reported to mediate recombination events. When the promoter of a gene is bracketed by inverted repeats, its orientation can be changed by recombination (inversion) between the repeats resulting in the reversible alteration of the 'on' and 'off' stages of gene expression[16]. It is not known if the complex rearrangement mediated by inverted repeats in phages and plasmids also occurs in the chromosomes of cells. Such events have been reported to occur in Leishmania chromosome[17].

To gain further insight into this interpretation, a series of future experiments have to be pursued. The initial and distal fragments of TRC4 would be subcloned in expression vectors to characterize the gene. The putative promoter sequences would also be confirmed by using primer extension studies. Further experiments have to be designed to confirm the role of inverted repeats.

The seven inverted repeats and their location bracketing the promoter-like sequences implicate the proteins coded by this sequence to play an important role in the metabolism of the organism and consequently in the pathogenesis of the disease. This also implicates that this fragment lies in the essential region of the genome. Elucidation of the role of such elements in the evolutionary process may provide valuable information on pathogenicity and this should be encouraged.

1. Iida, S., Meyer, J. and Arber, in *Mobile Genetic Elements* (ed. Shapiro, J. A.), Academic Press, New York, 1983.
2. Clark Curtis, J. E. and Docherty, M. A., *J. Inf. Dis.*, 1989, **159**, 7–15.
3. Grosskinsky, C. M., Jacobs Jr., W. R., Clark-Curtis, J. E. and Bloom, B. R., *Infect. Immun.*, 1989, **57**, 1535–1541.
4. Thierry, D., Cave, M. D., Eisenach, K. D., Crawford, J. T., Bates, J. H., Gicquel, B. and Guesdon, J. L., *Nucleic Acids Res.*, 1990, **18**, 188.
5. Hermans, P. W. M., van Soolingen, D., Dale, J. W., Schuitema, A. R. J., Mc Adam, R. A., Catty, D. and van Embden, J. D. A., *J. Clin. Microbiol.*, 1990, **28**, 2051–2058.
6. Eisenach, K., Crawford, J. T. and Bates, J. H., *J. Clin. Microbiol.*, 1988, **26**, 2240–2245.
7. Green, E. P., Tizard, M. C. V., Moss, M. T., Thompson, D. J., Winterboune, J. J., Mc Fadden and Hermon Taylor, J., *Nucleic Acids Res.*, 1989, **17**, 9063–9073.
8. Clark Curtis, J. E. and Gerald, P. W., *J. Bacteriol.*, 1989, **171**, 4844–4851.
9. Maniatis, T., Fritsch, E. F. and Sambrook, J., *Molecular Cloning: A Laboratory Manual*, Cold Spring Harbor Laboratory, Coldspring Harbor, 1982, 1st edn.
10. Oslevoska, E. and Homes, K., *Trends Genet.*, 1988, **5**, 92–94.
11. Southern, E. M., *J. Mol. Biol.*, 1975, **97**, 503–517.
12. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J., *J. Mol. Biol.*, 1990, **215**, 403–410.
13. Wayne, L. G. and Kubica, G. P., *Bergey's Manual of Systematic Bacteriology* (eds Sneath, P. H. A., Mair, N. and Sharpe, M. C.), Williams & Wilkins, Baltimore, Maryland, 1986, vol. 2, pp. 1436–1457.
14. Honore, N., Bergh, S., Chanteau, S., Doucet-Populaire, F., Eiglmeier, K., Gamier, T., Georges, C., Launois, P., Limpaiboon, T., Newton, S., Niang, K., del Portillo, P., Ramesh, G. R., Reddi, P., Ridel, P. R., Sittisombut, N., Wu-Hunter, S. and Cole, S. T., *Mol. Microbiol.*, 1993, **7**, 207–214.
15. Das, S., Paramasivan, C. N., Lowrie, D. B., Prabhakar, R. and Narayanan, P. R., *Tuberc. Lung. Dis.*, 1995, **76**, 550–554.
16. Bi, X. V. and Liu, L. F., *Proc. Natl. Acad. Sci. USA*, 1996, **93**, 819–823.
17. White, T. C., Fase-Fowler, F., Van Luenen, H., Calafat, J. and Borst, P., *J. Biol. Chem.*, 1988, **263**, 16977–16983.

ACKNOWLEDGEMENTS. We thank Dr Roy Gross, Lehrstuhl fur Microbiologie, Am Hubland, Wurzburg, Germany for 'T-fasta' searches. Part of the work has been supported by CSIR, Sr Research Fellowship to R.S. We thank Dr C. N. Paramasivan, Head, Bacteriology Department for giving the *M. tuberculosis* strains.

Received 19 May 1997; revised accepted 3 July 1997