Lepr Rev (2001) **72,** 449–461

Repetitive sequences in *Mycobacterium leprae* and their impact on genome plasticity

STEWART T. COLE*, PHILIPPE SUPPLY** & NADINE HONORÉ*

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

***Institut National de la Sante et de la Recherche Medicale, U447, Institut Pasteur de Lille, 1 Rue du Professor Calmette, F-59019 Lille Cedex, France*

Summary About 2% of the genome of *Mycobacterium leprae* is composed of repetitive DNA. There are more than 26 extinct IS elements together with four families of dispersed repeats, present in five copies or more, RLEP (37 copies), REPLEP (15 copies), LEPREP (eight copies), and LEPRPT (five copies). Although there is no sequence similarity to known transposable elements, RLEP occurs predominantly at the 3′-end of genes and, in several cases, within pseudogenes, suggesting that it was capable of dissemination. Strikingly, on comparison of the genome sequences of *M. leprae* and the closely related tubercle bacillus, *Mycobacterium tuberculosis* H37Rv, many of these repetitive sequences were found at sites of discontinuity in gene order. Evidence is presented that loss of synteny, inversion and genome downsizing may have resulted from recombination between dispersed copies of these repetitive elements.

Introduction

Repetitive sequences are common constituents of the genomes of all living organisms although they are far more prominent in higher eukaryotes where they can account for a substantial percentage of the chromosomal DNA. There are two principal forms of repetitive DNA in bacterial genomes: dispersed and tandem repeats. Dispersed repetitive sequences can correspond to duplicated genes, or to mobile genetic elements present in several copies like insertion sequences (IS). IS are often an important component of bacterial genomes and as a result of their ability to transpose have mutational potential based on their ability to locate within coding or regulatory regions. Hundreds of individual IS have been described and grouped into 17 families on the basis of their genetic organization, sequence similarities in their recombinases/transposases, the similarity of their ends (direct or terminal inverted repeats) and their target sites which are often duplicated during transposition.[1] The genome sequence of *Mycobacterium tuberculosis* H37Rv contains more than 56 IS, belonging to eight

* Correspondence to: Prof. S. T. Cole, Unité de Génétique Moléculaire Bactérienne, Institut Pasteur, 28 Rue du Dr Roux, 75724 Paris Cedex 15, France. Phone: 33-1-45 68 84 46. Fax: 33-1-40 61 35 83. E-mail: stcole@pasteur.fr

0305-7518/01/072449+13 $1.00

families and these are an important source of plasticity and genetic variability.[2–6] A novel repeated sequence, the REP13E12 family, is present in seven copies on the chromosome and contains a probable phage attachment site.[3] A large portion of the genome has also evolved from gene duplication events, followed by sequence divergence, leading to functional redundancy and expansion of the biological potential of the tubercle bacillus.[7]

Tandem repeats can be relatively simple, such as multiple repetitions of di- or tri-nucleotide sequences, or more complex such as the tandem duplication of large chromosomal segments,[8] like those described in *Mycobacterium bovis* BCG Pasteur.[9] Genetic variation is commonly associated with di- or tri-nucleotide repeats which are prone to amplification and contraction. These are often referred to as micro- or mini-satellites and are useful for typing purposes. One such mini-satellite that has been described in *M. tuberculosis* is the mycobacterial interspersed repetitive unit (MIRU), and this is also found in *M. leprae*.[10] A very promising epidemiological tool for tubercle bacilli has been developed that is based on variable number tandem repeats (VNTRs) of MIRU and this is capable of efficiently discriminating between outbreak strains.[11,12]

Here we describe the complete repertoire of repetitive DNA sequences identified in the genome of *M. leprae*[13] and discuss their potential impact on the evolution of the organism. In addition, attempts are being made to exploit some of these sequences for the development of a test that can distinguish between isolates of the leprosy bacillus.

Materials and methods

To identify repetitive DNA, the BLASTN program[14,15] was used to compare the genome sequence with itself. Areas showing >99% identity were then inspected visually and annotated using Artemis.[16] Potential IS elements were uncovered by database searches using BLASTX and tandem repeats identified using the program tandem repeats finder.[17] MIRUs were localized by BLASTN searches of the genome sequence using the consensus sequences of MIRU1-3 as strings, and all hits with scores >70 were investigated using a combination of Artemis the relational database, Lepromа.[18]

To investigate MIRU-based polymorphism, PCR primers were designed using the Oligo 5.0 software (National Biosciences, Plymouth, MN, USA), and the sequences are summarized in Table 1. *M. leprae* DNA was prepared by the freeze-boiling method.[19] For PCR reactions, 5 μl of DNA solutions was added to a final volume of 25 μl containing 10% DMSO, 0·5 mmol/l of each dATP, dCTP, dGTP and dTTP, 0·2 μmol/l of primers, 2·5 μl of PCR buffer [170 mmol/l $(NH4)_2SO_4$, 600 mmol/l Tris-HCL (pH 8·8), 20 mmol/l $MgCl_2$, 100 mmol/l β-mercaptoethanol] and 1·25 IU of *Taq* polymerase (Gibco-BRL). The PCR was performed using a PTC-100 (MJ Research, Inc.) for 35 cycles of 1 min at 94°C, 2 min at 59°C, 2 min at 72°C. The reactions were terminated by incubating for 10 min at 72°C and analysed by agarose gel electrophoresis using the appropriate controls.

