New insertion sequences and a novel repeated sequence in the genome of *Mycobacterium tuberculosis* H37Rv

Stephen V. Gordon,¹ Beate Heym,² Julian Parkhill,³ Bart Barrell³ and Stewart T. Cole¹

Author for correspondence: Stewart T. Cole. Tel: +33 1 45 68 84 46. Fax: +33 1 40 61 35 83.
e-mail: stcole@pasteur.fr

¹ Unité de Génétique Moléculaire Bactérienne, Institut Pasteur, 28 Rue Dr Roux, 75724 Paris, Cedex 15, France

² Hôpital Ambroise Paré, 9 Avenue Charles-de-Gaulle, 92104 Boulogne Cedex, France

³ The Sanger Centre, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK

The genome sequence of *Mycobacterium tuberculosis* H37Rv was found to contain 56 loci with homology to insertion sequences (ISs). As well as the previously described IS6110, IS1081, IS1547 and IS-like elements, new ISs belonging to the IS3, IS5, IS21, IS30, IS110, IS256 and ISL3 families were identified. In addition, six ISs created a grouping of their own to form a new family (the IS1535 family). Elements with similarity to ISs in other actinomycetes were identified, suggesting the movement of ISs between related genera. The location of ISs on the chromosome revealed that an approximately 600 kb region close to the origin of replication lacks ISs, pointing to the possible detrimental effect of insertions in this area. Analysis of the distribution of ISs through the tubercle strains *Mycobacterium africanum*, *M. microti*, *M. bovis*, *M. bovis* BCG Pasteur, *M. tuberculosis* H37Ra, *M. tuberculosis* CSU#93 and 29 clinical isolates revealed that only IS1532, IS1533, IS1534, and IS1561' were absent from some of the strains tested. A novel repeated sequence, the REP13E12 family, is described that is present in seven copies on the *M. tuberculosis* H37Rv chromosome and which contains a probable phage attachment site. This study therefore offers an insight into the possible role of ISs and repetitive elements in the evolution of the *M. tuberculosis* genome, as well as identifying genetic markers that may be useful for phylogenetic and epidemiological analysis of the tubercle complex.

Keywords: *Mycobacterium tuberculosis*, insertion sequences, prophage

INTRODUCTION

Insertion sequence (IS) elements are small (<2.5 kb) segments of DNA that can insert at multiple sites in a target molecule (Mahillon & Chandler, 1998). The transposition reaction is catalysed by the IS-encoded transposase enzyme, with homology between transposases, organization of the ORFs, and fate of the target sites allowing elements to be placed in different families. These families are generally named after their prototypical IS element, such as the IS3 family or the IS256 family.

Before the completion of the genome sequence of

Mycobacterium tuberculosis H37Rv, four IS elements had been described in the *M. tuberculosis* genome, namely IS6110 (Thierry *et al.*, 1990), IS1081 (Collins & Stephens, 1991), IS1547 (Fang *et al.*, 1998) and the IS-like element (Mariani *et al.*, 1993). These elements are present in multiple copies on the genome of *M. tuberculosis* H37Rv, with 16 copies of IS6110, 6 copies of IS1081, and 2 copies each of IS1547 and the IS-like element (Cole *et al.*, 1998). The ISB9 element described by Musa and colleagues as an *M. tuberculosis* IS (EMBL accession number MTU78639) contains multiple frame-shift mutations in comparison with the *M. tuberculosis* H37Rv sequence and so does not resemble an IS element.

The variable copy number of IS6110 between different strains of the tubercle bacillus has led to its extensive use as a genetic marker to investigate the epidemiology of tuberculosis (Poulet & Cole, 1995; Yang *et al.*, 1998; Small *et al.*, 1994). However, despite the undoubted

Abbreviations: IR, inverted repeat; IS, insertion sequence; DR, direct repeat.

The GenBank accession numbers for the sequences reported in this paper are given in Tables 3 and 5.

value of IS6110 RFLP typing, questions persist concerning the rate of change of IS profiles and the discriminating power of these techniques (Yeh *et al.*, 1998; McHugh *et al.*, 1997; Helleyer *et al.*, 1996). The identification of new genetic markers in the genome sequence, in particular new IS elements that may be absent in some strains or show differences in copy number between strains, would provide additional loci with which to tackle these questions.

The presence of closely related IS elements in different species or genera may reveal that the organisms have occupied the same niche at some stage and have exchanged DNA. This has important implications for pathogenic bacteria, as IS elements may allow horizontal transfer of virulence or drug-resistance genes. In fact, the transfer of IS elements from disparate species into the mycobacteria has already been observed, with IS6100 being present in *Mycobacterium fortuitum*, *Arthrobacter* sp. and *Pseudomonas aeruginosa* (Kato *et al.*, 1994). Analysis of IS-containing loci in *M. tuberculosis* may reveal regions that have been introduced by the incoming element, resulting in horizontal transfer of genes. As the work of Musser and colleagues (Sreevatsan *et al.*, 1997) suggests that the tubercle complex (*M. tuberculosis*, *M. africanum*, *M. bovis*, *M. microti*) is a highly homogeneous group, it is possible that movement of IS elements provides the driving force behind genomic plasticity and variation within the complex.

The preferential integration of IS elements into insertional 'hot-spots' on the *M. tuberculosis* genome has been reported for IS6110 (Hermans *et al.*, 1991; Fang & Forbes, 1997; McHugh & Gillespie, 1998). Such hot-spots may prevent indiscriminate gene disruption by IS elements, as has been suggested for *Rhizobium* (Perret *et al.*, 1997). Conversely, areas of the genome lacking IS elements may identify regions that cannot support integration of foreign DNA. Analysis of the distribution of IS elements in the *M. tuberculosis* genome may therefore reveal important features of genome architecture. Additionally, identification of integration hot-spots may have important implications for transposon mutagenesis experiments as they could complicate the generation of representative mutant libraries. To answer the questions raised above we undertook an examination of the new IS elements in the genome of the tubercle complex reference strains and clinical isolates of *M. tuberculosis*.

METHODS

Bacterial strains. These are listed in Table 1. Non-epidemic clinical isolates of *M. tuberculosis* were collected at Hôpital Ambroise Paré, 9 avenue Charles-de-Gaulle, 92014 Boulogne Cedex, France, over a 6 month period. Strains were maintained on Löwenstein-Jensen slopes (Difco) at 37 °C, and if required grown in liquid Middlebrook 7H9 medium with ADC enrichment (Difco) at 37 °C with mild agitation (100 r.p.m.).

