Microbiology (1998), **144**, 1189–1196 Printed in Great Britain

Genetic diversity in the *Mycobacterium tuberculosis* complex based on variable numbers of tandem DNA repeats

Richard Frothingham[1,2] and Winifred A. Meeker-O'Connell[1,2]

Author for correspondence: Richard Frothingham. Tel: +1 919 286 0411 ext. 6566. Fax: +1 919 286 0264. e-mail: richard.frothingham@duke.edu

[1] Veterans Affairs Medical Center, 508 Fulton Street, Building 4, Durham, NC 27705, USA

[2] Department of Medicine, Duke University Medical Center, Box 31080, Durham, NC 27710, USA

Genetic loci containing variable numbers of tandem repeats (VNTR loci) form the basis for human gene mapping and identification, forensic analysis and paternity testing. The variability of bacterial tandem repeats has not been systematically studied. Eleven tandem repeat loci in the *M. tuberculosis* genome were analysed. Five major polymorphic tandem repeat (MPTR) loci contained 15-bp repeats with substantial sequence variation in adjacent copies. Six exact tandem repeat (ETR) loci contained large DNA repeats with identical sequences in adjacent repeats. These 11 loci were amplified in 48 strains to determine the number of tandem repeats at each locus. The strains analysed included 25 wild-type strains of *M. tuberculosis*, *M. bovis*, *M. africanum* and *M. microti* and 23 substrains of the attenuated *M. bovis* BCG vaccine. One of the five MPTR loci and all six ETR loci had length polymorphisms corresponding to insertions or deletions of tandem repeats. Most ETR loci were located in intergenic regions where copy number may influence expression of downstream genes. Each ETR locus had multiple alleles in the panel. Combined analysis identified 22 distinct allele profiles in 25 wild-type strains of the *M. tuberculosis* complex and five allele profiles in 23 *M. bovis* BCG substrains. Allele profiles were reproducible and stable, as demonstrated by analyses of multiple isolates of particular reference strains obtained from different laboratories. VNTR typing may be generally useful for strain differentiation and evolutionary studies in bacteria.

Keywords: tuberculosis, *Mycobacterium tuberculosis*, tandem repeat DNA, bacterial strain differentiation, BCG vaccine

INTRODUCTION

Mycobacterium tuberculosis is the leading cause of adult death due to a single infectious agent worldwide (Raviglione *et al*., 1995). Other members of the *M. tuberculosis* complex (*Mycobacterium africanum*, *Mycobacterium bovis*, *Mycobacterium microti*) also cause disease in humans and animals (Wayne & Kubica, 1986). Although strains of the *M. tuberculosis* complex vary in host range, virulence and other phenotypes (National Institutes of Health, 1980; Wayne, 1982), they have highly conserved DNA sequences (Kapur *et al*., 1994; Frothingham *et al*., 1994).

Genetic loci containing variable numbers of tandem repeats (VNTR loci) form the basis for human genetic mapping. Many medically important genes have been identified based on their linkage to a mapped VNTR locus. VNTR loci are also used for human forensic and paternity testing. Individual VNTR loci have been identified in bacteria (Andersen *et al*., 1996; Frénay *et al*., 1994; Frothingham, 1995; Goyal *et al*., 1994) but no systematic analysis has been reported. We undertook a systematic analysis of the variability of tandem repeat loci in the *M. tuberculosis* complex.

We identified 11 tandem repeat loci in the genome of *M. tuberculosis* H37Rv (type strain) by reviewing published literature and by searching cosmid sequences from a

Abbreviations: ETR, exact tandem repeat; MPTR, major polymorphic tandem repeat; VNTR, variable number of tandem repeats.

The GenBank accession numbers for the sequences reported in this paper are listed in Table 1.

Table 1. Summary of tandem repeat loci studied

Locus name	Location in H37Rv map (kb)*	GenBank accession no.	Sequence of PCR primers (5′–3′)†	No. and size of repeat units in H37Rv (bp)‡	Size of PCR product in H37Rv (bp)§	Result of PCR analysis of 48 strains‖	Reference
ETR-A	3820	S77045	AAATCGGTCCCATCACCTTCTTAT CGAAGCCTGGGGTGCCCGCGATTT	(3 × 75) + 23	420	Variable, 7 alleles	Goyal *et al.* (1994)
ETR-B	4160	Z70283	GCGAACACCAGGACAGCATCATG GGCATGCCGGTGATCGAGTGG	(3 × 57) + 8	292	Variable, 6 alleles	This work
ETR-C	1480	Z77162	GTGAGTCGCTGCAGAACCTGCAG GGCGTCTTGACCTCCACGAGTG	(4 × 58) − 21	276	Variable, 5 alleles	This work
ETR-D	1480	Z77162	CAGGTCACAACGAGAGGAAGAGC GCGGATCGGCCAGCGACTCCTC	(3 × 77) + 7	310	Variable, 6 alleles	This work
ETR-E	450	Z74024	CTTCGGCGTCGAAGAGAGCCTC CGGAACGCTGGTCACCACCTAAG	(3 × 53) − 1	224	Variable, 6 alleles	This work
ETR-F	450	Z74697	CTCGGTGATGGTCCGGCCGGTCAC GGAAGTGCTCGACAACGCCATGCC	(3 × 79) − 13	476	Variable, 3 alleles	This work
MPTR-A	2160	M15467	GGTTACCACTTCGATGCGTCTGCG AGCCGCCGAAACCCATC	16 × 15	343	Variable, 3 alleles	Frothingham (1995)
MPTR-B	1010	X60430	CTGGTAGTCGCCGCGCCACAG CTTAATGCGGCAGTTTCAAC	12 × 15	227	Not variable	Hermans *et al.* (1992)
MPTR-C	1010	X60431	CTGAACAGCCTCGTGATCA GAATTCCGGAACGGGCAG	15 × 15	267	Not variable	Hermans *et al.* (1992)
MPTR-D	2640	Z73101	CAAGCCCGAGGTGAATCTG CGGTCACTCAAGGCGTCGC	10 × 15	214	Not variable	This work
MPTR-E	2160	M15467	CTCAAAGCCGCCGTGCTCATGC GATCACCAGATGGGTTTC	14 × 15	314	Not variable	Shinnick (1987)