Results and discussion

DIRECT REPEATS

On examination of the *M. leprae* genome sequence a series of perfect direct repeats was found ranging in size from 2 to 52 bp. All repeat sequences of >20 bp that were present in

Table 1. PCR primers used to study MIRU diversity

PRIMER F	SEQUENCE	PRIMER R	SEQUENCE	GENES
B937-MIRU1F	GTGCTGACCCGCTATCCTGA	B937-MIRU1R	CCCGCGACCCAGATTCTATC	*ML0534...carA*
B1308-MIRU1F	CGTTCTTGTGTGCGGGTGAGT	B1308-MIRU1R	TTACGACGCTGTTATGGAAACTGC	*ML0719...aldB*
B2235-MIRU1F	GCTGCGCCCGCGGTAGTCAC	B2235-MIRU1R	GAGGGGATGCCGACCATTTGG	*miaA...dapF*
B1764-MIRU1F	GGGCTTTCCATCGTCAACAG	B1764-MIRU1R	GCGTTAGGCACCCCAACA	*dut...ML1029*
B1764-MIRU2F	TTCACGCGAGTCCAGGTCAGAC	B1764-MIRU2R	CGTACGCAGGGAGGAGCAAAAC	*ML1042...ML1041*
L471-MIRU1F	CCAGGAGCCCACCAGGAC	L471-MIRU1R	GACGGCTGACATTGTCGGTCTAG	*ML1135...ML1136*
B1549-MIRU2F	TACCAGGAGCGGGATCGTAT	B1549-MIRU2R	CGGACGTGCTGACCATC	*cysM...ML1171*
B1549-MIRU1F	GTTCAGCGATACCAGCGTCA	B1549-MIRU1R	TCAGGGGACTGGTGAGGG	*rphA...ML1175*
B1133-MIRU1F	TGACGCTGGGTTTTTGGT	B1133-MIRU1R	GTTCGCGTGGAGTTCTTGTC	*argD...argF*
B2266-MIRU2F	GGAGGGAACTGGCAAGTCGT	B2266-MIRU2R	TGACCAGCCCAACAGACCTG	*ML2199...ML2200*
B2266-MIRU1F	TGAGCGGTCCACTAGCACAG	B2266-MIRU1R	CCGTCCAACGCGACTATCAC	*pabCML2203*
B2168-MIRU2F	CGCGGGTGGCTCGTAGAAGA	B2168-MIRU2R	TGACCGGCAAGCGACTTTGG	*ML2412...ML2413*
B2168-MIRU3F	TGGGCTCAAAACCTCCTTGC	B2168-MIRU3R	GGGCTGGCCATCGTCAAAC	*ML2439...ML2440*
B2168-MIRU1F	GGATGGCGTTGGTCTTGAG	B2168-MIRU1R	GCACTTTGGTGTTCGGACAT	*ML2442...ML2443*

two or more identical copies were annotated together with all perfect repeats present in three or more copies. There are far fewer tandem repeats in the genome of *M. leprae* compared to that of *M. tuberculosis*, mainly as a result of the much smaller number of PE-PGRS genes that are composed of such motifs. No tetranucleotide repeats are found in *M. leprae* nor in *M. tuberculosis* and dinucleotide repeats were only observed in the leprosy bacillus. A trinucleotide repeat (TTC) displaying copy number differences has been described in some isolates of the leprosy bacillus,[20] and this has 21 repetitions in the TN strain used for genome sequencing, while a hexanucleotide repeat in the *sigA* (*rpoT*) gene has been found recently to be present in three copies in most strains of *M. leprae*, including TN, but in four copies in others.[21] Promising discriminatory tests have been devised that target these polymorphisms,[20,21] and this encourages us to examine di- and trinucleotide repeats for variability in different isolates of *M. leprae*.

MYCOBACTERIAL INTERSPERSED REPETITIVE UNITS (MIRUS)

Prominent among the tandem repeats found in *M. tuberculosis*, are the MIRUs, as these can occur in from two to four tandem copies ranging in size from 46 to 101 bp, and are present at 41 loci.[10,12] MIRUs generally occur in intergenic regions and have the potential to encode small peptides as they contain short open reading frames whose start overlaps the stop codon of the upstream gene whereas the stop codon overlaps the iniation codon of the following gene. No MIRUs were detected by tandem repeat finder, but 20 single copies were found in the *M. leprae* genome by BLAST (Table 2). Eleven of these MIRUs had no counterparts in the *M. tuberculosis* genome, whereas four of the conserved loci contained multiple MIRUs in *M. tuberculosis* but only one in *M. leprae*.

To determine whether any of the sites harbouring MIRUs were occupied by multiple copies in other *M. leprae* strains, PCR primers were designed for 14 loci and used to screen a panel of 14 different isolates from Mali, Martinique, New Caledonia and the Philippines for diversity. In all cases, the size of the PCR fragments was consistent with the presence of a single MIRU, and when the DNA sequence was determined this was found to be identical to that of the TN strain. These results indicate that MIRUs are unlikely to represent a source of polymorphism in the leprosy bacillus, in contrast to the situation in *M. tuberculosis*.[11]

INSERTION SEQUENCES AND DUPLICATED GENES

Unlike *M. tuberculosis* H37Rv, which contains 56 IS elements, most of which are predicted to be functional,[3] *M. leprae* has only vestigial IS elements, as >26 transposase gene fragments were identified. These could not be classified reliably owing to the extensive levels of mutation and truncation incurred. All of these sequences appear in single copies.

