# Genomic Deletions Suggest a Phylogeny for the *Mycobacterium tuberculosis* Complex

**Serge Mostowy,[1] Debby Cousins,[2] Jacqui Brinkman,[1,a] Alicia Aranaz,[3] and Marcel A. Behr[1]**

*[1]McGill University Health Centre, Montreal, Quebec, Canada; [2]Australian Reference Laboratory for Bovine Tuberculosis, Department of Agriculture, South Perth, Australia; [3]Departamento de Patología Animal I (Sanidad Animal), Facultad de Veterinaria, Universidad Complutense de Madrid, Madrid, Spain*

**To better understand the evolution of the *Mycobacterium tuberculosis* complex, subspecies were tested for large sequence polymorphisms. Samples with greater numbers of deletions, without exception, were missing all the same regions that were deleted from samples with lesser numbers of deletions. Principal genetic groups based on single-nucleotide polymorphisms were restricted to one of the deletion-based groups, and isolates that shared genotypes based on molecular epidemiological markers were assigned almost exclusively to the same deletion type. The data provide compelling evidence that human tuberculosis did not originate from the present-day bovine form. Genomic deletions present themselves as an attractive modality to study the evolution of the *M. tuberculosis* complex.**

The *Mycobacterium tuberculosis* complex is composed of the closely related subspecies *M. tuberculosis, M. bovis, M. microti,* and *M. africanum*. In addition, strains not conforming to these classic subspecies have been isolated from other mammals, such as *M. tuberculosis* subspecies *caprae* (hereafter referred to as *"M. caprae"*) [1] and the so-called "seal bacillus" [2]. Although the subspecies of the *M. tuberculosis* complex show >99% DNA identity [3], they differ in their host range, epidemiology, clinical presentation in humans, and laboratory phenotype. However, close genetic relatedness, overlapping phenotypes, and slow growth make species assignment difficult [4]. As a result, little is understood about the differences among subspecies within the *M. tuberculosis* complex or how these differences have evolved.

Sequence analysis of *M. tuberculosis complex* isolates has revealed that allelic polymorphism is extremely rare, occurring in ~1 in 10,000 bp [5]. Three genetic groups of *M. tuberculosis* complex (hereafter referred to as principal genetic groups 1, 2, and 3) have been identified on the basis of 2 single-nucleotide polymorphisms (SNPs) that occur in the *katG* and *gyrA* genes [5]. Principal genetic group 1 *M. tuberculosis* was deemed to be evolutionarily old and was allied with the *M. tuberculosis* complex "ancestor," *M. microti, M. africanum,* and *M. bovis*. Principal genetic groups 2 and 3 of *M. tuberculosis* were thought to have phylogenetically followed from group 1. More recently, genomic comparisons between the completed genome sequence of *M. tuberculosis* H37Rv [6] and *M. bovis* have revealed a number of large sequence polymorphisms (LSPs) that appear to distinguish virulent *M. bovis* isolates relative to *M. tuberculosis* H37Rv [7–9]. To date, only deletions relative to the sequenced strains of *M. tuberculosis* (H37Rv and CDC1551) have been identified. However, analysis of the complete genome sequence of *M. bovis* 2122 (available at http://www.sanger.ac.uk/Projects/M_bovis/) reveals that the bovine strain is shorter by 66,037 bp, and no genomic region that is unique to *M. bovis* but is consistently absent from *M. tuberculosis* has been uncovered to date [10, 11]. Unlike the nonsynonymous SNPs described above, these deletions include genes predicted to affect a range of metabolic pathways and putative virulence factors [7, 8]. Coupled with the observation that some of these deletions are variably absent from other subspecies of the *M. tuberculosis* complex [8–11], deletions promise to be an important source of variability among complex members.

To explore the evolutionary relationship of members of the *M. tuberculosis* complex, we tested for the presence or absence of deletions within complex isolates derived from different hosts and isolated in various geographic locales. The same samples were also tested for the SNPs within the *katG* and *gyrA* genes, to determine the concordance of a deletion-based scenario with that proposed by analysis of SNPs. Because genomic deletions are expected to represent unidirectional genetic events [7, 12], the distribution of the deletions suggests a phylogeny for the *M. tuberculosis* complex.

Received 7 January 2002; revised 5 March 2002; electronically published 30 May 2002.

Presented in part: general meeting of the American Society for Microbiology, Salt Lake City, Utah, 19–23 May 2002 (poster R34).

Financial support: Canadian Institutes of Health Research (grant GOP 36054).

[a] Present affiliation: Micrologix Biotech, Vancouver, British Columbia, Canada.

Reprints or correspondence: Dr. Marcel A. Behr, Montreal General Hospital, Div. of Infectious Diseases and Medical Microbiology, A5-156, 1650 Cedar Ave., Montreal, Quebec H3G 1A4, Canada (marcel.behr@mcgill.ca).

**The Journal of Infectious Diseases 2002; 186:74–80**

0022-1899/2002/18601-0010$15.00

## Materials and Methods

*Bacterial isolates.* DNA from 66 *M. tuberculosis* complex isolates was provided by 2 laboratories (those of D.C. and A.A.). On the basis of classic tests of the *M. tuberculosis* complex [13], isolates were classified as *M. tuberculosis* ($n = 12$), *M. bovis* ($n = 17$), *M. caprae* ($n = 10$), *M. africanum* ($n = 10$), *M. microti* ($n = 7$), or seal bacilli ($n = 10$). Bacteria had been isolated in 8 different countries from 9 different animal hosts. Nearly all these samples had been genotyped previously by restriction fragment–length polymorphism (RFLP) [14, 15], using the molecular epidemiologic markers IS*6110* [16] (63 of 66 isolates), polymorphic GC-rich sequence (PGRS) [17] (58 of 66 isolates), and direct repeat (DR) [18] (54 of 66 isolates), and by spoligotyping [19, 20] (64 of 66 isolates). In total, 41 different IS*6110* genotypes, 40 different PGRS genotypes, 31 different DR genotypes, and 34 different spoligotypes were provided from these 66 samples. Samples were coded numerically for genetic analysis and were blinded to their origin, species designation, and genotype.