* Approximate *M. tuberculosis* H37Rv genome map locations based on Philipp *et al.* (1996).

† MPTR-A primers are from Frothingham (1995). Other primers were derived as part of this study. (Some strains were amplified using earlier versions of these primer sets, yielding the same results.)

‡ Each ETR locus had several complete repeats and one partial repeat in the type strain (*M. tuberculosis* H37Rv). For example, the ETR-A locus contained three complete 75-bp repeats followed by an additional 23 bp of repetitive sequence. The ETR-F locus contained three 79-bp repeats; the first two copies were complete but the third copy was lacking the last 13 bp.

§ PCR primers were complementary to DNA flanking each locus. Size includes the tandem repeat locus, flanking DNA and the primers.

‖ Each allele corresponds to a different number of tandem repeat units as listed in Tables 2 and 3.

genome sequencing project (Table 1). DNA from 48 strains of the *M. tuberculosis* complex in these 11 loci was amplified to determine the number of tandem repeats at each locus in each strain. We included a diverse collection of 25 reference strains with a broad geographic and host distribution, including all four species of the *M. tuberculosis* complex (*M. tuberculosis*, *M. bovis*, *M. africanum* and *M. microti*). We also included 23 *M. bovis* BCG strains which were known to be clonally derived (Fomukong *et al.*, 1992).

METHODS

Strains analysed. DNA samples from 48 strains of the *M. tuberculosis* complex, including 25 wild-type strains of *M. tuberculosis*, *M. bovis*, *M. africanum* and *M. microti* and 23 substrains of *M. bovis* BCG were used. The strain name, bacterial species, host species and geographic origin for each strain are listed in Tables 2 and 3. Primary references for these strains are cited in Frothingham (1995) and Talbot *et al.* (1997). Phenotypic and genetic data are available from previous publications, including the results of standard IS*6110* fingerprinting and PFGE for some strains (National Institutes of Health, 1980; Frothingham, 1995; Fomukong *et al.*, 1992; Cave *et al.*, 1991, 1992; Talbot *et al.*, 1997; Zhang *et al.*, 1995).

DNA preparation. We used DNA samples from several laboratories prepared by several methods. Strains were killed by heat or ethanol fixation (Williams *et al.*, 1995). DNA was released by agitation with glass beads (Frothingham, 1995), snap freeze-thawing (Williams *et al.*, 1995), boiling (Zwadyk *et al.*, 1994) or a complex protocol (Cave *et al.*, 1994). DNA was used in PCR with no further purification. DNA samples were coded to blind gel readers from strain identities.

Identification of tandem repeat loci. We identified tandem repeat loci by searching *M. tuberculosis* H37Rv sequence cosmids generated by the Sanger Centre in Cambridge, UK. The cosmids themselves originated at the Institut Pasteur, Paris, France (Philipp *et al.*, 1996). We used the REPEAT program of the Wisconsin Sequence Analysis Package (Genetics Computer Group, 1994) to search for DNA repeats with a length of $\geqslant$45 bp, an identity of $\geqslant$90% and with at least two copies of the repeat unit located within 200 bp. We also identified new tandem repeat loci by searching for sequences with homology to known loci.

PCR and sequencing. PCR was performed in a total volume of 25 µl. The PCR mix contained 2·5 µl GeneAmp 10× PCR Buffer II (Perkin-Elmer Cetus), 2 mM $MgCl_2$, 100 nM each primer, 200 µM each of the four dNTPs and 0·625 U AmpliTaq Gold DNA Polymerase (Perkin-Elmer Cetus). A primer pair was designed to anneal upstream and downstream of each tandem repeat locus (Fig. 1, Table 1). An initial denaturation of 12 min at 95 °C was followed by 35 cycles of denaturation at 94 °C for 30 s, annealing at 60 °C for 1 min and extension at 72 °C for 2 min, followed by a final extension at 72 °C for 10 min. Multiple interspersed negative controls (reagents only, no DNA) were included each time PCR was performed. The

Table 2. Allele profiles of 25 strains of the *M. tuberculosis* complex analysed by VNTR typing