Two identical copies of 16 dispersed repeats of >700 bp were detected and examined (Table 3). Three of these (1329, 1261 and 1179 bp) probably correspond to extinct IS, although it is interesting to note that despite loss of function their sequences are perfectly conserved, whereas two others (1054 and 753 bp) appear to be counterparts of the REP13E12 repeats described in the tubercle bacilli.[3] The remaining duplicated sequences correspond to genes or more rarely to pseudogenes (Table 3), and again it is unusual to find perfect conservation of the sequence in bacteria. This suggests that these duplication events may have occurred very recently or that sequence divergence occurs at an exceptionally slow rate in *M. leprae*.

Table 2. Features of MIRU in *M. leprae* and comparison with *M. tuberculosis*

Class	Position	Bases	Genes*	*M. tuberculosis*	Class (No.)°	Comments
MIRU2	153610...153666	57	*rfbe*...ML0112	*rfbE*...3781	–	
MIRU2	534710...534766	57	*scoA*...*scoB*	*scoA*...*scoB*	–	In-frame stop
MIRU1	648342...648425	84	ML0534...*carA*	Rv1382...*carA*	3 (2)	
MIRU1	862062...862114	>53	ML0719...*aldB*	Rv3292...*aldB*	–	Degenerate
MIRU3	877171...877223	53	*purK*...*purE*	*purK*...*purE*	3	Embedded in purK
MIRU2	1164265...1164320	57	*miaA*...*dapF*	*miA*...*dapF*	–	No ATG
MIRU1	1194835...1194886	52	*dut*...ML1029	*dut*...Rv2696c	1	Out-of-frame
MIRU2	1207116...1207172	57	ML1042...ML1041	Rv2680...*echA15*	2 (3)	–
MIRU2	1327764...1327823	60	ML1135...ML1136	Rv1300...Rv1301	2 (3)	–
MIRU1	1368453...1368519	67	*cysM*...ML1171	*cysM*...Rv1337	–	In-frame with ML1171
MIRU1	1371749...1371830	83	*rphA*...ML1175	*rphA*...Rv1341	2	Out-of-frame
MIRU2	1629692...1629746	55	ML1368...ML1369	Rv1709...Rv1710	2	Out-of-frame
MIRU2	1777816...1777872	57	ML1476...ML1475	Rv2454c...Rv2455c	–	–
MIRU1	1692100...1692180	81	*argD*...*argF*	*argD*...*argF*	–	
MIRU2	1915421...1915483	66	*pyrH*...*frr*	*pyrH*...*frr*	2 (2)	No ATG
MIRU2	2044283...2044339	57	*ilvb*...*ilvN*	*ilvB*...*ilvN*	–	No ATG
MIRU1	2613356...2613460	105	ML2199...ML2200	Rv0813c...Rv0814c	–	
MIRU1	2617793...2617873	81	*pabC*...ML2203	*pabC*...Rv0811c	–	–
MIRU2	2884042...2884095	54	ML2412...ML2413	Rv0525...Rv0526	2	No ATG
MIRU1	2917923...2918003	81	ML2442...ML2443	Rv0486...Rv0487	–	–

* Underlining indicates pseudogenes.
° –, denotes MIRU absent.

Table 3. Identical duplicated genes and sequences of >700 bp

Sequence	Genes*	Description
1329 bp	ML0040	Possible transposase remnant
1261 bp	ML1749	Possible transposase remnant
1179 bp	ML0444	Pseudogene similar to group II intron maturase
1054 bp	ML1290/ML1850	Pseudogenes orthologous to REP13E12 proteins
753 bp	ML1118/ML2236	Pseudogenes orthologous to REP13E12 proteins
1551 bp	*ansP1/ansP2*	L-asparagine transport proteins
1391 bp	ML2356/ML2357	Part of polyketide synthase
1219 bp	*mmpL2/mmpL4*	RND family transport proteins
1186 bp	ML0396/ML2692	Myo-inositol-1-phosphate synthase
1063 bp	ML1053/ML1183	PE proteins
1063 bp	ML1054/ML1182	PPE proteins
1006 bp	ML0125/ML0128	Putative glycosyl transferase
879 bp	ML1055/ML1180	QILSS family
879 bp	ML1056/ML1181	ESAT-6 family
871 bp	ML1047/ML1943	Pseudogene orthologous to Rv3714c
847 bp	*fadD5/fadD5*	Acyl-CoA synthase pseudogenes
740 bp	ML0447/ML2159	Similar to region of cytochrome P450s
704 bp	*umaA2/umaA1*	Mycolic acid synthase and pseudogene

* Underlining indicates pseudogenes.