*Deletions.* We tested samples for genomic regions present in *M. tuberculosis* H37Rv and absent from virulent isolates of *M. bovis* [7–9]. For RD (region deleted) regions, we followed the nomenclature of Gordon et al. [8], with the additional inclusion of N-RD (new region deleted) regions discovered by Salamon et al. [9]. LSPs uncovered to date are RD3, RD4, RD5, RD6, RD7, RD8, RD9, RD10, RD11, RD12, RD13, N-RD17, and N-RD25. The phages (RD3 and RD11) were excluded from testing for 2 reasons. First, these elements already have been observed to have a variable presence in *M. tuberculosis* [12, 21] and *M. bovis* isolates [7]. In addition, because of the variable genomic location of these phages within the chromosome (available at http://www.tigr.org), it is not possible to demonstrate by site-specific polymerase chain reaction (PCR) the deletion of these elements from a precise locus of the genome. Of the remaining LSPs, we included only those expected to represent deletions from *M. bovis,* as opposed to insertions into *M. tuberculosis*. To select these, our criteria were 2-fold. First, LSPs must have at least 1 junction point truncating an open-reading frame (ORF) in H37Rv. This was done because a genomic insertion into an intergenic region would not disrupt ORF structure, whereas such an insertion into an existing ORF would not be expected to complete a truncated ORF. Second, the LSP must extend beyond a single ORF. Otherwise, it would not be possible to distinguish an insertion into an ORF that does not disrupt the coding sequence from a deletion incurred within that same ORF. According to these criteria, N-RD17, which is located entirely within an ORF *(Rv3479),* was excluded from our analysis. The following LSPs met the stated criteria and were assessed as being present or absent in our samples: RD4, RD5, RD6, RD7, RD8, RD9, RD10, RD12, RD13, and N-RD25.

*PCR amplification and sequencing across deletions.* To search for the presence or absence of the deleted regions, we subjected each of our 66 samples to a 3-primer PCR, as described by Talbot et al. [22] for the study of bacille Calmette-Guérin (BCG)–specific RD1. In brief, for each deleted region, we designed one pair of PCR primers beyond the region (forward and reverse) that would amplify should the genomic region be absent. A third primer (reverse) was designed within the deleted region. Amplification resulting from this primer and the forward primer would result in a PCR product of a different size, indicating that the genomic region was present. For each experiment, *M. tuberculosis* H37Ra was used as the control for the presence of the region, and *M. bovis* BCG Pasteur was used as the control for the absence of the region. To confirm the precise genomic locus of the deletion, we performed PCR-based sequencing across deleted regions at the Montreal Genome Center using the ABI PRISM Dye Terminator Cycle sequencing kit and a Prism 3700 automated sequencer (both from PE Applied Biosystems).

*SNP analysis.* Sreevatsan et al. [5] assigned isolates of *M. tuberculosis* complex organisms to 1 of 3 genotypic groups, on the basis of the combinations of polymorphisms at *katG* codon 463 and *gyrA* codon 95. Principal genetic group 1 has the allele combination *katG* codon 463 CTG and *gyrA* codon 95 ACC, principal genetic group 2 has *katG* codon 463 CGG and *gyrA* codon 95 ACC, and principal genetic group 3 has *katG* codon 463 CGG and *gyrA* codon 95 AGC. To investigate the arrangement of the *katG* and *gyrA* SNPs within our samples, we designed forward and reverse primers to amplify a fragment of the *katG* and *gyrA* genes housing the site of polymorphism. The *katG* PCR product was digested with *Msp*I, a restriction enzyme that cleaves at a C/CGG sequence. Digestion with *Msp*I produces 4 DNA fragments in isolates with *katG* codon 463 CGG and 3 DNA fragments in isolates with *katG* codon 463 CTG. Because there is no appropriate enzyme to cleave *gyrA* around the desired region, we sequenced our PCR product (as described above) to determine polymorphism occurring at *gyrA* codon 95. Again, *M. tuberculosis* H37Ra (principal genetic group 3, *katG* codon 463 CGG and *gyrA* codon 95 AGC) and *M. bovis* BCG Pasteur (principal genetic group 1, *katG* codon 463 CTG and *gyrA* codon 95 ACC) were used as controls.

*Analysis of deletions and SNP results.* On the basis of review of gels by 2 independent readers, regions were deemed to be present or absent, without knowledge of subspecies assignment, host, or geographic provenance. PCR amplicons across deletions were sequenced, and the bridging sequence was compared with that of H37Rv by Tuberculist Blastserver (available at http://genolist.pasteur.fr/TubercuList/) to delineate the exact junctions of the deletion. The *katG* SNP was assessed by PCR-RFLP, with gels read by 2 independent readers. The *gyrA* SNP was assessed by sequencing and compared with H37Rv by Tuberculist Blastserver (http://genolist.pasteur.fr/TubercuList/). All these results were assembled into a single framework that provides the most parsimonious scenario (table 1). Subsequently, names of isolates and provenance were superimposed onto this framework (table 2).