PCR amplifications. Primers for PCR reactions were designed such that one of the primers was present in the DNA flanking the element, while the other primer was contained in the element itself. This permitted a 'location-specific' PCR to

Table 1. Mycobacterial strains

Strain	Source
<i>M. africanum</i>	GMB*
<i>M. bovis</i>	GMB
<i>M. bovis</i> BCG Pasteur	Unité de Physiopathologie de l'Infection, Institut Pasteur, Paris
<i>M. microti</i> OV254	NIMR, Mill Hill, London, UK
<i>M. tuberculosis</i> clinical isolates 1-29	Hôpital Ambroise Paré, Paris
<i>M. tuberculosis</i> CSU#93	J. Belisle, Dept of Microbiology, Colorado State University, Fort Collins, CO 80523, USA
<i>M. tuberculosis</i> H37Ra	GMB
<i>M. tuberculosis</i> H37Rv	GMB

*GMB, Unité de Génétique Moléculaire Bactérienne, Institut Pasteur, Paris, France.

determine if the element was present in the same location in diverse strains. Sequences of the primers used are presented in Table 2. Strains that gave a negative result on the first screen were rechecked at least two more times. Strains that gave consistently negative results were checked with primers internal to the IS to verify that the IS was truly absent and that the result was not due to a mismatch of the primers or to the presence of the IS at a different chromosomal location. These internal primers are listed in Table 2 with the suffix '-int'.

Template DNA from clinical isolates was prepared by scraping colonies from the surface of Löwenstein-Jensen slopes, then fracturing the cells by subjecting them to five cycles of 1 min at 100 °C followed by 1 min in liquid nitrogen. Reactions were carried out in a total volume of 50 µl containing 5 µl PCR buffer (100 mM β-mercaptoethanol, 600 mM Tris/HCl, pH 8.8), 20 mM MgCl₂, 170 mM (NH₄)₂SO₄, 5 µl 20 mM nucleotide mix, 5 µl of each primer at a concentration of 2 µM, 5 µl of template DNA (20-50 ng), 5 µl DMSO, 0.5 units *Taq* polymerase (Boehringer Mannheim) and sterile distilled water to 50 µl. Thermal cycling was performed on a PTC-100 amplifier (MJ Research) with an initial denaturation step of 90 s at 95 °C, followed by 35 cycles of 30 s at 95 °C, 1 min at 55 °C, and 2 min at 72 °C.

Computer analysis. ORFs that showed similarity to transposase enzymes, but that were not identical to the previously described IS6110, IS1081 or IS1547 elements, were searched for inverted repeats (IRs) and direct repeats of the target sequence (DRs). Searching for IRs was performed using a combination of approaches: (i) by using the 'search for long-range inverted repeats' option of the Staden nucleotide interpretation program (NIP, version 7.3; Staden 1996); (ii) by using BESTFIT (Genetics Computer Group) to align one strand of the element with its reverse sequence, permitting local homology between the ends to be determined; (iii) where more than one copy of the IS existed the copies were aligned using GAP (Genetics Computer Group), allowing identification of homologous regions. Once IRs were identified, visual inspection of the sequence flanking the IRs, using the DIANA (Display and Analysis) software developed at the Sanger Centre, Cambridge, UK, was used to determine the presence of DRs. In cases where neither IRs nor DRs could be found,

the end of the element was delineated either by the break in identity with other copies of the element, or simply by the start and end of the ORFs.

Multiple alignments of DNA sequences were performed using CLUSTAL W (Thompson *et al.*, 1994). All other programs used to generate dendrograms were from the PHYLIP phylogeny suite (J. Felsenstein, University of Washington, USA). DNADIST was used to bootstrap the aligned sequences 100 times and generate distance matrices. NEIGHBOR was used to generate a treefile from the matrices using the method of neighbour-joining. A consensus tree was generated from the treefile using CONSENSE which was then edited in TREETOOL (M. Maciukenas, University of Illinois, USA).

IS designations. IS numbers were kindly provided by Dr Esther M. Lederberg, Plasmid Reference Center, Stanford University School of Medicine, Stanford, CA, USA.

RESULTS AND DISCUSSION

The criteria used to identify a sequence as being an IS element were the similarity (>25% identity) of the putative ORF product to transposase enzymes and the presence of IRs and DRs flanking the ORF. Analysis of the genome revealed 30 loci that fulfilled these criteria and that appeared to code for previously uncharacterized IS elements (Table 3). These new IS elements included members of the IS3, IS5, IS21, IS30, IS110, IS256 and ISL3 families. Six elements that showed high similarity to each other, but to no other ISs, were classified into a new family, the IS1535 family. Ten defective ISs that appear to be vestiges of once functional elements were marked with a prime (') and classed with the family that their transposase most resembled. The salient features of the new IS elements are dealt with below on a family-by-family basis.

IS3 family

The IS3 family forms the largest group of IS elements so far characterized in prokaryotes (Ohtsubo & Sekine, 1996; Mahillon & Chandler, 1998). The family is defined as containing two ORFs flanked by variable tandem repeats, although some variants, such as IS1076 and IS1141, have been described that contain a single ORF encoding the transposase (Mahillon & Chandler, 1998). This ORF arrangement was present in IS1540 and IS1604 (Table 3). The transposase of IS1540 shared high similarity with the IS1411 element of *Mycobacterium intracellulare* (GenBank accession no. L10239), which may point to the movement of IS elements between these strains. The transposase of IS1604 showed low similarity to the transposases of Tn552 and IS3 family members, so its grouping with the IS3 family is tentative. No appropriate IRs or DRs could be found for IS1540 or IS1604.

IS5 family

The IS5 family was represented in the genome by two distinct elements, IS1560 and the IS-like element described by Mariani *et al.* (1993). IS1560 was present in

two copies on the chromosome, with one copy appearing defective. The intact copy contained two minimally overlapping ORFs encoding proteins of 234 aa and 225 aa respectively. The overlap between the ORFs suggested that a 'ribosome slippage' mechanism may function to generate the mature transposase. Ribosome slippage sites are found in many overlapping ORFs encoding transposases, with the mechanism consisting of a -1 backward slippage of the translating ribosome at a heptanucleotide sequence (Chandler & Fayet, 1993). However, no appropriate slippage site could be found in the IS1560 sequence, suggesting that this mechanism may not function in the generation of the mature IS1560 transposase. The defective IS1560' was almost identical to the other copy at the left-hand end, but this similarity broke down after ~500 bp.

Analysis of the sequence of the IS-like element (Mariani *et al.*, 1993) for the presence of a ribosome slippage allowed us to identify a possible error in the published sequence, where there was a base missing between positions 410–411. This introduced a -1 frameshift into the ORF, making it appear that the IS element carried a single ORF. In fact, the IS element carries two overlapping ORFs, with a putative slippage site (ACA GGG GGC) between positions 432–437. This site may allow a -1 frameshift that would generate a single protein from the two ORFs.

IS21 family

Three IS sequences, IS1532, IS1533 and IS1534, were found that showed similarity to members of the IS21 family. Members of this family are generally large, in the range of 2–2.5 kb, and contain two ORFs flanked by related IRs of 11–50 bp (Mahillon & Chandler, 1998). The IRs at the ends of IS1532, IS1533 and IS1534 were found to contain duplicated internal sequences (Table 4), a feature found in many members of the IS21 family which is thought to play some role in the binding of the transposase to the ends of the element (Mahillon & Chandler, 1998).