Of particular interest are two regions of 1063 and 879 bp as these encode proteins of the PE, PPE, and ESAT-6 families.[2,7,22] In *M. tuberculosis* there are 11 regions containing ESAT-6 genes and these show two configurations comprising blocks of four or 10 conserved genes. There are three ESAT-6 regions of the larger type in *M. leprae* and two blocks of four genes. The latter consist of two identical repeat sequences of 1063 and 879 bp (Table 2; Figure 1). Interestingly, in one of these ESAT-6 regions, an additional unique sequence of 619 bp is present within the ML1182 gene, encoding a PPE protein, but one cannot tell whether this has been acquired by ML1182 or lost from ML1054. This is further evidence indicating that the ESAT-6 regions are dynamic[7,23] and that PPE proteins can undergo variation.[2,24]

RLEP

The RLEP element was intitially detected as a repetitive sequence in *M. leprae* by means of Southern blotting[25,26] and subsequently characterized at the molecular level by Woods

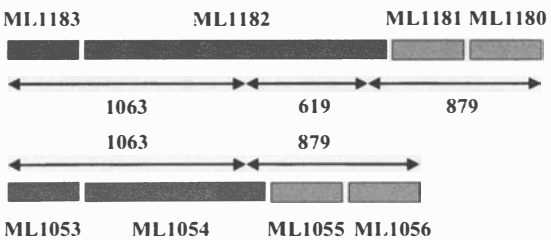

Figure 1. Organization of repeated loci encoding PE, PPE, and ESAT-6 proteins. Gene names are given and repeat sizes indicated in bp.

et al.[27] These authors estimated that there were at least 28 copies of RLEP in the genome and demonstrated that there was a central portion, common to all copies of RLEP, flanked by additional sequences whose presence was variable. With the complete genome sequence at our disposal we were able to perform the definitive bioinformatic analysis and this revealed that the TN strain of *M. leprae* contains 37 copies of RLEP, one of which, RLEP_29, lacks part of the central domain and will not be discussed further here. There is a conserved segment of 488 bp found in all intact copies of RLEP and this is flanked by additional sequences present in two or more independent RLEP elements. Consequently, the total length can vary from 601 to 1075 bp (Figure 2) and, as described previously, no open reading frames capable of coding for transposases, resolvases or other IS-associated functions could be found.[27] Further comparisons uncovered six polymorphic sites in the 488 bp conserved segment, three of which occurred only once while the remainder were found in numerous copies of RLEP. All of these polymorphisms can be accounted for by C-T transitions. On construction of a tree of RLEP sequences (Figure 2) by phylogenetic analysis using parsimony routines (PAUP), three large branches were established together with several outliers. However, there was little clear association between the length of the RLEP element or the presence of particular polymorphic nucleotides and its position in the tree (Figure 2). This is consistent with the complex organization of these sequences.

Roughly 1% of the chromosome is composed of RLEP DNA and these elements are distributed fairly randomly.[13] It is clear, however, that RLEP has contributed extensively to the remodelling of the *M. leprae* genome as copies are often found at breaks in synteny with *M. tuberculosis*. This will be discussed further below. There is a marked over-representation of RLEP elements, in either orientation, at the 3′-ends of genes, since ~30 of the copies are within 80 bp of the stop codon of the nearest gene. In several instances, RLEP is situated within the coding sequence at the 3′-end or overlapping the stop codon. Examples of this may be found in the *truA*, *truB* and *polA* genes. In two cases, important genes such as *glnA* and *polA* are flanked by inverted pairs of RLEPs in a configuration resembling that of a transposon. Data have been published that show that this composite *polA* structure is polymorphic between isolates of *M. leprae*[28] and it is conceivable that similar variability may also be associated with *glnA*. Transcription of *polA* may also have been impaired by RLEP,[29] as RLEP_22 is situated 6 bp upstream of the *polA* initiation codon. Some copies of RLEP are found within the sequences of pseudogenes that have intact functional orthologues in *M. tuberculosis*, notable examples are RLEP_8 in the spermidine biosynthetic gene, *speE*, RLEP_29 in the phosphoglucomutase gene, *pgmA*, and RLEP_28 in ML1722, a pseudogene orthologous to Rv3037c, a conserved hypothetical gene of *M. tuberculosis*. These observations suggest that RLEP may have been capable of transposition at one time although it is quite unclear how this was mediated. Furthermore, attempts to detect restriction fragment length polymorphisms linked to RLEP have revealed no diversity suggesting that RLEP is no longer capable of movement.[30]

REPLEP

There are 13 essentially intact sequences belonging to the REPLEP family and two large fragments (Figure 3). The largest elements are 881 bp long, with extensive complementarity between bases 1–95 and 783–880 (68% identity). REPLEP is bounded in most cases by an 8 bp inverted repeat (5′-GTTGTGGG) and contains no open reading frames. In several cases the inverted repeats continue past this octamer with certain REPLEP elements