## Results

We tested a total of 66 DNA samples but obtained data for all deletions from only 63 samples because of insufficient DNA in 3 of them. All 63 samples providing deletion data were amenable to sequence confirmation of deletion junction and to SNP analysis by PCR-RFLP or sequencing.

Sequence analysis of all deletions but RD6 confirmed that these represent the same genetic event, because the flanking sequences observed are identical for all isolates (data not

**Table 1.** Large sequence polymorphisms among isolates of the *Mycobacterium tuberculosis* complex: distribution of regions (RD [region deleted] and N-RD [new region deleted]) present (+) or absent (−) in isolates of the *M. tuberculosis* complex.

| Isolate | RD9 | RD10 | RD8 | RD7 | RD12 | RD5 | RD13 | RD6 | N-RD25 | RD4 | *gyrA* | *katG* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| H37Ra | + | + | + | + | + | + | + | + | + | + | AGC | CGG |
| 15 | + | + | + | + | + | + | + | + | + | + | AGC | CGG |
| 55 | + | + | + | + | + | + | + | + | + | + | AGC | CGG |
| 13 | + | + | + | + | + | + | + | + | + | + | ACC | CGG |
| 16 | + | + | + | + | + | + | + | + | + | + | ACC | CGG |
| 26 | + | + | + | + | + | + | + | + | + | + | ACC | CGG |
| 40 | + | + | + | + | + | + | + | + | + | + | ACC | CGG |
| 17 | + | + | + | + | + | + | + | + | + | + | ACC | CTG |
| 27 | + | + | + | + | + | + | + | + | + | + | ACC | CTG |
| 50 | + | + | + | + | + | + | + | + | + | + | ACC | CTG |
| 44 | + | + | + | + | + | + | + | + | + | + | ACC | CTG |
| 52 | + | + | + | + | + | + | + | + | + | + | ACC | CTG |
| 24 | + | + | + | + | + | + | + | + | + | + | ACC | CTG |
| 7 | + | + | + | + | + | + | + | + | + | + | ACC | CTG |
| 8 | + | + | + | + | + | + | + | + | + | + | ACC | CTG |
| 10 | − | + | + | + | + | + | + | + | + | + | ACC | CTG |
| 1 | − | − | − | − | + | + | + | + | + | + | ACC | CTG |
| 19 | − | − | − | − | + | + | + | + | + | + | ACC | CTG |
| 2 | − | − | − | − | + | + | + | + | + | + | ACC | CTG |
| 3 | − | − | − | − | + | + | + | + | + | + | ACC | CTG |
| 20 | − | − | − | − | + | + | + | + | + | + | ACC | CTG |
| 38 | − | − | − | − | + | + | + | + | + | + | ACC | CTG |
| 4 | − | − | − | − | + | + | + | + | + | + | ACC | CTG |
| 21 | − | − | − | − | + | + | + | + | + | + | ACC | CTG |
| 22 | − | − | − | − | + | + | + | + | + | + | ACC | CTG |
| 39 | − | − | − | − | + | + | + | + | + | + | ACC | CTG |
| 5 | − | − | − | − | + | + | + | + | + | + | ACC | CTG |
| 6 | − | − | − | − | + | + | + | + | + | + | ACC | CTG |
| 28 | − | − | − | − | + | + | + | + | + | + | ACC | CTG |
| 29 | − | − | − | − | + | + | + | + | + | + | ACC | CTG |
| 30 | − | − | − | − | + | + | + | + | + | + | ACC | CTG |
| 31 | − | − | − | − | + | + | + | + | + | + | ACC | CTG |
| 32 | − | − | − | − | + | + | + | + | + | + | ACC | CTG |
| 41 | − | − | − | − | + | + | + | + | + | + | ACC | CTG |
| 42 | − | − | − | − | + | + | + | + | + | + | ACC | CTG |
| 43 | − | − | − | − | + | + | + | + | + | + | ACC | CTG |
| 51 | − | − | − | − | + | + | + | + | + | + | ACC | CTG |
| 56 | − | − | − | − | + | + | + | + | + | + | ACC | CTG |
| 57 | − | − | − | − | − | − | − | − | − | + | ACC | CTG |
| 58 | − | − | − | − | − | − | − | − | − | + | ACC | CTG |
| 59 | − | − | − | − | − | − | − | − | − | + | ACC | CTG |
| 60 | − | − | − | − | − | − | − | − | − | + | ACC | CTG |
| 61 | − | − | − | − | − | − | − | − | − | + | ACC | CTG |
| 62 | − | − | − | − | − | − | − | − | − | + | ACC | CTG |
| 63 | − | − | − | − | − | − | − | − | − | + | ACC | CTG |
| 64 | − | − | − | − | − | − | − | − | − | + | ACC | CTG |
| 65 | − | − | − | − | − | − | − | − | − | + | ACC | CTG |
| 66 | − | − | − | − | − | − | − | − | − | + | ACC | CTG |
| 11 | − | − | − | − | − | − | − | − | − | − | ACC | CTG |
| 12 | − | − | − | − | − | − | − | − | − | − | ACC | CTG |
| 25 | − | − | − | − | − | − | − | − | − | − | ACC | CTG |
| 9 | − | − | − | − | − | − | − | − | − | − | ACC | CTG |
| 23 | − | − | − | − | − | − | − | − | − | − | ACC | CTG |
| 33 | − | − | − | − | − | − | − | − | − | − | ACC | CTG |
| 34 | − | − | − | − | − | − | − | − | − | − | ACC | CTG |
| 35 | − | − | − | − | − | − | − | − | − | − | ACC | CTG |
| 36 | − | − | − | − | − | − | − | − | − | − | ACC | CTG |
| 45 | − | − | − | − | − | − | − | − | − | − | ACC | CTG |
| 46 | − | − | − | − | − | − | − | − | − | − | ACC | CTG |
| 47 | − | − | − | − | − | − | − | − | − | − | ACC | CTG |
| 48 | − | − | − | − | − | − | − | − | − | − | ACC | CTG |
| 49 | − | − | − | − | − | − | − | − | − | − | ACC | CTG |
| 53 | − | − | − | − | − | − | − | − | − | − | ACC | CTG |
| 54 | − | − | − | − | − | − | − | − | − | − | ACC | CTG |
| BCG | − | − | − | − | − | − | − | − | − | − | ACC | CTG |