Species	Strain name and description (no. of isolates analysed)	IS*6110* RFLP*		No. of tandem repeat units at locus†							Allele profile‡
		Copy no.	Pattern	MPTR-A	ETR-A	ETR-B	ETR-C	ETR-D	ETR-E	ETR-F	
M. africanum	ATCC 25420, human, Senegal, type	?	?	15	6	4	5	4	4	3	5645443
M. africanum	34/92, human, Sierra Leone	?	?	15	6	4	5	4	4	3	5645443
M. africanum	31/92, human, Sierra Leone	?	?	15	6	4	5	4	5	NP	564545?
M. bovis	243, human, Switzerland	5	Unique	16	2	2	2	3	3	3	6222333
M. bovis	10983, human, Switzerland	1	A	16	2	2	2	3	3	3	6222333
M. bovis	TMC 409, host unknown, France	2	Unique	16	5	NP	3	4	3	2	65?3432
M. bovis	4306, human, Switzerland	4	Unique	16	6	4	3	4	3	2	6643432
M. bovis	TMC 401, bovine, Wisconsin (2)	3	Unique	17	7	5	5	4	3	3	7755433
M. bovis	TMC 410, bovine, Iowa, type	1	A	16	7	5	5	4	2	2	6755422
M. bovis	TMC 412, bovine, England	2	Unique	16	5	5	5	4	2	1	6555421
M. microti	ATCC 19422, vole, England, type	?	?	16	5	3	5	7	1	2	6535712
M. tuberculosis	CSU-22, human, India	?	?	16	4	1	4	3	5	2	6414352
M. tuberculosis	T5, human, Arkansas	13	Unique	17	3	2	3	3	3	2	7323332
M. tuberculosis	T1, human, Arkansas	13	Unique	16	3	2	3	3	3	2	6323332
M. tuberculosis	3908-83, human, New York	?	?	16	3	2	3	3	3	3	6323333
M. tuberculosis	TMC 107, human, Minnesota	?	?	17	3	2	3	3	3	3	7323333
M. tuberculosis	T2, human, Arkansas	12	Unique	16	3	2	3	3	4	3	6323343
M. tuberculosis	T3, human, Arkansas	10	Unique	16	2	2	3	3	3	3	6223333
M. tuberculosis	T4, human, Arkansas	14	Unique	16	3	2	4	3	3	3	6324333
M. tuberculosis	3194, human, New York	?	?	16	1	2	4	3	2	1	6124321
M. tuberculosis	CSU-24, human, Japan	?	?	16	4	2	4	3	5	3	6424353
M. tuberculosis	H37Rv, human, New York, type (3)	9	Unique	16	3	3	4	3	3	3	6334333
M. tuberculosis	H37Ra, human, New York (4)	9	Unique	16	3	3	4	3	3	3	6334333
M. tuberculosis	CSU-23, human, India	?	?	16	4	6	4	6	4	3	6464643
M. tuberculosis	CSU-28, human, Thailand	?	?	16	4	6	4	6	2	3	6464623

* Data are listed for strains whose IS*6110* fingerprints have been published (Fomukong *et al*., 1992; Cave *et al*., 1991, 1992). ?, no data available. Patterns listed as 'unique' differ from all other patterns in Tables 2 and 3. Patterns A and B correspond to particular one- and two-band IS*6110* fingerprints.

† NP, No PCR product.

‡ Each digit of the seven-digit allele profile represents the number of copies at one of the seven VNTR loci.

positive control was 500 pg DNA from *M. tuberculosis* H37Rv (TMC 102, type strain). The presence and size of each PCR product was determined by electrophoresis on an agarose gel in Tris/boric acid/EDTA buffer followed by staining with ethidium bromide. PCR conditions were not specifically optimized for any of the primer pairs. Selected PCR products were sequenced directly with the *Taq* Dye Deoxy Terminator Cycle Sequencing Kit on a 373A DNA Sequencer (Applied Biosystems). Both PCR primers were used as sequencing primers in each locus.

RESULTS

We searched *M. tuberculosis* DNA sequences present in databases, as described in Methods, and identified 11 tandem repeat loci for analysis, including five major polymorphic tandem repeat (MPTR) loci and six exact tandem repeat (ETR) loci. These loci were distributed widely in the 4600 kb *M. tuberculosis* H37Rv genome (Table 1). We determined the variability of each locus by amplifying DNA from 48 strains. The PCR product amplified from *M. tuberculosis* H37Rv in each locus had the predicted size. PCR products from other strains were either the same size as the H37Rv product, or had sizes which corresponded to insertion or deletion of tandem repeat units. Length polymorphisms were easily identified on agarose gels (Fig. 2). Based on the size of the PCR products, we determined the exact number of tandem repeats at each locus in each strain.

MPTR loci

The MPTR consists of 15-bp repeats with a single consensus sequence but with substantial sequence variability between adjacent repeats (Hermans *et al*., 1992). We amplified the 48 strains in our panel in five MPTR loci, including four published loci (Hermans *et al*., 1992; Shinnick, 1987) and one locus identified by our search of cosmid sequences. One of these MPTR loci (MPTR-A) demonstrated length polymorphism, corresponding to 15, 16 or 17 copies of the 15-bp repeat unit (Tables 2 and 3). Results in the MPTR-A locus for most of these strains were published by Frothingham (1995). The other four MPTR loci demonstrated no length polymorphisms in our panel of 48 strains.