IS30 family

IS1603 contained a single ORF encoding a protein of 334 aa that showed similarity with members of the IS30 family. However, the element possessed 63 bp IRs, unusually long for members of this family (Mahillon & Chandler, 1998; Olaz *et al.*, 1998). Close inspection of the IRs revealed that they contained a 32 bp segment with 100% identity to the IRs of IS1534 (Table 4). The significance of this observation is not known, but it may point to recombination between the elements.

IS110 family

IS1558 is present as an intact and a defective copy on the chromosome, with the intact copy encoding a protein of 333 aa flanked by 13 bp IRs. While the defective IS1558' had identical ends to the intact copy, the identity

Table 2. Primers used to screen for the presence of IS and REP elements in *M. tuberculosis* strains

IS element	Primer sequences (5'–3')	Product size (bp)
IS1532	GCTGCGTAATGCGCTACGTACC AAACCGAGATGACGATCAACAACC	898
IS1532-int*	TGCCCTTGATGGTGGGATAG GGCAACCCGATAGCACCG	1140
IS1533	CATCGCAGACACCACACTTGGTAG TTGATTCCAACGCTGACTTCACC	705
IS1533-int*	TTTCCAAGAACACGGTGAACTCAG GACGCTAAGGTTGCTGATGAGC	880
IS1534	GCACTTGCCATACTCACCACATC GGACTTCGTCACCCTAGCAATC	659
IS1534-int*	TGTAGTGTGGCGGCGGCAAC GATAAGCGCTGGCGGAGCGT	1719
IS1535	TTGCCTCCAGCCAAATCCAC TCGCCGAATTCCTTGACCAC	881
IS1536	GCCCCGTTGCATCGAATCAG GCCACAGCCGAACAGATCCC	1277
IS1537	GGATATAACGCCGCCGTCAG CCAGTTCAGTACGCATCGACAG	1022
IS1538	TACCCCGATCACCGCCGCC CCGCAAACGCTGCCTGCAC	694
IS1539	TCGATGACGTGCTGTTGAACGC TACAAGACAGGCTGGTACGGATCGAG	784
IS1540	AACACTCCAGCAATCGCAGCAC GCTGCTCGACGAGGGCATCTAC	944
IS1552	CGGTCAGCAACGACGAATACAG TCGAGGTGCAGATCCCCAAG	760
IS1553	GCCGATCAGCGATACCGAAC TCAACGCCTTCACCAGCCC	983
IS1554	CCAACCAATCACCGACCACAG CGGGTCATCGAGGAGATGCAG	978
IS1604	TGCGCGAAAAACCGTAATACGAG CGATGAACGCACCTGCAAC	1065
IS1556	GAGTATGGGACGTGAGCGAC GCTGATGAACCGCCACGATG	1060
IS1557-1	GCATTTCTGTCGTGGTTGGAG GGTGTGCGAGTTGATGCGG	1053
IS1557-2	TCCGACGATGACGACACCAC CCGCATCAACTCGCACACC	866
IS1558	GCAAAACCGACCTCTTGATG CGACCGGTGCCACTAGCTGGT	979
IS1561'	GACCTGACGCCGCTGACAC CACCTACACCGCTTCCTGCC	530
IS1602	CGTAGACGTAGCCGACGTAG GCACCACCGAAATATCCC	772
IS1603	CGATACAACAAGGGCGGACCTAC CAGCCGATGTTGTGATTACAC	905
IS1607	GCGAGTCAAGGTCAGCGAACAAC CGAACCCGAGCCGAGCAG	605
REP09F9	GCACGCACCTGCTCTATGCTC GCCTACCAACGCCTCGTTGTC	1585
REPI125	ACGCTAAGCGATGACCGTCAC GCAACTACAACGACACCGACC	895

Table 2. (cont.)

IS element	Primer sequences (5'-3')	Product size (bp)
REP13E12	CCAGGTGGGAACAGACGCAG TCGAGGTGGGGTGGTGGTAG	1442
REP22G8	TCAATCCCAGCTTGCCGCCAG TTCGCCAGCCAGAGAGGATCGTC	675
REP251	TGCCGTCGTATTGCTGGTTG TCTGGGTAAATCCCTTGTGTCATTTC	702
REP336	CGACTTCGTTGCGAATCGCC GTGCCGCCAGTTCTTCCTC	1130
REPI65	GCCGACCAATAGCGAAGAAATC GTGAGCACCACGCTAAAGGAG	671

* int primers are both internal to the IS element.

between the two elements broke down after the first 40 bp, resuming after a further 335 bp. This variant region caused the IS1558' transposase to be only 221 aa in length, lacking the first 112 aa, and most likely renders it inactive.

Another element with homology to the IS110 family was IS1607. This IS contained two overlapping ORFs that encoded proteins of 233 aa and 222 aa respectively. While the upstream ORF of IS1607 showed similarity to the N-terminal region of the transposases encoded by IS110, IS900, IS902 and the IS117 mini-circle element from *Streptomyces coelicolor*, the downstream ORF is similar to the C-terminal end of the above proteins. Although this would suggest that a form of translational frameshifting may function to make the mature transposase, no clear frameshifting site was found.

Two copies of a 1013 bp element, designated IS1608', that encoded a protein of 163 aa were located on the genome. The fact that only 489 bp of the element appeared to be coding suggested that the element may be defective, hence the prime suffix. IS1608' showed low similarity to IS116 from *Streptomyces clavuligerus*, and no IRs or DRs could be detected.

IS256-like family

IS1552, IS1553 and IS1554 showed marked similarity to members of the IS256 family. The transposase from IS1554 contained the 'mutator family' transposase signature as defined in the PROSITE database of protein sites and patterns (Bairoch *et al.*, 1997). This conserved domain is present in many members of the IS256 family, although its function is not clear.

A striking example of the possible exchange of IS elements between genera was illustrated by IS1552. The transposase of this element, as well as being similar to IS256 elements, shared 80% identity (91% similarity)

over a 278 aa overlap with a transposase encoded on the pHG201 plasmid of *Rhodococcus opacus* (Grzeszik *et al.*, 1997; Fig. 1). Members of the genus *Rhodococcus*, like the mycobacteria, are members of the nocardioform taxonomic grouping, and contain mycolic acids, have a genomic G+C content of 63–73 mol% and form substrate mycelia. Horizontal transfer of ISs between the mycobacteria and *Rhodococcus* has also been suggested by Picardeau *et al.* (1997), underlining the possible ability of these bacteria to swap genetic information.

ISL3 family

The IS1557 element was present as two complete copies and with a third truncated by approximately 1 kb. The complete copies contained a single long ORF that coded for a transposase with high similarity (67.9% identity in 168 aa overlap) to a transposase enzyme from the dibenzofuran-degrading bacterium *Terrabacter* (GenBank accession DBU57649), and with ISAE1 from *Alcaligenes eutrophus* (30.9% identity in 369 aa). The IRs of the two intact copies showed some variations (Table 4), perhaps reflecting errors introduced during transposition. The defective IS1561' showed homology to the ISL3 family, with the ORF products displaying similarity to IS204 from *Norcardia asteroides*.