NOTE. The last 2 columns present the allele combination at *katG* codon 463 and *gyrA* codon 95 for each isolate. BCG, bacille Calmette-Guérin.

shown). We could not sequence confirm the exact site of RD6 because it represents a highly repetitive region. Therefore, we have not assigned the same level of confidence that absence of RD6 consistently represents the same LSP.

We observed that certain samples contain all regions ($n$ = 14), certain samples lack all regions deleted from virulent *M. bovis* ($n$ = 16), and the remaining samples provide intermediate deletion results ($n$ = 33; table 1). Of the samples demonstrating deletions ($n$ = 49), samples with greater numbers of deletions, without exception, were missing all the same regions that were deleted from samples with lesser numbers of deletions. For example, certain samples lack 4 regions (RD7, RD8, RD9, and RD10), and others are missing 9 regions, which include the aforementioned 4 regions plus 5 extra deletions (RD5, RD6, RD12, RD13, and N-RD25). Analysis of deletions suggests that there are at least 5 deletion types within the complex, based on no deletions, 1 deletion, 4 deletions, 9 deletions, and all 10 regions missing. Analysis of SNPs revealed that the group with no deletions could be further subdivided into principal genetic groups 1, 2, and 3, whereas all other samples fell into principal genetic group 1 (table 1).

Associating isolates with various molecular epidemiological markers (IS*6110* RFLP, PGRS RFLP, DR RFLP, and spoligotype) revealed that these typing modalities grouped isolates almost exclusively within deletion types (data available at http://www.molepi.mcgill.ca/Mtbcomplex). In other words, isolates grouped together by deletion type had different IS*6110* RFLP patterns, but all isolates with the same IS*6110* RFLP pattern had the same deletion type (IS*6110* patterns common to >1 sample were seen 9 times). This same scenario is also observed for PGRS RFLP, in which 9 patterns were seen in >1 sample, but each pattern was found only in 1 deletion type. There were 9 DR RFLP patterns shared by >1 sample; 8 DR patterns were each restricted to a single deletion type, but 1 DR pattern was observed in 2 samples that had different deletion types. There were 12 spoligotype patterns shared by >1 sample; 11 spoligotype patterns were each restricted to a single deletion type, but 1 spoligotype pattern was observed in 2 samples that had different deletion types.

When we superimposed the assigned names to DNA samples (table 2), we observed that the deletion groups correspond to the following subspecies groupings: *M. tuberculosis, M. microti*/seal bacillus, *M. caprae,* and *M. bovis*. Notably, samples with a number of different deletion types were called *M. africanum*. Specifically, samples with all regions present were assigned to *M. tuberculosis* or *M. africanum*. There was 1 sample in which only 1 region was absent, called *M. africanum*. Samples missing 4 regions were called *M. africanum, M. microti,* or seal bacillus. Samples missing 9 regions were called *M. caprae*. Samples missing 10 regions were assigned to *M. bovis*. A phylogenetic scenario suggested by this pattern of deletions is put forward in figure 1.

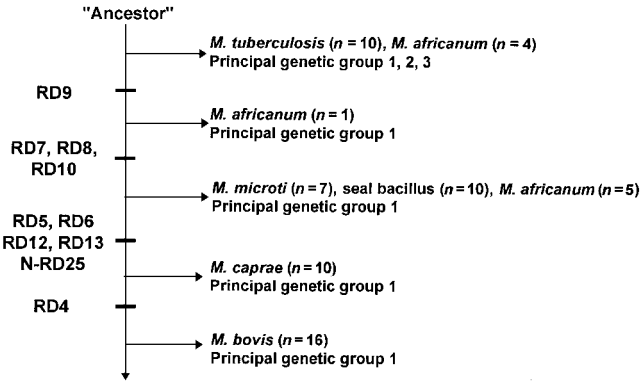


**Figure 1.** Hypothetical phylogeny of the *Mycobacterium tuberculosis* complex. Suggested phylogeny is derived from analysis of genomic deletions (RD [region deleted] and N-RD [new region deleted]). The vertical axis represents the loss of genomic regions within isolates of the *M. tuberculosis* complex. Each horizontal axis clusters isolates on the presence or absence of genomic regions. Principal genetic groups (1, 2, and 3) are indicated for isolates within each genomic grouping. Under this phylogeny, isolates labeled *M. tuberculosis* are characterized by no genomic deletions and are closest to the "ancestor" of the *M. tuberculosis* complex. Deletions accumulate stepwise within isolates labeled *M. microti*/seal bacillus, *M. caprae,* and *M. bovis*. Isolates labeled *M. africanum* could not be assigned to a single deletion type.