ETR loci

One ETR locus was previously identified as a VNTR locus (Goyal *et al*., 1994). We identified five additional ETR loci by searching *M. tuberculosis* cosmid sequences. Each of the six ETR loci contained large tandem repeats with identical DNA sequences in ad-

Table 3. Allele profiles of 23 *M. bovis* BCG strains analysed by VNTR typing

Strain name and description (no. of isolates analysed)	IS*6110* RFLP*		No. of tandem repeat units at locus							Allele profile†
	Copy no.	Pattern	MPTR-A	ETR-A	ETR-B	ETR-C	ETR-D	ETR-E	ETR-F	
TMC 1012, BCG Montreal	?	?	16	5	5	5	1	3	2	6555132
TMC 1025, BCG Prague	1	A	16	5	5	5	1	3	2	6555132
TMC 1030, BCG Connaught	?	?	16	5	5	5	1	3	2	6555132
ATCC 35736, BCG Brazilian	2	B	16	5	5	5	1	3	2	6555132
TMC 1103, BCG Montreal INH-R	?	?	16	5	5	5	1	3	2	6555132
ATCC 19274, BCG Calmette	?	?	16	5	3	5	2	3	2	6535232
KE1003, BCG, human, Tennessee	?	?	16	5	3	5	2	3	2	6535232
B140, BCG, human, California	?	?	16	5	5	5	2	3	2	6555232
B142, BCG, human, California	?	?	16	5	5	5	2	3	2	6555232
TMC 1010, BCG Danish	1	A	16	5	5	5	2	3	2	6555232
TMC 1020, BCG Mexican	?	?	16	5	5	5	2	3	2	6555232
TMC 1021, BCG Australian	?	?	16	5	5	5	2	3	2	6555232
TMC 1028, BCG Tice	1	A	16	5	5	5	2	3	2	6555232
TMC 1029, BCG Phipps	?	?	16	5	5	5	2	3	2	6555232
ATCC 35746, BCG Montreal SM-R	?	?	16	5	5	5	2	3	2	6555232
TMC 1022, BCG Russian	2	B	16	5	5	5	2	3	2	6555232
ATCC 27290, BCG Copenhagen	1	A	16	5	5	5	3	3	2	6555332
ATCC 35741, BCG Glaxo (2)	1	A	16	5	5	5	3	3	2	6555332
TMC 1002, BCG Birkhaug	?	?	16	5	5	5	3	3	2	6555332
TMC 1009, BCG Swedish	?	?	16	5	5	5	3	3	2	6555332
TMC 1019, BCG Japanese	2	B	16	5	5	5	3	3	2	6555332
ATCC 35734, BCG Pasteur (2)	1	A	16	5	5	6	2	3	2	6556232
TMC 1108, BCG Pasteur SM-R	?	?	16	5	5	6	2	3	2	6556232

*Data are listed for strains whose IS*6110* fingerprints have been published (Fomukong *et al*., 1992; Cave *et al*., 1991, 1992). ?, No data available. Patterns A and B correspond to particular one- and two-band IS*6110* fingerprints.

†Each digit of the seven-digit allele profile represents the number of copies at one of the seven VNTR loci.

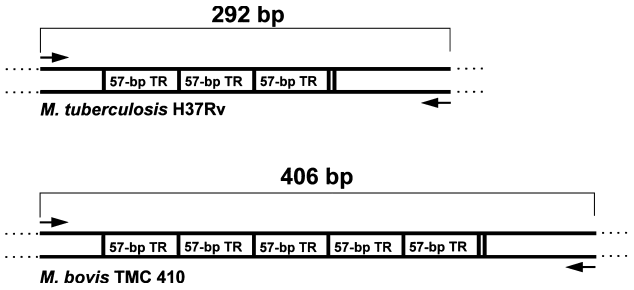

Fig. 1. Example of a VNTR locus. The figure shows genomic DNA at the ETR-B locus in *M. tuberculosis* H37Rv and *M. bovis* TMC 410. We amplified this locus using PCR primers complementary to flanking DNA (arrows) and sequenced the respective 292 and 406 bp PCR products. *M. tuberculosis* H37Rv DNA contains three complete copies of the 57-bp tandem repeat, plus eight additional bases corresponding to the beginning of another tandem repeat. *M. bovis* TMC 410 DNA has five complete copies plus the same eight additional bases.

jacent repeats (Fig. 1). Each locus had a unique repeat sequence; repeat units ranged from 53 to 79 bp. PCR products from all six ETR loci demonstrated substantial length polymorphism in our panel of 48 strains. Thus each ETR locus was a useful VNTR locus. The number of alleles found in our panel is listed for each locus in Table 1. Each allele corresponds to a different number of repeat units as determined by PCR. The exact number of tandem repeats at each locus in each strain is listed in Tables 2 and 3. For example, locus ETR-D is a VNTR locus with 6 alleles in our panel. These 6 alleles consist of 1, 2, 3, 4, 6 or 7 tandem copies of the 53-bp repeat unit.

PCR product sequencing

We sequenced PCR products from some loci to confirm that our PCR products corresponded to the expected regions. Sequences of the H37Rv PCR products from loci ETR-B, -C and -D were identical to the published sequences. We also sequenced PCR products from the ETR-B and -C loci in *M. bovis* TMC 410. These sequences differed from those of H37Rv by exact insertions of tandem repeats (Fig. 1). Sequence analysis of multiple PCR products in the MPTR-A and ETR-A loci also confirmed that length polymorphisms corresponded to insertion or deletion of complete tandem repeat units (Frothingham, 1995; Goyal *et al*., 1994).