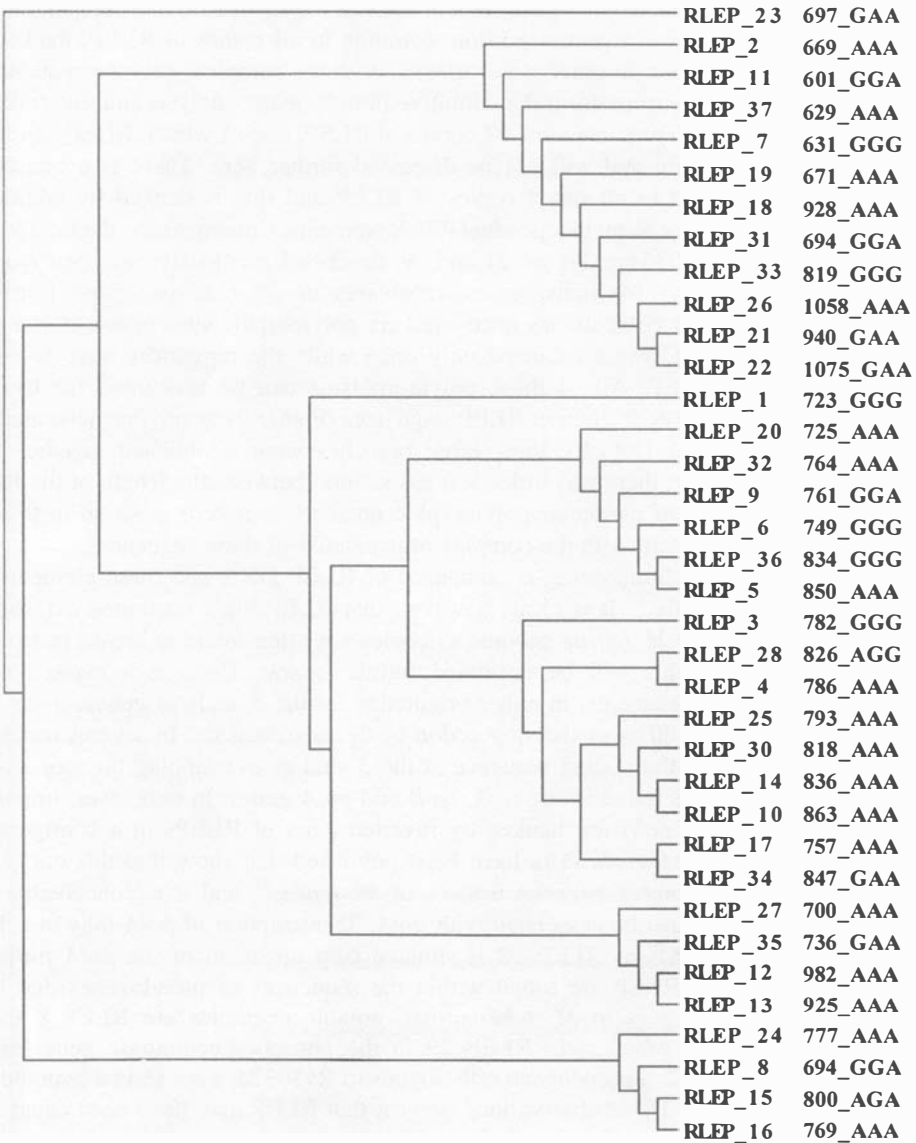

Figure 2. Phylogenetic tree of RLEP elements established using the phylogenetic analysis using parsimony routine of the GCG package. RLEP identifiers, sequence lengths and the sequence present at the concatenated polymorphic sites are indicated on the right. The sequence gaa signifies the presence of G, A and A at positions 501, 583 and 592 in the multiple alignment.

displaying a 45 bp sequence, or subsequence thereof, at the 3′-end that is also wholly or partially present at the 5′-end of some copies (Figure 3). With the exception of a single site, where six copies of REPLEP have a GGG tract whereas the remaining nine have GG, the nucleotide sequences are identical. No information as to a possible function is available.

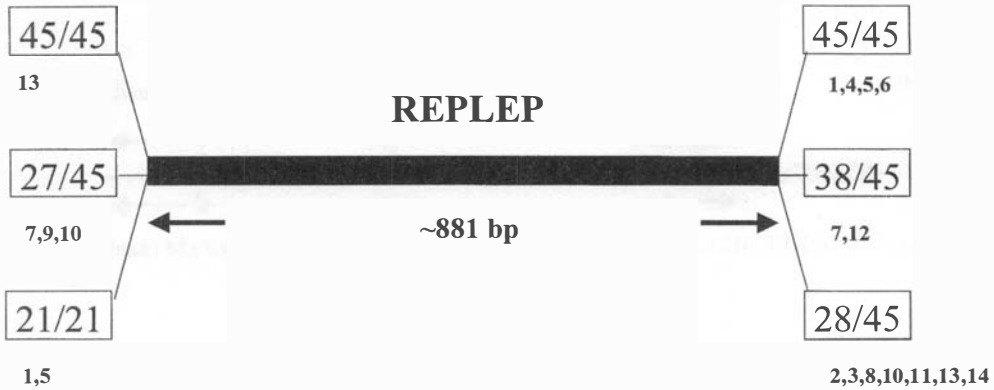

Figure 3. Schematic organization of REPLEP elements. The variable segments associated with the ends are shown as boxes with the numbers corresponding to the length in bp; 27/45 denote that 27 bp of the 45 bp segment are present. Identifiers for each REPLEP are shown below the boxes.