## Discussion

The availability of the whole genome sequence of *M. tuberculosis* H37Rv and comparative genomic tools has provided an unprecedented opportunity to make sequence comparisons among the *M. tuberculosis* complex [6–9]. From these comparisons, several regions have been identified that are present in *M. tuberculosis* H37Rv but absent from virulent *M. bovis* isolates. Without having the ancestor of the *M. tuberculosis* complex, we cannot say with absolute certainty whether these LSPs are deletions from the *M. bovis* genome, as opposed to insertions into the *M. tuberculosis* genome. However, it is compelling that the LSPs are deletions, because the ORF structures at junction points are truncated [11]. This notion follows from the reasoning that insertion of a polygenic region into the *M. bovis* genome, thereby completing an existing truncated ORF, is much less likely than disruption of that ORF during a deletion from H37Rv. Furthermore, genes at the junction points have homologues in *M. avium* and *M. smegmatis* (available at http://www.tigr.org). It would therefore be difficult to postulate that these genes are present in mycobacteria with larger genomes, were absent in the ancestral member of the *M. tuberculosis* complex, but were subsequently reacquired only in the *M. tuberculosis* lineage of the *M. tuberculosis* complex.

Empirical support for genomic deletions representing unidirectional genetic events has been established recently by 2 separate studies and is supported by data presented here. In a genomic study of BCG vaccine strains, Behr et al. [7] documented that BCG-specific deletions superimpose perfectly on

**Table 2.** Association of deletion types with subspecies of the *Mycobacterium tuberculosis* complex and molecular typing data.

| Isolate | No. of deletions | Principal genetic group | Subspecies assigned | Host | Geographic origin | IS*6110* RFLP | PGRS RFLP | DR RFLP | Spoligotype |
|---|---|---|---|---|---|---|---|---|---|
| H37Ra | 0 | 3 | *M. tuberculosis* | Reference strain H37Ra | TMC | ND | ND | ND | ND |
| 7 | 0 | 1 | *M. africanum* | Human | Australia | IS49 | PG97 | DR48 | Sp57 |
| 8 | 0 | 1 | *M. africanum* | Human | Australia | IS50 | PG98 | DR48 | Sp58 |
| 17 | 0 | 1 | *M. tuberculosis* | Human | Australia | IS54 | PG125 | DR62 | Sp72 |
| 24 | 0 | 1 | *M. africanum* | Human | Australia | IS52 | PG100 | DR41 | Sp68 |
| 27 | 0 | 1 | *M. tuberculosis* | Human | Australia | IS64 | ND | DR71 | Sp65 |
| 44 | 0 | 1 | *M. tuberculosis* | Bovine | Iran | IS11 | PG29 | DR22 | Sp56 |
| 50 | 0 | 1 | *M. tuberculosis* | Human | Australia | IS54 | PG125 | DR62 | Sp07 |
| 52 | 0 | 1 | *M. tuberculosis* | Human | Australia[a] | IS12 | PG30 | DR18 | Sp17 |
| 13 | 0 | 2 | *M. africanum* | Human | Australia | IS67 | PG144 | DR66 | Sp79 |
| 16 | 0 | 2 | *M. tuberculosis* | Human | Australia | IS77 | ND | DR53 | SP60 |
| 26 | 0 | 2 | *M. tuberculosis* | Human | Australia | IS79 | ND | DR68 | Sp54 |
| 40 | 0 | 2 | *M. tuberculosis* | Human | Australia | IS26 | PG127 | DR59 | Sp59 |
| 15 | 0 | 3 | *M. tuberculosis* | Reference strain H37Rv | WA reference laboratory | IS22 | PG126 | DDR60 | Sp16 |
| 55 | 0 | 3 | *M. tuberculosis* | Human | Australia | IS76 | ND | DR60 | Sp54 |
| 10 | 1 | 1 | *M. africanum* | Human | Australia | IS53 | PG101 | DR62 | Sp69 |
| 1 | 4 | 1 | *M. microti* | Reference strain | TMC | IS27 | PG123 | DR64 | Sp67 |
| 2 | 4 | 1 | *M. microti* | Dassie | Australia[b] | IS23 | PG121 | DR38 | Sp71 |
| 3 | 4 | 1 | *M. microti* | Dassie | South Africa | IS24 | PG122 | DR63 | Sp70 |
| 4 | 4 | 1 | *M. microti* | Surikat | Sweden (zoo) | IS25 | PG122 | DR63 | Sp70 |
| 5 | 4 | 1 | *M. africanum* | Reference strain | TMC | IS20 | PG92 | DR36 | ND |
| 6 | 4 | 1 | *M. africanum* | Human | Australia | IS21 | PG96 | DR61 | Sp61 |
| 19 | 4 | 1 | *M. microti* | Dassie | Australia[b] | IS23 | PG121 | DR38 | Sp71 |
| 20 | 4 | 1 | *M. microti* | Vole | Wales | IS28 | PG124 | DR56 | Sp67 |
| 21 | 4 | 1 | *M. africanum* | Reference strain | TMC | IS19 | PG94 | DR35 | Sp63 |
| 22 | 4 | 1 | *M. africanum* | Human | Australia | IS51 | PG95 | DR33 | Sp62 |
| 28 | 4 | 1 | Seal bacillus | Seal | Australia | IS07 | PG28 | DR17 | Sp22 |
| 29 | 4 | 1 | Seal bacillus | Seal | Australia | IS07 | PG28 | DR17 | Sp22 |
| 30 | 4 | 1 | Seal bacillus | Seal | Australia | IS18 | PG33 | DR17 | Sp22 |
| 31 | 4 | 1 | Seal bacillus | Human | Australia | IS07 | PG28 | DR17 | Sp22 |
| 32 | 4 | 1 | Seal bacillus | Seal | Australia | IS07 | PG28 | DR17 | Sp22 |
| 38 | 4 | 1 | *M. microti* | Vole | Wales | ND | ND | ND | Sp67 |
| 39 | 4 | 1 | *M. africanum* | Reference strain | TMC | IS19 | PG93 | DR34 | Sp64 |
| 41 | 4 | 1 | Seal bacillus | Seal | Australia | IS17 | PG32 | DR17 | Sp22 |
| 42 | 4 | 1 | Seal bacillus | Seal | Australia | IS07 | PG28 | DR17 | Sp22 |
| 43 | 4 | 1 | Seal bacillus | Seal | Australia | IS07 | PG28 | DR17 | Sp22 |
| 51 | 4 | 1 | Seal bacillus | Seal | Australia | IS07 | PG28 | DR17 | Sp22 |
| 56 | 4 | 1 | Seal bacillus | Seal | Australia | IS07 | PG28 | DR17 | Sp22 |
| 57 | 9 | 1 | *M. caprae* | Goat | Spain | 5D | 27 | ND | Spc-1 |
| 58 | 9 | 1 | *M. caprae* | Goat | Spain | 5D | 27 | ND | Spc-1 |
| 59 | 9 | 1 | *M. caprae* | Goat | Spain | 5D | 27 | ND | Spc-1 |
| 60 | 9 | 1 | *M. caprae* | Goat | Spain | 5E | 26 | ND | Spc-3 |
| 61 | 9 | 1 | *M. caprae* | Goat | Spain | 5E | 26 | ND | Spc-3 |
| 62 | 9 | 1 | *M. caprae* | Goat | Spain | ND | ND | ND | Spc-3 |
| 63 | 9 | 1 | *M. caprae* | Sheep | Spain | 6A | 28 | ND | Spc-2 |
| 64 | 9 | 1 | *M. caprae* | Goat | Spain | 6A | 28 | ND | Spc-2 |
| 65 | 9 | 1 | *M. caprae* | Goat | Spain | 7C | 28 | ND | Spc-2 |
| 66 | 9 | 1 | *M. caprae* | Goat | Spain | 5E | 26 | ND | Spc-3 |
| 9 | 10 | 1 | *M. bovis* | Bovine | Malawi | IS58 | PG135 | DR10 | Sp73 |
| 11 | 10 | 1 | *M. bovis* | Bovine | Malawi | IS58 | PG138 | DR10 | Sp73 |
| 12 | 10 | 1 | *M. bovis* | Bovine | Malawi | IS59 | PG142 | DR10 | Sp73 |
| 23 | 10 | 1 | *M. bovis* | Bovine | Malawi | IS01 | PG141 | DR07 | Sp07 |
| 25 | 10 | 1 | *M. bovis* | Bovine | Malawi | IS58 | PG140 | DR10 | Sp73 |
| 33 | 10 | 1 | *M. bovis* | Buffalo | Australia | IS48 | PG87 | DR50 | Sp25 |
| 34 | 10 | 1 | *M. bovis* | Bovine | Australia | IS02 | PG01 | DR25 | Sp19 |
| 35 | 10 | 1 | *M. bovis* | Buffalo | Australia | IS01 | PG04 | DR09 | Sp08 |
| 36 | 10 | 1 | *M. bovis* | Human | Australia | IS04 | PG11 (H14) | DR10 | Sp15 |
| 45 | 10 | 1 | *M. bovis* | Human | Australia | IS06 | PG01 | DR01 | Sp01 |
| 46 | 10 | 1 | *M. bovis* | Bovine | Australia | IS13 | PG83 | DR05 | Sp05 |
| 47 | 10 | 1 | *M. bovis* | Bovine | Australia | IS41 | PG57 | DR44 | Sp40 |
| 48 | 10 | 1 | *M. bovis* | Buffalo | Australia | IS01 | PG151 | DR09 | Sp08 |