Reproducibility

We amplified multiple DNA samples of *M. tuberculosis* H37Ra (four samples), *M. tuberculosis* H37Rv (three samples), *M. bovis* TMC 401 (two samples), *M. bovis* BCG Pasteur (two samples) and *M. bovis* BCG Glaxo (two samples). Each sample was obtained from a different source laboratory. Multiple DNA samples from each strain yielded PCR products of identical sizes at all loci. *M. africanum* 31/92 and *M. bovis* TMC 409

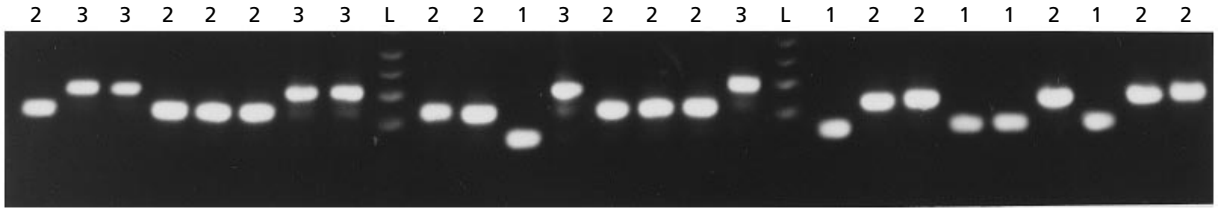

Fig. 2. Length polymorphisms in multiple *M. bovis* BCG substrains at the ETR-D locus. Locus ETR-D was amplified by PCR and subjected to electrophoresis on an agarose gel. Each lane contains a PCR product from a different *M. bovis* BCG substrain. Length polymorphisms corresponded to additions or deletions of the 77-bp tandem repeat unit. The number of tandem repeats at locus ETR-D is indicated for each substrain.

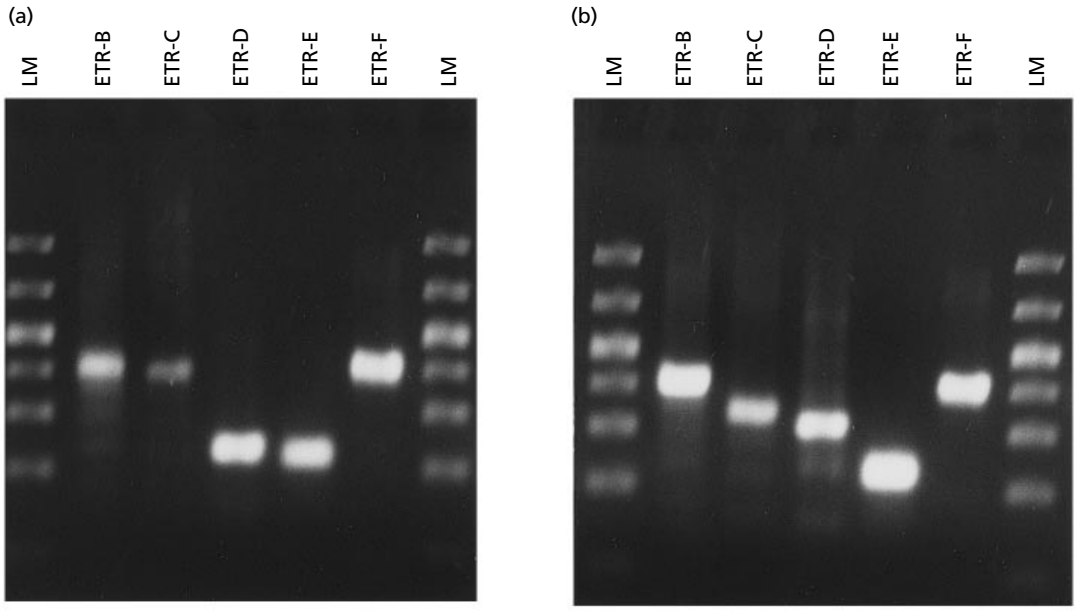

Fig. 3. Comparison of the allele profiles of *M. bovis* BCG Pasteur (a) and BCG Swedish (b). ETR loci were amplified by PCR and subjected to electrophoresis on an agarose gel. These two BCG substrains have a common ancestor but were serially passaged in separate countries from 1926 to 1961. The two strains have identical numbers of tandem repeats at five loci [MPTR-A and ETR-A (not shown), -B, -E and -F] and vary at two loci (ETR-C and -D). Large restriction fragment patterns from these two strains have many identical bands and a few bands which differ.

yielded no product at one locus each, possibly due to mutations or poor DNA samples.

VNTR allele profiles

When results for all seven VNTR loci were combined, 22 allele profiles were identified in the 25 *M. tuberculosis* complex strains (Table 2). The allele profiles from each species were distinct in this panel. Three *M. africanum* strains were divided into two profiles. Seven non-BCG *M. bovis* strains had 6 allele profiles. Fourteen strains of *M. tuberculosis* were divided into 13 profiles and the single strain of *M. microti* tested had a unique allele profile. Two pairs of strains with identical allele profiles have epidemiological links. *M. tuberculosis* H37Rv and H37Ra were clonally derived from H37R and *M. bovis* 243 and 10983 were both isolated from humans in Switzerland. Each of these pairs was distinguished by IS*6110* fingerprinting (Table 2). Five allele profiles were identified in the 23 substrains of *M. bovis* BCG (Table 3). As expected the VNTR allele profiles of these clonally derived strains were very similar to each other, differing at only one or two loci (Fig. 3).