IS1535 family

Six IS elements, IS1535 to IS1539 and IS1602, failed to show any significant similarity with IS elements from the main groupings. However, their putative transposases showed from 40.6 to 86.1% identity with each other, and they were therefore placed in a new grouping, the IS1535 family. These IS elements contained two overlapping ORFs where the stop codon from the upstream ORF overlapped the start codon of the downstream

Table 3. New IS elements in the genome of *M. tuberculosis* H37Rv

Family	Element*	Size (bp)	Structure†	IR size (bp)	DR size (bp)	Similarity‡	Genome position	Accession no.
IS3	IS1540	1162		ND	ND	<i>Mycobacterium intracellulare</i> IS1141 (64.1%/340 aa)	3847640–3848802	Z95389
	IS1604	1408		ND	ND	<i>Staphylococcus aureus</i> Tn552 (28%/254 aa)	3116814–3118222	Z81331
IS5	IS1560	1567		25	2	<i>Pseudomonas cepacia</i> IS402 (26.3%/186 aa)	3799983–3801550	AL009198
	IS1560'	1512		ND	ND	<i>Mycobacterium tuberculosis</i> IS1560 (82.8%/87 aa)	1158919–1160431	Z92539
IS21	IS1532	2609		48	4	<i>Pseudomonas alcaligenes</i> IS1491 (35.5%/391 aa)	3843787–3846396	Z77165
	IS1533	2212		54	5	<i>Bacillus stearothermophilus</i> IS5376 (23.9%/410 aa)	3288363–3290575	Z83858
	IS1534	2129		49	5	<i>Escherichia coli</i> IS21 (87.4%/135 aa)	4075657–4077786	Z95436
IS30	IS1603	1327		63	ND	<i>Pseudomonas alcaligenes</i> IS1394 (42.8%/334 aa)	3557290–3556817	AL021646
IS110	IS1558	1212		13	ND	<i>Streptomyces coelicolor</i> A3(2) IS110 (27.7%/238 aa)	2720642–2721854	Z81451
	IS1558'	803		ND	ND	<i>Streptomyces coelicolor</i> A3(2) IS110 (32.9%/149 aa)	2439143–2439946	AL021957
	IS1607	1227		ND	ND	<i>Streptomyces coelicolor</i> A3(2) IS117 (33.3%/207 aa)	2260441–2261668	Z74025
	IS1608'	1031		ND	ND	<i>Streptomyces clavuligerus</i> IS116 (23.5%/158 aa)	3753327–3754360	AL009198
	IS1608	1031		ND	ND	<i>Streptomyces clavuligerus</i> IS116 (23.5%/158 aa)	4318338–4319369	Z83864
IS256	IS1553	1292		13	ND	<i>Mycobacterium tuberculosis</i> IS1081 (46.8%/365 aa)	4078503–4079795	Z95436
	IS1554	1435		15	ND	<i>Burkholderia cepacia</i> IS1356 (41.9%/378 aa)	102546–1026891	Z95210
	IS1552'	844		ND	ND	<i>Rhodococcus opacus</i> Tpase (80.6%/278 aa)	3849292–3850136	Z95389
IS1535	IS1535	2322		17	ND	<i>Mycobacterium tuberculosis</i> IS1602 (40.6%/465 aa)	1027039–1029361	Z95210
	IS1536	1391		ND	ND	<i>Mycobacterium tuberculosis</i> IS1539 (83.5%/237 aa)	701363–702754	Z97182
	IS1537	1889		ND	ND	<i>Mycobacterium tuberculosis</i> IS1539 (84%/368 aa)	4301540–4303429	Z97188
	IS1538	2055		ND	ND	<i>Mycobacterium tuberculosis</i> IS1602 (86.1%/459 aa)	3333755–3335810	Z83018
	IS1539	2057		ND	ND	<i>Mycobacterium tuberculosis</i> IS1538 (83%/458 aa)	3194136–3196193	Z74024
	IS1602	2052		ND	ND	<i>Mycobacterium tuberculosis</i> IS1538 (86.1%/459 aa)	3100172–3102224	AL008967
	IS1605'	287		ND	ND	<i>Mycobacterium tuberculosis</i> IS1539 (98.9%/93aa)	921573–921860	AL022004
	IS1557	1513		20	ND	<i>Terrabacter</i> Tpase (67.9%/168 aa)	1468140–1469653	Z73419
ISL3	IS1557	1513		28	ND	<i>Terrabacter</i> Tpase (67.9%/168 aa)	4252851–4254364	AL022076
	IS1557'	516		ND	ND	<i>Mycobacterium tuberculosis</i> IS1557 (99.0%/104 aa)	832350–832866	AL021958
	IS1555'	398		ND	ND	<i>Mycobacterium smegmatis</i> IS1096 (56.3%/96 aa)	3115741–3116139	Z81331
	IS1561'	1319		ND	ND	<i>Nocardia asteroides</i> IS204 (48.5%/97 aa)	3754292–3755611	AL009198
	IS1606'	330		ND	ND	<i>Corynebacterium glutamicum</i> IS31831 (38.6%/57 aa)	947309–947639	AL022004
Unknown	IS1556	1468		ND	ND	<i>Mycobacterium bovis</i> IS987 (32.3%/133 aa)	2342940–2344408	Z73966

ND, Not detected.

* IS elements marked with a prime are assumed to be defective.

† The DNA of the IS is represented by a thick black line; arrows above the line represent ORFs, and short arrows under the line represent IRs.

‡ The IS which showed the highest similarity to the *M. tuberculosis* IS is shown, followed in parentheses by the percentage identity over the indicated stretch of amino acids.