*(continued)*

**Table 2.** *(Continued.)*

| Isolate | No. of deletions | Principal genetic group | Subspecies assigned | Host | Geographic origin | IS*6110* RFLP | PGRS RFLP | DR RFLP | Spoligotype |
|---|---|---|---|---|---|---|---|---|---|
| 49 | 10 | 1 | *M. bovis* | Buffalo | Australia | IS01 | PG04 | DR72 | Sp08 |
| 53 | 10 | 1 | *M. bovis* | Bovine | Australia | IS01 | PG01a | DR01 | Sp01 |
| 54 | 10 | 1 | *M. bovis* | Bovine | Australia | IS01 | PG42f | DR01 | Sp01 |
| BCG | 10 | 1 | BCG Pasteur | Bovine | France | ND | ND | ND | ND |

NOTE. Isolates shown here are the same as those shown in table 1. Control isolates (H37Ra, with no deletions, and bacille Calmette-Guérin [BCG] Pasteur, with all 10 regions deleted) are placed at top and bottom of table, respectively. Isolates are ordered by no. of genomic deletions and then by principal genetic group. The final 4 columns provide arbitrary strain types according to IS*6100*-based restriction fragment–length polymorphism (RFLP), polymorphic GC-rich sequence (PGRS)–based RFLP, direct repeat (DR)–based RFLP, and spoligotyping. ND, not done; TMC, Trudeau mycobacterial collection; WA, Western Australia.

[a] Born in Singapore.