Additional mutations in tandem repeat loci

For this analysis we determined only the number of tandem repeat units at each locus, based on agarose gel mobility. Our previous report also identified a nucleotide substitution in the MPTR-A locus, which was found only in BCG substrains (Frothingham, 1995). Some PCR products from the ETR-D and -F loci varied from the predicted sizes by about 20 bp. These variations are less than the size of the tandem repeat units (77 and

79 bp, respectively) and may represent independent mutations at these loci. We plan to sequence these PCR products to identify these mutations.

Additional ETR loci

We report here detailed results for the first six ETR loci we evaluated. We have identified additional ETR loci as additional *M. tuberculosis* sequences have become available. As of October 1997, we have identified a total of 34 ETR loci in 111 cosmid sequences from *M. tuberculosis* H37Rv (data not shown). Most are similar to the ETR loci described here, consisting of large repeat units with identical or nearly identical sequence in adjacent units and lacking spacer DNA. The sizes of the repeat units range from 51 to 111 bp. We have amplified a total of 15 ETR loci in multiple *M. tuberculosis* complex strains. At least 13 of these loci are informative VNTR loci. It appears that most ETR loci in *M. tuberculosis* are VNTR loci.

DISCUSSION

DNA sequences from various *M. tuberculosis* strains have highly conserved DNA sequences, despite considerable phenotypic variability (National Institutes of Health, 1980; Wayne, 1982; Hoffner *et al*., 1993). We analysed strains of all four species of the *M. tuberculosis* complex and found no nucleotide substitutions in an intergenic region which is variable in other mycobacteria (Frothingham *et al*., 1994; Frothingham & Wilson, 1993). Kapur *et al*. (1994) sequenced a total of 200000 bp of DNA from diverse *M. tuberculosis* strains and found only four synonymous nucleotide substitutions. The variability we observed in seven tandem repeat loci contrasts with the sequence conservation observed in other parts of the genome. Polymorphism in tandem repeat loci may be a significant genetic mechanism underlying phenotypic variations among *M. tuberculosis* strains.

VNTR loci in coding regions

Both MPTR-A and ETR-A are located within large ORFs. If these ORFs are expressed, then the DNA repeats encode amino acid repeats and variations in repeat numbers lead to size variations in the proteins. Other examples of bacterial VNTR loci occur inside ORFs (Andersen *et al*., 1996; Frénay *et al*., 1994). As expected all reported VNTR loci within bacterial coding regions have repeat sizes which are multiples of three.

VNTR loci in intergenic regions

The other five VNTR loci in Table 1 (ETR-B to -F) were located in intergenic spacers. The ETR-B locus is located downstream from two ORFs and contains a region of hairpin symmetry which may represent a bi-directional transcription terminator. ETR-C, -D, -E and -F were all located upstream from ORFs and in some cases overlapped the beginning of the ORF. ETR-D, -E and -F were similar to each other (about 80% identity) and are all located upstream from ORFs in the same orientation. These upstream tandem repeat loci may contain regulatory elements. Each repeat of ETR-D contains an ATG start codon and each repeat of ETR-E contains a consensus bacterial ribosome-binding site (5′ TGAGGAGGAGC) adjacent to an ATG start codon. In both cases, the final start codon is followed by a large ORF but the start codons in the other repeats are followed by short and probably spurious ORFs. Variations in the number of tandem repeats in these loci may influence expression of downstream genes.

VNTR loci in other mycobacteria

Large amounts of sequence data are currently available for *Mycobacterium leprae*. We identified segments of *M. leprae* DNA with substantial homology to several *M. tuberculosis* ETR loci. In each case, the *M. leprae* sequence contained only one copy of the DNA segment with similarity to the *M. tuberculosis* tandem repeats. We also searched *M. leprae* cosmids for tandem repeat DNA but found none. However, preliminary experiments (not shown) suggest that *Mycobacterium avium* has tandem repeats similar to the *M. tuberculosis* ETR loci.

Applications of VNTR strain typing

VNTR typing is the standard method for human forensic and paternity testing and may be useful for bacterial typing. Combined analysis of seven VNTR loci differentiated *M. tuberculosis* complex strains with reasonable power (Table 2). Potential clinical applications include epidemiological investigations, identification of outbreak-associated strains and recognition of laboratory cross-contamination. However, the clinical significance of *M. tuberculosis* clusters identified by VNTR typing is not yet known. VNTR typing by PCR has several advantages. It is a rapid and reproducible method. It can be performed on mycobacteria killed by heat or alcohol, reducing biohazards. Results are intrinsically digital, simplifying the comparison of large numbers of strains.

Evolutionary studies

VNTR analysis may be useful for tracing the phylogeny of the *M. tuberculosis* complex. Each VNTR locus represents a different portion of the genome and the loci identified so far appear to be independent. VNTR analysis thus provides multiple independent characters for phylogenetic analysis. PCR-based VNTR analysis may be applicable to archaic DNA, allowing direct confirmation of evolutionary hypotheses.