LEPREP

There are five intact sequences belonging to the LEPREP family and three fragments (Figure 4). These display near identical sequences with only three base differences being detected, two C–T transitions in REPLEP5 and one in REPLEP3. Unlike RLEP, REPLEP and LEPRPT, LEPREP displays a number of features commonly associated with IS elements and most probably corresponds to a degenerate version. The complete LEPREP sequence is 2383 bp long, contains a 54 bp palindromic inverted repeat and has a 6 bp inverted repeat (5′-CTAGTG) at its ends. Although there are no open reading frames that could code for

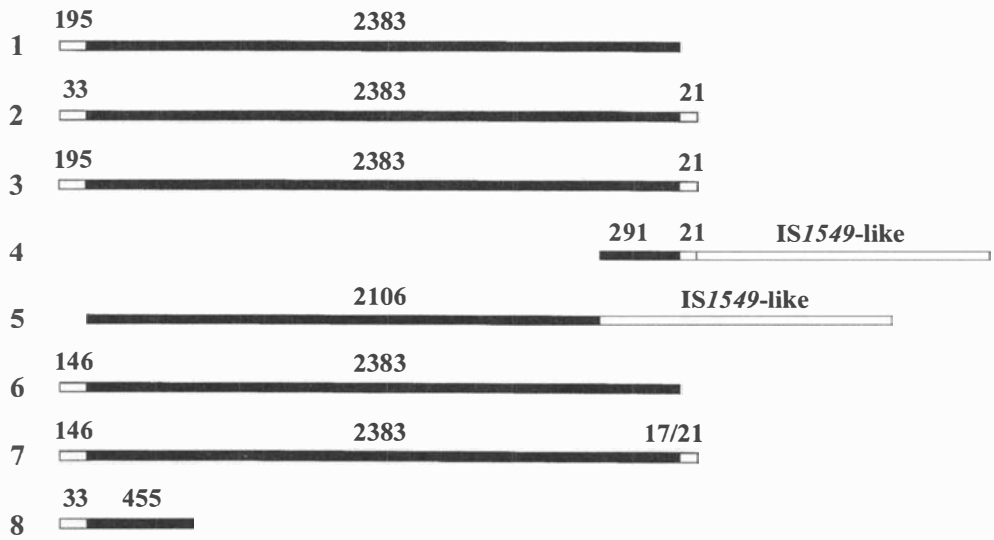

Figure 4. Schematic organization of LEPREP elements. The length in bp of each element is indicated. The variable segments associated with the ends are shown as stippled boxes with the numbers corresponding to the length in bp; 17/21 denotes that 17 bp of the 21 bp segment are present. Identifiers for each REPLEP are shown to the left.