[b] Animals from South Africa.

the historical record, demonstrating that deletions represent unidirectional events in the evolution of BCG. In a study of genomic deletions within clinical isolates of *M. tuberculosis,* mycobacterial clones shared the same genomic deletions, again suggesting that deletions can be used to reconstruct phylogenetic trees [12]. In the present analysis, samples sharing typing patterns have the same deletion profile with just 2 exceptions, both involving the DR region of the genome. These different studies all support the potential value of genomic deletions as evolutionary markers, despite the possibility that deletions offer a medium for selection to act upon. Selective pressure leads to the theoretical concern that different strains may converge to delete the same genomic region. If this were to occur, such a phenomenon would confuse any phylogenetic inference. However, deletion convergence is highly unlikely, because independent deletion of a set of genes would not be expected to reproducibly occur at the exact same base pair, and, empirically, such an event has not been observed in samples studied to date. It therefore follows that regions of *M. tuberculosis* H37Rv missing from virulent *M. bovis* can suggest a phylogeny for the *M. tuberculosis* complex.

The distribution of deletions suggests their order of occurrence during bacterial evolution (figure 1). By overlaying SNP results onto the deletion-based phylogeny, it can be seen that the 3 principal genetic groups are distinguished within isolates harboring no genomic deletions, whereas all samples missing $\geqslant$1 genomic region fell into principal genetic group 1. A phylogeny based on genomic deletions therefore supplements the previous SNP classification system proposed by Sreevatsan et al. [5]. From figure 1, it appears that the human form of tuberculosis (TB) preceded the form observed in voles, seals, goats, and cows. This scenario concurs with deletion data presented by Gordon et al. [8], arguing against a bovine origin of TB [10], and is supported by the recent archeological find of an isolate from a 17,000-year-old bison that grouped more closely with *M. tuberculosis* and *M. africanum* than *M. bovis* [23]. Thus, the distribution of genomic deletions suggests that present-day *M. bovis* is not the evolutionary precursor of *M. tuberculosis* and argues against the commonly expressed belief that human TB originated with the domestication of cattle [24].

There are several important limitations in this analysis that suggest the need for further study. First, the deletions tested derive from a monophyletic comparison of *M. tuberculosis* with *M. bovis* and do not include deletions found in *M. tuberculosis* isolates or expected deletion polymorphisms that one can anticipate in the lineages of *M. microti,* seal bacillus, *M. caprae,* and *M. bovis*. Clearly, the demonstration of deletions specific to each of these subspecies will help provide a more complete picture of the *M. tuberculosis* complex. At that point, it will be possible to revisit the phylogenetic scenario suggested by the current data to determine whether the deletions studied here still provide a backbone for the *M. tuberculosis* complex. A second limitation relates to the number of samples tested. An example of this is the independent deletion of RD9 that was observed in only one sample (*M. africanum;* table 2). Although this finding is in accordance with previous genomic characterization of *M. africanum* [8], we cannot assign the same level of confidence to the phylogenetic occurrence of this deletion on the basis of just one observation. A third limitation is the diversity of samples studied. We have tested each of the classic subspecies, and the 63 samples come from a total of 8 different countries, 9 different hosts, 39 different IS*6110* RFLP patterns, 39 different PGRS RFLP patterns, 30 different DR RFLP patterns, and 33 different spoligotype patterns. However, the fact that intermediate numbers of deletions (2, 3, 5, 6, 7, or 8 deletions) were not detected suggests that other forms of the *M. tuberculosis* complex were not sampled, either because they were geographically absent in the isolates tested or because they represent transition forms that are no longer present. To resolve this, a larger number of samples from a greater geographic range will be needed. A final limitation comes from the difficulty in assigning a link between organisms called *M. africanum* and deletion-based genotyping. This discrepancy is best explained as an artifact of ambiguous classification, because no simple set of phenotypic features is generally accepted for *M africanum,* with some researchers suggesting that isolates not man-

ifesting either of the classic phenotypes for *M. tuberculosis* or *M. bovis* are arbitrarily assigned to *M. africanum* [25].

Because subspecies of the *M. tuberculosis* complex are thought to have a clonal genetic structure [5, 12, 26, 27], genomic deletions among them promise to be a consequential source of variability. In the present study, which examines isolates of the *M. tuberculosis* complex, samples with greater numbers of genomic deletions, without exception, were missing all the same regions that were deleted from samples with lesser numbers of deletions, suggesting that the deletions tested are unidirectional events and can serve as clonal genetic markers. Genomic deletions may assist in the evolutionary analysis of other clonal organisms and, coupled with the strong phylogenetic relevance of gene content from whole-genome comparison of divergent bacteria [28], may hold promise as an important tool in the study of bacterial evolution.

**Acknowledgments**

We thank Carol Dore and Pierre LePage (Montreal Genome Center), for their sequencing efforts, and Chris Daborn and Goran Bolske, for supplying isolates from Malawi and the surikat, respectively, to Debby Cousins.