VNTR typing compared with IS*6110* fingerprinting

We used reference strains for our analyses so that VNTR typing could be compared to other methods. IS*6110* is a mobile insertion element present in multiple copies in *M. tuberculosis*. When used as a hybridization probe, IS*6110* identifies restriction fragment length

polymorphisms in clinical isolates. IS*6110* fingerprinting is the current standard method for *M. tuberculosis* strain differentiation (van Embden *et al*., 1993). We compared published IS*6110* fingerprints (Fomukong *et al*., 1992; Cave *et al*., 1991, 1992) with the results of VNTR typing (Tables 2 and 3, Fig. 2 legend). IS*6110* fingerprinting appears to be more discriminative than VNTR analysis for *M. tuberculosis* strains with high copy numbers. IS*6110* fingerprinting distinguished the H37Rv and H37Ra strains which have identical VNTR allele profiles. Also, strains T1 and T5 differed by multiple IS*6110* bands but by only one VNTR allele.

Population-based epidemiological surveys have shown that IS*6110* typing has poor discriminative power in low-copy-number *M. tuberculosis* and *M. bovis* strains (Braden, 1997; van Soolingen *et al*., 1993, 1994). IS*6110* fingerprinting was less discriminative than VNTR analysis for the *M. bovis* BCG substrains, all of which had one or two copies of IS*6110*. Also, two *M. bovis* (not BCG) strains with identical one-band IS*6110* fingerprints had distinct allele profiles. VNTR typing may be a useful adjunct to standard IS*6110* fingerprinting. IS*6110* fingerprinting and VNTR typing are directed at different targets and should provide independent information. For example, the nine BCG substrains typed by both methods segregated into two groups by *IS6110* fingerprinting, into four groups by VNTR typing and into seven groups when both analyses were combined (Table 3).

Comparison to PFGE

Zhang and colleagues analysed 25 BCG substrains, including many of the substrains in Table 3, using PFGE (Zhang *et al*., 1995). They digested chromosomal DNA using enzymes with rare sites (*Dra*I, *Asn*I, *Xba*I and *Spe*I) yielding large restriction fragment fingerprints. PFGE was more discriminative than either VNTR typing or IS*6110* fingerprinting in these BCG substrains. However, PFGE is a difficult technique to apply to mycobacteria.

Future directions

Further research is necessary to determine the clinical significance of clusters of *M. tuberculosis* strains with identical VNTR allele profiles. Inclusion of additional loci will increase the discriminative power of VNTR typing. VNTR typing is probably applicable to *M. avium* and may be useful for other bacterial species.

NOTE ADDED IN PROOF

While this work was under review, a report was published which describes the ETR-D locus in detail and discusses intergenic loci homologous to ETR-D, ETR-E and ETR-F (Supply *et al*., 1997).

ACKNOWLEDGEMENTS

We thank Percy L. Strickland for his characterization of additional ETR loci and for serving as a third independent gel interpreter for this work. We thank Elizabeth A. Talbot and Alison J. Cobb for critical review of early versions of this manuscript. We thank the following individuals who provided bacterial strains or DNA used in this study: M. Donald Cave and Kathleen D. Eisenach, University of Arkansas for Medical Sciences; Suzanne Glickman and W. Ray Butler, Centers for Disease Control and Prevention; Wayne M. Dankner and Charles E. Davis, University of California, San Diego; Norman J. Waeker, Naval Hospital, San Diego; Diana L. Williams, Louisiana State University; Gisela Bretzel and V. Sticht-Groh, Armauer-Hansen Institute, Würzburg, Germany; John T. Belisle and Patrick J. Brennan, Colorado State University; Kathy H. McDonald, Nancy Myers and Peter Zwadyk Jr, Durham VA Medical Center; Celeste M. McKnight and Robert P. Gruninger, Duke University Medical Center; Phyllis Pienta, American Type Culture Collection; and Kathryn M. Edwards, Vanderbilt University. This work was supported by NIH grant AI35230, the Durham VA Medical Center's Research Center on AIDS and HIV Infection and the Department of Veterans Affairs.

REFERENCES

Andersen, G. L., Simchock, J. M. & Wilson, K. H. (1996). Identification of a region of genetic variability among *Bacillus anthracis* strains and related species. *J Bacteriol* **178**, 377–384.

Braden, C. R. (1997). Current concepts in *Mycobacterium tuberculosis* DNA fingerprinting. *Infect Dis Clin Pract* **6**, 89–95.

Cave, M. D., Eisenach, K. D., McDermott, P. F., Bates, J. H. & Crawford, J. T. (1991). IS*6110*: conservation of sequence in the *Mycobacterium tuberculosis* complex and its utilization in DNA fingerprinting. *Mol Cell Probes* **5**, 73–80.

Cave, M. D., Eisenach, K. D., Salfinger, M., Bates, J. H. & Crawford, J. T. (1992). Usefulness of IS*6110* in fingerprinting DNA of *Mycobacterium bovis*. *Med Microbiol Lett* **1**, 96–102.

Cave, M. D., Eisenach, K. D., Templeton, G., Salfinger, M., Mazurek, G., Bates, J. H. & Crawford, J. T. (1994). Stability of DNA fingerprint pattern produced with IS*6110* in strains of *Mycobacterium tuberculosis*. *J Clin Microbiol* **32**, 262–266.

van Embden, J. D. A., Cave, M. D., Crawford, J. T., Dale, J. W., Eisenach, K. D., Gicquel, B., Hermans, P., Martin, C., McAdam, R., Shinnick, T. M. & Small, P. M. (1993). Strain identification of *Mycobacterium tuberculosis* by DNA fingerprinting: recommendations for a standardized methodology. *J Clin Microbiol* **31**, 406–409.

Fomukong, N. G., Dale, J. W., Osborn, T. W. & Grange, J. M. (1992). Use of gene probes based on the insertion sequence IS*986* to differentiate between BCG vaccine strains. *J Appl Bacteriol* **72**, 126–133.