Table 4. Inverted repeats

Element	Inverted repeat
IS1535*	IRL GGGTTCAGAGCTGTTGCGT <u>GTTGAGTGTGTTTTAGTGTGCGTTAGTGTG</u> IRR <u>CCGTTGAGTGTGTTTTAGTGTG</u> CACTCTCATGCGGCGTCCCCTTTGCGGG
IS1532†	IRL GTTGTCCGTCACGTAATT <u>CCCCACCCCG</u> -GCTGTCACGTAATT <u>CCCCA</u> IRR GTCGACCGCCACGTAATT <u>CCCCAGGGGTTGGCGTCACGTAATTCCCCA</u>
IS1533†	IRL TGTGACCGGCACGTGAAAACTG <u>ACCCCGGCGCGGCACCCGAATTTTGACCCCT</u> IRR TGTCAACGGCACCCGAAAACTG <u>ACCCCTGACGGCATCTGAAAATTGACCCCT</u>
IS1534†	IRL <u>CGGCAACTGAATACTGACCAGAGCGCGGCAACTGAAAATTGACCAGCTT</u> IRR <u>CGGCAACCGAAAACTGATCAGGTGT</u> <u>CGGCAATCGAAAATTGACCAGCTT</u>
IS1553	IRL GAGTTCGTCCGTG IRR GAGATCGTCCGTG
IS1554	IRL ATTCGGTGTAAGTGG IRR ATTCGGTGTAAGTGG
IS1560	IRL TAATTACTAGGACCTGAAAAAGTCG IRR TAATTACTAAGACCTGAAAAAGTCG
IS1557‡	IRL AGCAGACGCAAAAGCCCCCA IRR AGCAGACGCGAAAGCCCCCA
IS1557§	IRL CGAGCAGACGTAAAAGCCCCC-AATTCTG IRR CGAGCAGACGTAAAAGCCCCCAAATTCTG
IS1558	IRL GAAAATGCTGACG IRR GAAAACGCTGACG
IS1603	IRL TGTGCGCGGCAACTGAATACTGACCAGAGCGCGGCAAGGTGGTTTCTAGGCGGTGTCGCAACA IRR TGTGAGCGGCAACCGAAAACTGATCAGGTGTGCGGCAAGGTGGTTTCTAGTCAACGTGCGCAACA

*Inverted repeats in the IS1535 sequence are underlined.

†Direct repeats in the IRs of IS1532, IS1533 and IS1534 are underlined.

‡Copy at genome position 1468140–1469653.

§Copy at genome position 4252851–4254364.

|| The 32 bp identical region between the IRs of IS1534 and IS1603 is shown in bold.

ORF (TGATG). The upstream ORF product of IS1536, IS1537 and IS1538 had a site-specific resolvase motif as defined by the PROSITE database (Bairoch *et al.*, 1997), indicating that the protein may have resolvase activity. The downstream ORF coded for a protein with weak homology to bacterial transposases, with the highest score generally being in the range of 31–38% identity over an 80 aa overlap with the IS1136 element from *Saccharopolyspora erythraea* (Donadio & Staver, 1993). Although the transposase of IS1536 was truncated by ~200 aa, its left-hand end was intact, suggesting that the element may have been truncated during transposition.

This is similar to class II transposons such as Tn3, where the transposon carries genes for resolvase and transposase activity (Hatfull & Grindley, 1988). In such transposons, the initial cointegrate formation is catalysed by the transposase enzyme, with resolution of the cointegrate performed by the resolvase. It is possible that a similar mechanism functions during transposition of the IS1535 family. Furthermore, the very basic isoelectric points of the putative transposases, from 11.23 to 11.78, would give the proteins a strong net positive charge, facilitating binding of the enzyme to the target DNA. Although the ends of the IS1535 element

contain perfect internal 17 bp IRs (Table 4), no similar IRs could be found for the other members.

IS elements of unknown family

The element IS1556 was unusual, containing three overlapping ORFs that coded for proteins of 101, 217, and 121 aa respectively and that displayed low homology with transposases belonging to the IS3 family, in particular IS6110. The ORFs were frameshifted with respect to one another, with the downstream ORFs lacking an initiation codon. Although this suggested that a frameshifting mechanism may be involved in the production of the mature transposase, no sites for ribosome slippage could be found. The ORF structure of IS1556 was similar to that seen with the defective IS1560', suggesting that it is non-functional.

Novel repeated sequence: the 'REP13E12' family

In *M. tuberculosis* H37Rv the biotin operon is interrupted by the phiRv1 prophage (Cole *et al.*, 1998; Fig. 2). Examination of the sequence flanking the prophage revealed that it had integrated into an ORF that bore no resemblance to the other genes of the biotin operon. The ORF was in fact part of a family of repeated sequences,

IS1552	-----MFAELIR
<i>R. opacus</i>	-----MTHDHSALI-----AQLDGLKGADSASVFAELIR
IS1553	-----MALPQSALS-----ELLDAPRTGDGVLDLRDAVR
IS256	-----MTQVHFTLKS-----EETQSIIEYSVKDDVSKNILT
IS1554	MDAAQVIEPAHAGQDVDEAAVAARELSGAERALVGLVRQARABGVALTGPDGLLKALTK
:	
IS1552	AGLQALIEAEATEAIGAGRYERS-DGRIVHRNGHR-PKTVSTTAGDIEVQIPKLRAGSFF
<i>R. opacus</i>	AGMQALIDAEATEALGAGRYERA-EGRTTYRNGSR-AKTVATTSGDVELKIPKLRSGSFF
IS1553	LVLQELSELEATERIGAARYERS-DTRVTDNRGAR-SRVLSTQAGDVELRIPKLRKGSFF
IS256	TVFNQIMENQRTYIQAKEYERT-ENRQSQRINGY-ERSFTTRVGTLLELKVPRTRDGHFS
IS1554	TVLEAALQEEMTEHLGYDRHAAAGRGSGNSRNGSRNKKVITLACGQVEIAVPRDRNGTFE
	: : : * : : : * : : * : : * : *
IS1552	PSLLERRRRIDKALHAVIMEAYVHGVSSTRVDDLVAAMGVQAGVSKSEVSRI CAGLDTEI
<i>R. opacus</i>	PSLLERRRRIDRALHAVIMEAYVHGVSSTRNVDDL VGALGVSGISKSEVSRI CQQLDTEI
IS1553	PAILEPRRRIDQALYAVVMEAYVHGISTRAVDLVEAMGVETGISKSEVSRI CAGLDEIV
IS256	PTVFERYQRNEKALMASMLEMYVSGVSTRKVKI VEEL-CGKSVSKSFVSSILTEQLEPMV
IS1554	PVIVGKRKRRTVDVDRVLSLYAKGLTTGEIAAHFADV-YGVSVSKDTISRI TRDRIEEM
	* : : * : : : * : * :
IS1552	EAFRTSLTHTEFPYVFCDATFCKVRVG-AHVVSQALVATGVSIDGTREVLGTAVG--D
<i>R. opacus</i>	GRFRDRTLHTTFFPYVFLDATYCKVRVG-AHVVSALVATGVSMGDTREVLGTAVG--D
IS1553	GAFRTSLTHTEFPYVYLDATYLVNRNGTGQVSMMAVIVASGIAADGSRREILGLDVG--D
IS256	NEWQNRLLSEKNYPYLMTDVLYIKVREE-NRVLSKSHIAIGITKDGDRREIIGFMIQ--S
IS1554	QAWWSRPL-EKVYAAVFIDAIMVKIRDG--QVRNRPVYAAIGVDLDGHKDI LGMWAGEGD
	: * * . : : : * . : * : * . : * : * : * : * : * : * : * : * : *
IS1552	SESEYFWEFLASLKARGLTGVLVLSDAHAGLKAAVAQQFSGASWQRCRVHFMRLNYTA
<i>R. opacus</i>	SESFDWFREFLAGLRAGLCGVHLVLSDAHAGLKAAVAQQFSGSSWQRCRVHFMRLQSA
IS1553	SEDETFRWGLTSLKRGGLGVRLVLSQHAGLVKALKRCFQGAGHQRVHFAFARNLLAH
IS256	GESEETWITTFEYLKERGLQGTTELVISDAHKLVSARKSEINVSORCOVHFLRNIFTT
IS1554	GESAKFWLAVLTDLRNRGVKDIFFLVCDGLKGLPDSVSAAFPLATVOTCIHLIRNTFRY
	. * . * : : * : * : . : : * * : : * : * : * : * : * : * : * : *
IS1552	VAKHAPAVTVAVKTI FAHTDP-EEVGAQWDRVADPLCQP*
<i>R. opacus</i>	VHSHAPAVMAAVKTVFAHTDP-DEVAAQWDQVITDLAGS-FPKVSSMMREAKCDVLAFS
IS1553	VPKDKADMVAMFRMIF SAPDA-EAVHATWEGVRDLAAS-FPKIGPLMDDARAEVLAFT
IS256	IPKNSKSFREAVKGI FKFPTDINLAREAKNRLIHDYIDQPKYSKACASLDDGFEDAFQYT
IS1554	ASRKYWDKISVDLKPIYTAASA-AEARLYEEFAEKWGKP-YPAITRLWDSAWEEFI PFL
	. :
<i>R. opacus</i>	GFPRSHWQKIWSNNPIERLNKEIKRRADVVEIFPNPASFLRLATAVVIEAHDEWQVT-RR
IS1553	AFPKAHWQKIWSNPLERLNKEIKRRSRVVGIFPNPAVIRLVGAVLADMHDEWQASERR
IS256	VQGNH-NRLKSTNLIERLNQEVRRREKIRIFPNQTSANHLIGAVLMDLHDEWYSSRK
IS1554	DYDVEIRRVPCSTNAIESLNARYRRAVRARGHFPNEQSALKITLYLVTRSLDPKGTGQTKW
	. * : * : * : * : * : * : * : * : * : * : * : * : * : * : * : * : * : *
<i>R. opacus</i>	YVSDVSMDELRTIVIVAKERAAAQHATAESLTAETA
IS1553	YLSEASMLLYPDSINAVVAISGGQ-----
IS256	YINFDK-----
IS1554	AVRWKPALNALAITFADRMPEAER-----
	:

Fig. 1. Multiple alignment of the new IS256 family elements with the prototype IS256. The *R. opacus* pHG201-borne transposase is shown as *R. opacus*. In the alignment, an asterisk (*) denotes identity, a colon (:) that the position is identical in greater than half the proteins, and a period (.) that the position contains conserved amino acids in all sequences. A dot (•) marks the truncation of IS1552. The 'mutator family signature' as defined by the Prosite database is underlined.

Fig. 2. The phiRv1 prophage integrated into the REP336 locus. ORFs are denoted either by their proper gene name or by their 'Rv' designation. Black ORF boxes represent the REP336 element, with the truncated fragment marked *REP336'*. Open ORF boxes represent phage genes, with *Rv1586c* the putative phage integrase. Hatched ORF boxes represent other genes in the vicinity of the phage integration site.

Table 5. REP13E12 family

Sequence	Structure	Genome position	Accession no.
REP251	97 aa → 317 aa	103711–105099	Z74410
REP336	222 aa → 317 aa	1779195–1779263 1788512–1789834	Z95586
	phage		
REP13E12	222 aa → 317 aa	3883521–3884914	Z95390
REPI65	454 aa →	1276291–127727	Z95584
REPI125	454 aa →	1927216–1928587	Z98268
REP09F9	454 aa →	2195922–2197358	Z84998
REP22G8	451 aa →	1251619–1252943	Z95585

which was designated the REP13E12 family. The biotin operon had therefore been initially interrupted by the insertion of a REP13E12 element, which itself was then disrupted by the insertion of the phage.

The REP13E12 elements ranged in size from 1352 to 1436 bp and were present in seven copies throughout the genome (Table 5). Two elements, REP13E12 and REP336, contained a putative translational frame-shifting site, CCCCCCA, situated at the overlap of their two ORFs. The REP251 upstream ORF appeared truncated relative to the REP336 and REP13E12 elements due to a nonsense mutation. The elements did not show similarity to other sequences in the protein and nucleic acid databases. However, REP13E12 was identical to the 453 bp repeated sequence described by Lee *et al.* (1997) that is specific for the *M. tuberculosis* complex

Fig. 3. Dendrogram generated from multiple alignments of the DNA sequences of the REP13E12 family. The values at the nodes represent the values obtained by bootstrap analysis with 100 data sets.

and present in three or four copies in various strains (Lee *et al.*, 1997). This 453 bp repeat corresponded to an internal segment of the REP13E12 repeat and the difference in copy number observed experimentally by Lee *et al.* for H37Rv (three) and in the genome sequence (seven) is probably due to sequence variation, with four copies not detected under the stringent hybridization conditions used.

Fig. 4. (a) Distribution of IS, prophage and REP13E12 elements on the *M. tuberculosis* chromosome. *oriC* marks the origin of replication. The scale in megabases is marked inside the circle. The variable IS6110 in the *plcD* gene is shown in bold type and boxed. (b) Histogram showing the number of IS and REP13E12 (left-hand axis) present in each size of interval between the elements. The cumulative percentage of elements found in the intervals is also shown (♦, right-hand axis).

Fig. 5. Variable locus for the insertion of IS6110. ORFs are denoted either by their proper gene name or by their 'Rv' designation, with IS6110 shown in black. The interrupted *plcD* encodes a protein with similarity to phospholipase enzymes. The distance between the two insertions is 621 bp.

Alignment of the DNA sequences of the REP13E12 family revealed that they were very similar to each other, in particular REP165 with REP09F9 and REP336 with REP251 (Fig. 3). As REP336 serves as an attachment site for ϕ Rv1 it is possible that the prophage will be found in other members of the REP13E12 family in different strains of the tubercle bacillus. Data presented by Lee *et al.* (1997) suggest that this may be the case in the Erdman strain as it presents a different hybridization pattern from H37Rv when probed with the repeated sequence.

Distribution of IS elements in the H37Rv genome

From the physical and genetic map of *M. tuberculosis* H37Rv it was apparent that the distribution of IS6110 was not random throughout the genome, but that the copies were clustered (Philipp *et al.*, 1996). This is similar to the situation described by Boyd & Hartl (1997) for the distribution of IS1 on the chromosome of natural isolates of *Escherichia coli*, where one section of the chromosome lacks IS1 in 11 strains. The position of each of the new IS and REP13E12 elements on the genome is shown in Fig. 4(a), from which it appears that

insertions are selected against in the region close to *oriC*, with a 600 kb segment of the chromosome lacking even vestigial IS elements. Statistical analysis of the distribution of IS and REP13E12 elements on the genome revealed a mean spacing of 67.8 ± 56.6 kb (SEM). This distribution can be represented by a histogram (Fig. 4b) where the interval between each IS or REP13E12 (0–10 bp, 10–100 bp, 100–1000 bp, etc.) is plotted against the number of IS and REP13E12 elements found in each class. This further suggests that the interval of 600 kb represents a chromosomal region where insertions are detrimental. A transposon mutagenesis study by Bardarov *et al.* (1997) revealed a similar bias for the insertion of Tn5367 in the *M. tuberculosis* genome.