**References**

1. Aranaz A, Liébana E, Gomez-Mampaso E, et al. *Mycobacterium tuberculosis* subsp. *caprae* subsp. nov.: a taxonomic study of a new member of the *Mycobacterium tuberculosis* complex isolated from goats in Spain. Int J Syst Bacteriol **1999**;49:1263–73.
2. Cousins DV, Williams SN, Reuter R, et al. Tuberculosis in wild seals and characterization of the seal bacillus. Australian Veterinary Journal **1993**; 70:92–7.
3. Brosch R, Gordon SV, Pym A, Eiglmeier K, Garnier T, Cole ST. Comparative genomics of the mycobacteria. Int J Med Microbiol **2000**;290:143–52.
4. Sales MP, Taylor GM, Hughes S, et al. Genetic diversity among *Mycobacterium bovis* isolates: a preliminary study of strains from animal and human Sources. J Clin Microbiol **2001**;39:4558–62.
5. Sreevatsan S, Pan X, Stockbauer KE, et al. Restricted structural gene polymorphism in the *Mycobacterium tuberculosis* complex indicates evolutionarily recent global dissemination. Proc Natl Acad Sci USA **1997**;94: 9869–74.
6. Cole ST, Brosch R, Parkhill J, et al. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. Nature **1998**;393: 537–44.
7. Behr MA, Wilson MA, Gill WP, et al. Comparative genomics of BCG vaccines by whole-genome DNA microarray. Science **1999**;284:1520–3.
8. Gordon SV, Brosch R, Billault A, Garnier T, Eiglmeier K, Cole ST. Identification of variable regions in the genomes of tubercle bacilli using bacterial artificial chromosome arrays. Mol Microbiol **1999**;32:643–55.
9. Salamon H, Kato-Maeda M, Small PM, Drenkow J, Gingeras TR. Detection of deleted genomic DNA using a semiautomated computational analysis of GeneChip data. Genome Res **2000**;10:2044–54.
10. Brosch R, Pym AS, Gordon SV, Cole ST. The evolution of mycobacterial pathenogenicity: clues from comparative genomics. Trends Microbiol **2001**;9:452–8.
11. Gordon SV, Eiglmeier K, Garnier T, et al. Genomics of *Mycobacterium bovis*. In: Ellner JJ, Brennan PJ, Young D, eds. Tuberculosis. Vol. 81. Edinburgh, United Kingdom: Harcourt Publishers, **2001**:157–63.
12. Kato-Maeda M, Rhee JT, Gingeras TR, et al. Comparing genomes within the species *Mycobacterium tuberculosis*. Genome Res **2001**;11:547–54.
13. Heifets LB, Good RC. Current laboratory methods for the diagnosis of tuberculosis. In: Bloom BR, ed. Tuberculosis: pathogenesis, protection and control. Washington, DC: American Society for Microbiology Press, **1994**:85–110.
14. Cousins D, Williams S, Liebana E, et al. Evaluation of four DNA typing techniques in epidemiological investigations of bovine tuberculosis. J Clin Microbiol **1998**;36:168–78.
15. Liébana E, Aranaz A, Dominguez L, et al. The insertion element IS*6110* is a useful tool for DNA fingerprinting of *Mycobacterium bovis* isolates from cattle and goats in Spain. Vet Microbiol **1997**;54:223–33.
16. van Embden JD, Cave MD, Crawford JT, et al. Strain identification of *Mycobacterium tuberculosis* by DNA fingerprinting: recommendations for a standardized methodology. J Clin Microbiol **1993**;31:406–9.
17. Ross BC, Raios K, Jackson K, Dwyer B. Molecular cloning of a highly repeated DNA element from *Mycobacterium tuberculosis* and its use as an epidemiological tool. J Clin Microbiol **1992**;30:942–6.
18. Hermans PW, van Soolingen D, Bik EM, de Haas PE, Dale JW, van Embden JD. Insertion element IS*987* from *Mycobacterium bovis* BCG is located in a hot-spot integration region for insertion elements in *Mycobacterium tuberculosis* complex strains. Infect Immun **1991**;59:2695–705.
19. Aranaz A, Liébana E, Mateos A, et al. Spoligotyping of *Mycobacterium bovis* strains from cattle and other animals: a tool for epidemiology of tuberculosis. J Clin Microbiol **1996**;34:2734–40.
20. Kamerbeek J, Schouls L, Kolk A, et al. Simultaneous strain detection and differentiation of *Mycobacterium tuberculosis* for diagnosis and epidemiology. J Clin Microbiol **1997**;35:907–14.
21. Mahairas GG, Sabo PJ, Hickey MJ, Singh DC, Stover CK. Molecular analysis of genetic differences between *Mycobacterium bovis* BCG and virulent *M. bovis*. J Bacteriol **1996**;178:1274–82.
22. Talbot EA, Williams DL, Frothingham R. PCR identification of *Mycobacterium bovis* BCG. J Clin Microbiol **1997**;35:566–9.
23. Rothschild BM, Martin LD, Lev G, et al. *Mycobacterium tuberculosis* complex DNA from an extinct bison dated 17,000 years before the present. Clin Infect Dis **2001**;33:305–11.
24. Stead WW. The origin and erratic global spread of tuberculosis: how the past explains the present and is the key to the future. Clin Chest Med **1997**; 18:65–77.
25. Grange JM. *Mycobacterium bovis* infection in human beings. In: Ellner JJ, Brennan PJ, Young D, eds. Tuberculosis. Vol. 81. Edinburgh, United Kingdom: Harcourt Publishers, **2001**:71–7.
26. Kremer K, van Soolingen D, Frothingham R, et al. Comparison of methods based on different molecular epidemiological markers for typing of *Mycobacterium tuberculosis* complex strains: interlaboratory study of discriminatory power and reproducibility. J Clin Microbiol **2000**;37:2607–18.
27. Musser JM, Amin A, Ramaswamy S. Negligible genetic diversity of *Mycobacterium tuberculosis* host immune system protein targets: evidence of limited selective pressure. Genetics **2000**;155:7–16.
28. Snel B, Bork P, Huynen MA. Genome phylogeny based on gene content. Nature Genetics **1999**;21:108–10.