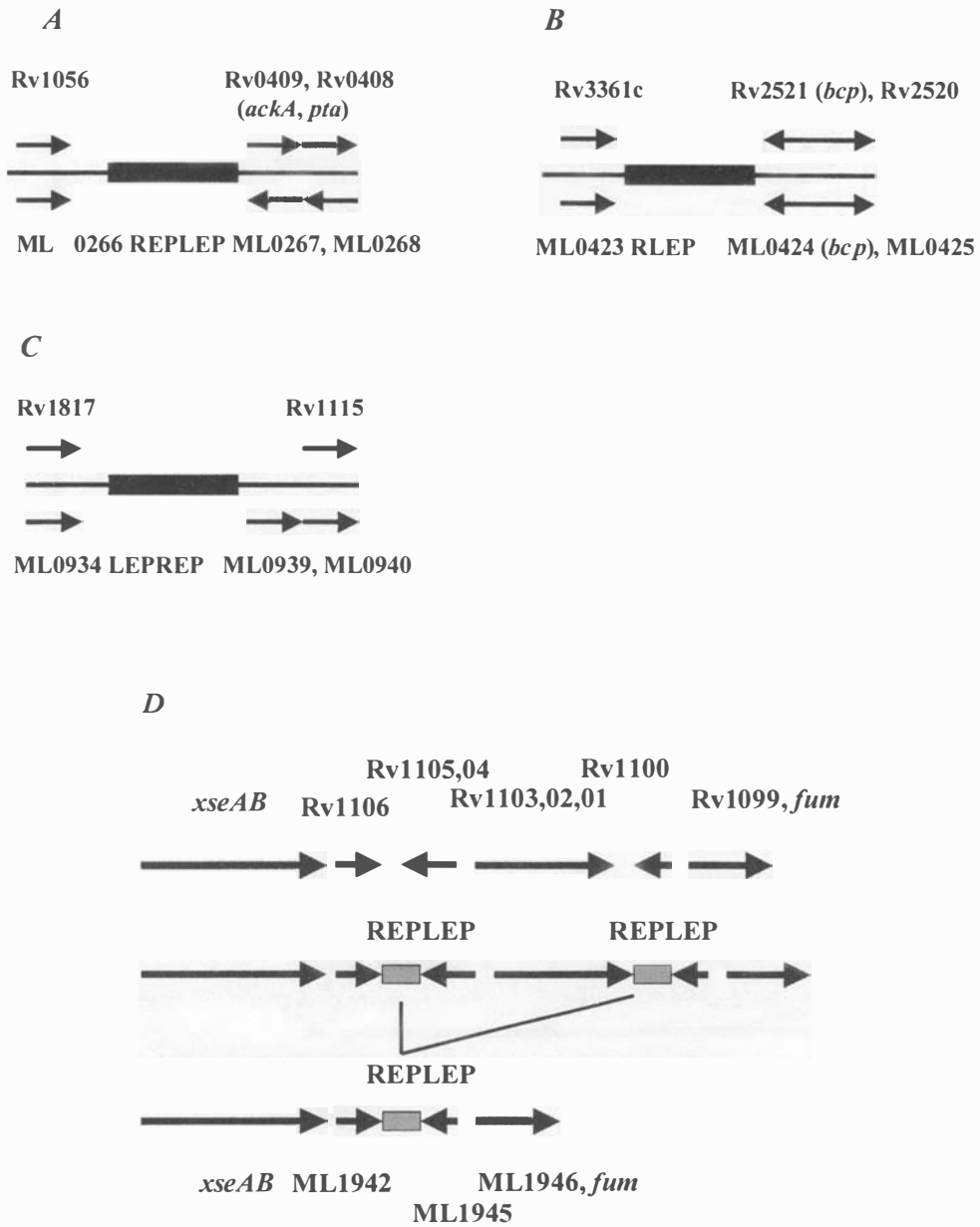

Figure 5. Repetitive elements and genome discontinuities. The three main repetitive elements in the *M. leprae* genome are shown together with examples of flanking genes and their counterparts in *M. tuberculosis*. The *M. tuberculosis* genes are designated with Rv prefixes, and the *M. leprae* genes with ML prefixes. Note the breaks in continuity of number of *M. tuberculosis* genes that indicate translocation event. **A.** REPLEP, 12 complete copies of ~875 bp, plus 2 fragments; **B.** RLEP, 36 complete copies of 545–700 bp plus 1 fragment; **C.** LEPREP, 6 complete copies of 2,400 bp plus 3 fragments. **D.** Example of gene loss by deletion following transposition of a REPLEP element and homologous recombination between the two copies. The gene organization in *M. tuberculosis* H37Rv is shown at the top followed by the corresponding region in an ancestor of *M. leprae*. The present situation in the TN strain of *M. leprae* is shown at the bottom.

functional proteins, BLASTX searches revealed extensive sequence similarity to parts of transposases from *Pseudomonas putida* (EMBL:AJ245436) and *Agrobacterium tumefaciens* (EMBL:Z18270), and to putative group II intron maturase-related proteins such as that of the fungus, *Cryphonectria parasitica* (EMBL:AF218567). Copies 4, 5 and 8 of LEPREP have been truncated and in two cases this appears to have resulted from the insertion of another IS element of 1261 bp that is now degenerate but still shows extensive similarity to IS*1549* from *Mycobacterium smegmatis*. Copies 4 and 5 are truncated at their 3′- and 5′-ends, respectively, share a 14 bp residual sequence and are followed by the IS*1549*-like element (Figure 4). They may once have comprised part of the same LEPREP element. Four copies of LEPREP are followed by the same 21 (or 17) bp sequence whereas two copies are preceded by common sequences of 195, 146 and 33 bp, respectively. It is conceivable that these sequences represent preferential sites of insertion for LEPREP.

LEPRPT

There are five sequences belonging to the LEPRPT family. Copies 1, 2 and 4 are 1252–1254 bp in length whereas copies 3 and 5 appear to have been shortened, as they only comprise 707 and 533 bp. The sequences of the LEPRPT elements are identical and contain no significant open reading frames. Copies 4 and 5 are preceded by the same 51 bp segment while copies 2 and 3 both have identical 7 bp sequences at their 5′-ends. Although the size of LEPRPT is consistent with that of an IS element, there is no other evidence to this effect.

REMODELLING THE GENOME

When whole genome comparisons of the tubercle and leprosy bacilli were performed it became apparent that there were ~65 conserved chromosomal segments common to both bacteria with loss of gene synteny occurring at sites occupied by repetitive elements in most cases.[13] This is illustrated in Figure 5, where one can see that gene order changes abruptly at sites harbouring RLEP, REPLEP and LEPREP. It is probable that this resulted from recombination events between dispersed repetitive sequences of the same family. If the elements were arranged in inverted orientation this would result in displacement and inversion of segments of the chromosome whereas recombination events between directly oriented repeats would result in deletion of the intervening segment. A potential example of this is shown in Figure 5D.

To conclude, it is likely that chromosomal rearrangements, gene deletions and duplications have had a profound effect on the biology of *M. leprae* and in turn on leprosy itself. One of the major forces that shaped this process was undoubtedly the dispersion of repetitive DNA, which may have been catalysed by enzymes encoded by the elements, followed by homologous recombination between these dispersed repeats effected by recombinases such as RecA. Characterization of the residual repetitive sequences has helped us to understand the past of the leprosy bacillus and may provide us with new tools to track its dissemination in the future.

Acknowledgements

We thank Edouard Yeramian for help with repeat analysis. STC wishes to acknowledge the financial support of the Institut Pasteur, the Association Française Raoul Follereau, the

Programme de Recherche Fondamentale en Microbiologie et Maladies Infectieuses, and the National Institutes of Health, National Institute of Allergy and Infectious Diseases (grant no. 1 RO1 AI47197-01A1) for parts of this work.

References

[1] Mahillon J, Chandler M. Insertion sequences. *Microbiol Mol Biol Rev*, 1998; **62:** 725–774.

[2] Cole ST, Brosch R, Parkhill J *et al.* Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature*, 1998; **393:** 537–544.

[3] Gordon SV, Heym B, Parkhill J *et al.* New insertion sequences and a novel repeated sequence in the genome of *Mycobacterium tuberculosis* H37Rv. *Microbiology*, 1999; **145:** 881–892.

[4] Brosch R, Gordon SV, Eiglmeier K *et al.* Genomics, biology, and evolution of the *Mycobacterium tuberculosis* complex. In: Hatfull GF, Jacobs WR Jr (eds) *Molecular genetics of mycobacteria.* ASM Press, Washington D.C., 2000, pp. 19–36.