The arrangement of IS elements on the *M. tuberculosis* genome can be regarded as a 'natural' transposon mutagenesis experiment, identifying regions of the genome where the insertion of IS elements is selected against. Areas lacking IS elements may therefore contain genes vital for survival. The distribution also identified loci that are hot spots for insertion, such as at ~ 3.85 Mb, where IS1532, IS1540 and IS1552 are grouped together. Some clustering of insertion events near tRNA genes was seen, for example at ~ 2.97 Mb, where ϕ Rv2, IS6110, and IS1081' were inserted next to a group of four tRNA genes, or at ~ 4.08 Mb, where IS1553 and IS1534 were located close to *thrU*. Furthermore, both copies of IS1547 were located downstream of IS6110, suggesting linkage of the elements.

Cases where an ORF has been disrupted by the insertion of an IS element were apparently uncommon in the genome. *plcD* (Rv1755c), a phospholipase C homologue, and the adjacent Rv1758, a possible cutinase, were truncated by an IS6110 insertion, while IS6110 had also interrupted Rv2353c, a PPE protein (Cole *et al.*, 1998). IS1081 had inserted in the molybdopterin biosynthesis operon located at ~ 3.48 Mb, but it is not clear whether this has disrupted the operon. The rarity of insertional inactivation of genes suggests targeting of IS elements for insertional hot spots, as described for *Rhizobium* sp. (Perret *et al.*, 1997). In fact, the *plcD* locus may be one

Fig. 6. PCR analysis of *M. tuberculosis* complex for the presence of IS1532. Lane 1, *M. africanum*; lane 2, *M. bovis*; lane 3, *M. bovis* BCG Pasteur; lane 4, *M. microti* OV254; lane 5, *M. tuberculosis* CSU#93; lane 6, *M. tuberculosis* H37Ra; lane 7, *M. tuberculosis* H37Rv; lane 8, negative control.

such hot spot. Sequence data from cosmid clones, constructed from a different batch of *M. tuberculosis* genomic DNA to a bacterial artificial chromosome (BAC) and whole genome shotgun library (Philipp *et al.*, 1996; Brosch *et al.*, 1998), revealed a second copy of IS6110 at the *plcD* locus relative to the same position in the BAC or whole-genome-shotgun library (Fig. 5). This suggests that the locus may be preferentially targeted by IS6110, presumably as insertions into the already inactive *plcD* have no effect on the bacillus.

Distribution of IS elements in tubercle bacilli

To determine whether the substantial repertoire and distribution of IS and repetitive elements in *M. tuberculosis* H37Rv was maintained in other tubercle bacilli, *M. africanum*, *M. microti*, *M. bovis*, *M. bovis* BCG Pasteur, *M. tuberculosis* H37Ra, *M. tuberculosis* CSU#93 and 29 clinical isolates of *M. tuberculosis* were examined for the presence of the newly described elements.

This analysis revealed that IS1532 was absent from *M. bovis*, *M. bovis* BCG Pasteur, *M. tuberculosis* CSU#93, and 10 of the 29 clinical isolates tested (Fig. 6). The only other elements that were absent from some of the strains tested were IS1533 (absent from one of the clinical isolates tested), IS1534 (absent from three clinical isolates), and IS1561' (not found in *M. microti* OV254). These results are somewhat surprising, in that loci containing IS elements would be expected to show greater strain-to-strain differences (Arber, 1993). Indeed, IS variation is the basis for IS6110 typing of tubercle clinical isolates (Yeh *et al.*, 1998; Blazquez *et al.*, 1997), showing that *M. tuberculosis* can support the movement of IS elements through its genome. It may be significant that the three elements absent in some clinical isolates all belonged to the IS21 family, indicating that these elements may be the most active in the tubercle complex after IS6110. The absence of IS1561' from *M. microti* OV254 suggests that this locus may be useful for the differentiation of *M. microti* from the other members of the tubercle complex, although a greater number of strains will have to be tested to see if this holds true.

Conclusions

The new IS elements described here provide insights into the biology of the tubercle bacilli on a number of levels. Their spatial distribution on the chromosome identifies regions that can support the insertion of foreign DNA without disrupting the activity of vital genes, while regions lacking IS elements probably contain loci that must remain intact for the correct functioning of the organism. It will be interesting to compare the distribution of IS elements in more clinical isolates to see if some areas of the chromosome consistently lack these elements, hence allowing the identification of loci that may be critical for virulence. Horizontal transfer of elements into the tubercle complex is also evident, showing that *M. tuberculosis* can exchange genetic

information with other bacteria. However, it is not apparent if IS movement introduced novel genes into the bacillus.

ACKNOWLEDGEMENTS

Sincerest thanks to Thierry Garnier for help with statistical analysis, and to Alexander Pym and Roland Brosch for advice and discussion. We are particularly grateful to Michael Chandler for his interest and guidance. S.V.G. is a Wellcome Trust travelling research fellow. This work was supported by the Wellcome Trust, the Association Française Raoul Follereau, and the Institut Pasteur.