Frénay, H. M. E., Theelen, J. P. G., Schouls, L. M., Vandenbroucke-Grauls, C. M. J. E., Verhoef, J., van Leeuwen, W. J. & Mooi, F. R. (1994). Discrimination of epidemic and non epidemic methicillin-resistant *Staphylococcus aureus* strains on the basis of protein A gene polymorphism. *J Clin Microbiol* **32**, 846–847.

Frothingham, R. (1995). Differentiation of strains in *Mycobacterium tuberculosis* complex by DNA sequence polymorphisms, including rapid identification of *M. bovis* BCG. *J Clin Microbiol* **33**, 840–844.

Frothingham, R. & Wilson, K. H. (1993). Sequence-based differentiation of strains in the *Mycobacterium avium* complex. *J Bacteriol* **175**, 2818–2825.

Frothingham, R., Hills, H. G. & Wilson, K. H. (1994). Extensive DNA sequence conservation throughout the *Mycobacterium tuberculosis* complex. *J Clin Microbiol* **32**, 1639–1643.

Genetics Computer Group (1994). *Wisconsin Sequence Analysis Package Program Manual*, 8th edn. Madison, WI: Genetics Computer Group.

Goyal, M., Young, D., Zhang, Y., Jenkins, P. A. & Shaw, R. J. (1994). PCR amplification of variable sequence upstream of *katG* gene to subdivide strains of *Mycobacterium tuberculosis* complex. *J Clin Microbiol* **32**, 3070–3071.

Hermans, P. W. M., van Soolingen, D. & van Embden, J. D. A. (1992). Characterization of a major polymorphic tandem repeat in *Mycobacterium tuberculosis* and its potential use in the epidemiology of *Mycobacterium kansasii* and *Mycobacterium gordonae*. *J Bacteriol* **174**, 4157–4165.

Hoffner, S. E., Svenson, S. B., Norberg, R., Dias, F., Ghebremichael, S. & Källenius, G. (1993). Biochemical heterogeneity of *Mycobacterium tuberculosis* complex isolates in Guinea-Bissau. *J Clin Microbiol* **31**, 2215–2217.

Kapur, V., Whittam, T. S. & Musser, J. M. (1994). Is *Mycobacterium tuberculosis* 15,000 years old? *J Infect Dis* **170**, 1348–1349.

National Institutes of Health (1980). *Mycobacterial Culture Collection*. NIH Publication No. 80-289. Bethesda, MD: National Institutes of Health.

Philipp, W. J., Poulet, S., Eiglmeier, K., Pascopella, L., Balasubramanian, V., Heym, B., Bergh, S., Bloom, B. R., Jacobs, W. R., Jr & Cole, S. T. (1996). An integrated map of the genome of the tubercle bacillus, *Mycobacterium tuberculosis* H37Rv, and comparison with *Mycobacterium leprae*. *Proc Natl Acad Sci USA* **93**, 3132–3137.

Raviglione, M. C., Snider, D. E., Jr & Kochi, A. (1995). Global epidemiology of tuberculosis: morbidity and mortality of a worldwide epidemic. *JAMA* **273**, 220–226.

Shinnick, T. M. (1987). The 65-kilodalton antigen of *Mycobacterium tuberculosis*. *J Bacteriol* **169**, 1080–1088.

van Soolingen, D., de Haas, P. E. W., Haagsma, J., Eger, T., Hermans, P. W. M., Ritacco, V., Alito, A. & van Embden, J. D. A. (1994). Use of various genetic markers in differentiation of *Mycobacterium bovis* strains from animals and humans and for studying epidemiology of bovine tuberculosis. *J Clin Microbiol* **32**, 2425–2433.

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

Supply, P., Magdalena, J., Himpens, S. & Locht, C. (1997). Identification of novel intergenic repetitive units in a mycobacterial two-component system operon. *Mol Microbiol* **26**, 991–1003.

Talbot, E. A., Williams, D. L. & Frothingham, R. (1997). PCR identification of *Mycobacterium bovis* BCG. *J Clin Microbiol* **35**, 566–569.

Wayne, L. G. (1982). Microbiology of tubercle bacilli. *Am Rev Respir Dis* **125 suppl.**, 31–41.

Wayne, L. G. & Kubica, G. P. (1986). The Mycobacteria. In *Bergey's Manual of Systematic Bacteriology*, vol. 2, pp. 1435–1457. Edited by P. H. A. Sneath, N. S. Mair, M. E. Sharpe & J. G. Holt. Baltimore: Williams & Wilkins.

Williams, D. L., Gillis, T. P. & Dupree, W. G. (1995). Ethanol fixation of sputum sediments for DNA-based detection of *Mycobacterium tuberculosis*. *J Clin Microbiol* **33**, 1558–1561.

Zhang, Y., Wallace, R. J., Jr & Mazurek, G. H. (1995). Genetic differences between BCG substrains. *Tubercle Lung Dis* **76**, 43–50.

Zwadyk, P., Jr, Down, J. A., Myers, N. & Dey, M. S. (1994). Rendering of mycobacteria safe for molecular diagnostic studies and development of a lysis method for strand displacement amplification and PCR. *J Clin Microbiol* **32**, 2140–2146.

Received 22 September 1997; revised 20 December 1997; accepted 21 January 1998.