[5] Brosch R, Philipp W, Stavropolous E *et al.* Genomic analysis reveals variation between *Mycobacterium tuberculosis* H37Rv and the attenuated *M. tuberculosis* H37Ra. *Infect Immun*, 1999; **67:** 5768–5774.

[6] Fang Z, Doig C, Kenna DT *et al.* IS6110-mediated deletions of wild-type chromosomes of *Mycobacterium tuberculosis*. *J Bacteriol*, 1999; **181:** 1014–1020.

[7] Tekaia F, Gordon SV, Garnier T *et al.* Analysis of the proteome of *Mycobacterium tuberculosis* in silico. *Tubercle Lung Dis*, 1999; **79:** 329–342.

[8] Lupski JR, Roth JR, Weinstock GM. Chromosomal duplications in bacteria, fruit flies, and humans. *Am J Hum Genet*, 1996; **58:** 21–27.

[9] Brosch R, Gordon SV, Buchrieser C *et al.* Comparative genomics uncovers tandem chromosomal duplications in some strains of *Mycobacterium bovis* BCG: implications for vaccination. *Comp Funct Genom (Yeast)*, 2000; **17:** 111–123.

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

[11] Mazars E, Lesjean S, Banuls AL *et al.* High-resolution minisatellite-based typing as a portable approach to global analysis of *Mycobacterium tuberculosis* molecular epidemiology. *Proc Natl Acad Sci USA*, 2001; **13:** 1901–1906.

[12] Supply P, Mazars E, Lesjean S *et al.* Variable human minisatellite-like regions in the *Mycobacterium tuberculosis* genome. *Mol Microbiol*, 2000; **36:** 762–771.

[13] Cole ST, Eiglmeier K, Parkhill J *et al.* Massive gene decay in the leprosy bacillus. *Nature*, 2001; **409:** 1007–1011.

[14] Altschul SF, Boguski MS, Gish W, Wooton JC. Issues in searching molecular sequence databases. *Nature Genet*, 1994; **6:** 119–129.

[15] Altschul S, Gish W, Miller W *et al.* A basic local alignment search tool. *J Mol Biol*, 1990; **215:** 403–410.

[16] Rutherford K, Parkhill J, Crook J *et al.* Artemis: sequence visualization and annotation. *Bioinformatics*, 2000; in press.

[17] Benson G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res*, 1999; **27:** 573–580.

[18] Jones LM, Cole ST, Moszer I. Leproma: a *Mycobacterium leprae* genome browser. *Lepr Rev*, 2001; **72:** 470–477.

[19] Woods SA, Cole ST. A rapid method for the detection of potentially viable *Mycobacterium leprae* in human biopsies: a novel application of PCR. *FEMS Microbiol Lett*, 1989; **65:** 305–310.

[20] Shin YC, Lee H, Walsh GP *et al.* Variable numbers of TTC repeats in *Mycobacterium leprae* DNA from leprosy patients and use in strain differentiation. *J Clin Microbiol*, 2000; **38:** 4535–4538.

[21] Matsuoka M, Maeda S, Kai M *et al. Mycobacterium leprae* typing by genomic diversity and global distribution of genotypes. *Int J Lepr Other Mycobact Dis*, 2000; **68:** 121–128.

[22] Berthet FX, Rasmussen PB, Rosenkrandt I *et al.* A *Mycobacterium tuberculosis* operon encoding ESAT-6 and a novel low-molecular-mass culture filtrate protein (CFP-10). *Microbiology*, 1998; **144:** 3195–3203.

[23] Gordon SV, Brosch R, Billault A *et al.* Identification of variable regions in the genomes of tubercle bacilli using bacterial artificial chromosome arrays. *Mol Microbiol*, 1999; **32:** 643–656.

[24] Gordon SV, Eiglmeier K, Garnier T *et al.* Genomics of *Mycobacterium bovis. Tubercle Lung Dis*, 2001; **6:** in press.

[25] Clark-Curtiss JE, Docherty MA. A species-specific repetitive sequence in *Mycobacterium leprae* DNA. *J Infect Dis*, 1989; **159:** 7–15.

[26] Grosskinsky CM, Jacobs Jr WR, Clark-Curtiss JE, Bloom BR. Genetic relationships among *Mycobacterium leprae*, *Mycobacterium tuberculosis*, and candidate leprosy vaccine strains determined by DNA hybridization: identification of an *M. leprae* specific repetitive sequence. *Infect Immun*, 1989; **57:** 1535–1541.

[27] Woods SA, Cole ST. A family of dispersed repeats in *Mycobacterium leprae*. *Mol Microbiol*, 1990; **4:** 1745–1751.

[28] Fsihi H, Cole ST. The *Mycobacterium leprae* genome: systematic sequence analysis identifies key catabolic enzymes, ATP-dependent transport systems and a novel polA locus associated with genomic variability. *Mol Microbiol*, 1995; **16:** 909–919.

[29] Gordhan BG, Mizrahi V. The RLEP-flanked polA gene from *Mycobacterium leprae* is not transcribed in Mycobacterium smegmatis. *Gene*, 1997; **187:** 63–66.

[30] Williams DL, Gillis TP, Portaels F. Geographically distinct isolates of *Mycobacterium leprae* exhibit no genotypic diversity by restriction fragment-length polymorphism analysis. *Mol Microbiol*, 1990; **4:** 1653–1659.