REFERENCES

- Arber, W. (1993). Evolution of prokaryotic genomes. *Gene* 135, 49–56.
- Bairoch, A., Bucher, P. & Hofmann, K. (1997). The PROSITE database, its status in 1997. *Nucleic Acids Res* 25, 217–221.
- Bardarov, S., Kriakov, J., Carriere, C., Yu, S., Vaamonde, C., McAdam, R., Bloom, B. R., Hatfull, G. F. & Jacobs, W. R., Jr (1997). Conditionally replicating mycobacteriophages: a system for transposon delivery to *Mycobacterium tuberculosis*. *Proc Natl Acad Sci USA* 94, 10961–10966.
- Blazquez, J., Espinosa de Los Monteros, L. E., Samper, S., Martin, C., Guerrero, A., Cobo, J., Van Embden, J., Baquero, F. & Gomez-Mampaso, E. (1997). Genetic characterization of multidrug-resistant *Mycobacterium bovis* strains from a hospital outbreak involving human immunodeficiency virus-positive patients. *J Clin Microbiol* 35, 1390–1393.
- Boyd, E. F. & Hartl, D. L. (1997). Nonrandom location of IS1 elements in the genomes of natural isolates of *Escherichia coli*. *Mol Biol Evol* 14, 725–732.
- Brosch, R., Gordon, S. V., Billault, A., Garnier, T., Eiglmeier, K., Soravito, C., Barrell, B. G. & Cole, S. T. (1998). Use of a *Mycobacterium tuberculosis* H37Rv bacterial artificial chromosome library for genome mapping, sequencing, and comparative genomics. *Infect Immun* 66, 2221–2229.
- Chandler, M. & Fayet, O. (1993). Translational frameshifting in the control of transposition in bacteria. *Mol Microbiol* 7, 497–503.
- Cole, S. T., Brosch, R., Parkhill, J. & 39 other authors (1998). Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* 393, 537–544.
- Collins, D. M. & Stephens, D. M. (1991). Identification of an insertion sequence, IS1081, in *Mycobacterium bovis*. *FEMS Microbiol Lett* 67, 11–15.
- Donadio, S. & Staver, M. J. (1993). IS1136, an insertion element in the erythromycin gene cluster of *Saccharopolyspora erythraea*. *Gene* 126, 147–151.
- Fang, Z. & Forbes, K. J. (1997). A *Mycobacterium tuberculosis* IS6110 preferential locus (ipl) for insertion into the genome. *J Clin Microbiol* 35, 479–481.
- Fang, Z., Morrison, N., Watt, B., Doig, C. & Forbes, K. J. (1998). IS6110 transposition and evolutionary scenario of the direct repeat locus in a group of closely related *Mycobacterium tuberculosis* strains. *J Bacteriol* 180, 2102–2109.
- Galas, D. J. & Chandler, M. (1989). Bacterial insertion sequences. In *Mobile DNA*, pp. 109–162. Edited by D. E. Berg & M. M. Howe. Washington, DC: American Society for Microbiology.

- Grzeszik, C., Lubbers, M., Reh, M. & Schlegel, H. G. (1997). Genes encoding the NAD-reducing hydrogenase of *Rhodococcus opacus* MR11. *Microbiology* **143**, 1271–1286.
- Hatfull, G. F. & Grindley, N. D. F. (1988). Resolvases and DNA-invertases: a family of enzymes active in site-specific recombination. In *Genetic Recombination*, pp. 357–396. Edited by R. Kucherlapati & G. R. Smith. Washington, DC: American Society for Microbiology.
- Hellyer, T. J., DesJardin, L. E., Assaf, M. K., Bates, J. H., Cave, M. D. & Eisenach, K. D. (1996). Specificity of IS6110-based amplification assays for *Mycobacterium tuberculosis* complex. *J Clin Microbiol* **34**, 2843–2846.
- Hermans, P. W., van Soolingen, D., Bik, E. M., de Haas, P. E., Dale, J. W. & van Embden, J. D. (1991). Insertion element IS987 from *Mycobacterium bovis* BCG is located in a hot-spot integration region for insertion elements in *Mycobacterium tuberculosis* complex strains. *Infect Immun* **59**, 2695–2705.
- Kato, K., Ohtsuki, K., Mitsuda, H., Yomo, T., Negoro, S. & Urabe, I. (1994). Insertion sequence IS6100 on plasmid pOAD2, which degrades nylon oligomers. *J Bacteriol* **176**, 1197–1200.
- Lee, T. Y., Lee, T. J., Belisle, J. T., Brennan, P. J. & Kim, S. K. (1997). A novel repeat sequence specific to *Mycobacterium tuberculosis* complex and its implications. *Tubercle Lung Dis* **78**, 13–19.
- Mahillon, J. & Chandler, M. (1998). Insertion sequences. *Microbiol Mol Biol Rev* **62**, 725–774.
- McHugh, T. D. & Gillespie, S. H. (1998). Nonrandom association of IS6110 and *Mycobacterium tuberculosis*: implications for molecular epidemiological studies. *J Clin Microbiol* **36**, 1410–1413.
- McHugh, T. D., Newport, L. E. & Gillespie, S. H. (1997). IS6110 homologs are present in multiple copies in mycobacteria other than tuberculosis-causing mycobacteria. *J Clin Microbiol* **35**, 1769–1771.
- Mariani, F., Piccolella, E., Colizzi, V., Rappuoli, R. & Gross, R. (1993). Characterization of an IS-like element from *Mycobacterium tuberculosis*. *J Gen Microbiol* **139**, 1767–1772.
- Ohtsubo, F. & Sekine, Y. (1996). Bacterial insertion sequences. *Curr Top Microbiol Immunol* **204**, 1–26.
- Olasz, F., Kiss, J., König, P., Buzàs, Z., Stalder, R. & Arber, W. (1998). Target specificity of insertion element IS30. *Mol Microbiol* **28**, 691–704.
- Perret, X., Viprey, V., Freiberg, C. & Broughton, W. J. (1997). Structure and evolution of NGRS-1, a complex, repeated element in the genome of *Rhizobium* sp. strain NGR234. *J Bacteriol* **179**, 7488–96.
- Philipp, W. J., Poulet, S., Eiglmeier, K. & 7 other authors (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.
- Picardeau, M., Bull, T. J. & Vincent, V. (1997). Identification and characterization of IS-like elements in *Mycobacterium goodii*. *FEMS Microbiol Lett* **154**, 95–102.
- Poulet, S. & Cole, S. T. (1995). Repeated DNA sequences in mycobacteria. *Arch Microbiol* **163**, 79–86.
- Small, P. M., Hopewell, P. C., Singh, S. P., Paz, A., Parsonnet, J., Ruston, D. C., Schecter, G. F., Daley, C. L. & Schoolnik, G. K. (1994). The epidemiology of tuberculosis in San Francisco. A population-based study using conventional and molecular methods. *N Engl J Med* **330**, 1703–1709.
- Sreevatsan, S., Pan, X., Stockbauer, K. E., Connell, N. D., Kreiswirth, B. N., Whittam, T. S. & Musser, J. M. (1997). Restricted structural gene polymorphism in the *Mycobacterium tuberculosis* complex indicates evolutionarily recent global dissemination. *Proc Natl Acad Sci USA* **94**, 9869–9874.
- Staden, R. (1996). The Staden sequence analysis package. *Mol Biotechnol* **5**, 233–241.
- Thierry, D., Cave, M. D., Eisenach, K. D., Crawford, J. T., Bates, J. H., Gicquel, B. & Guesdon, J. L. (1990). IS6110, an IS-like element of *Mycobacterium tuberculosis* complex. *Nucleic Acids Res* **18**, 188.
- Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22**, 4673–4680.
- Yang, Z., Barnes, P. F., Chaves, F., Eisenach, K. D., Weis, S. E., Bates, J. H. & Cave, M. D. (1998). Diversity of DNA fingerprints of *Mycobacterium tuberculosis* isolates in the United States. *J Clin Microbiol* **36**, 1003–1007.
- Yeh, R. W., Ponce de Leon, A., Agasino, C. B., Hahn, J. A., Daley, C. L., Hopewell, P. C. & Small, P. M. (1998). Stability of *Mycobacterium tuberculosis* DNA genotypes. *J Infect Dis* **177**, 1107–1111.

Received 19 August 1998; revised 19 November 1998; accepted 9 December 1998.